ISSN: 0974-8571 Vol.14 No. 1 June, 2022

International Journal of Computational Intelligence in Control

Machine Learning Techniques for Cardiovascular Disease Diagnosis Using Clinical Data

Muhammad Iqbal[1*], Muhammad Zubair[1] Asghar, Sahar Batool[2]
1. Institute of Computing and Information Technology (ICIT), Gomal University, Dera Ismail Khan, 29220, Pakistan.
2. Department of Computer Science & IT, Institute of Southern Punjab, Multan, Pakistan
malikiqbalprince1@gmail.com , mzubair@gu.edu.pk , bksaharbatool@gmail.com

*Corresponding Author: (malikiqbalprince1@gmail.com)

Date of Submission: 15th March 2022 Revised: 25th May 2022 Accepted:13th June 2022

Abstract - **Cardiovascular disease (CVD) remains one of the major causes of death worldwide, which creates a strong need for accurate and early diagnostic methods to support clinical decision making. Machine learning techniques offer useful tools for analyzing medical data and identifying patterns linked to disease risk. This study presents a comparative analysis of several machine learning algorithms for cardiovascular disease prediction using clinical features obtained from the UCI heart disease dataset. The proposed framework includes data preprocessing, feature selection, and the implementation of multiple supervised learning models, including Logistic Regression, Support Vector Machine, K Nearest Neighbor, Decision Tree, Random Forest, Gradient Boosting, and Naive Bayes. The models were evaluated using comprehensive performance measures such as accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve. Experimental results show that Logistic Regression achieved the highest performance, with an accuracy of 87.91%, which indicates its suitability for predicting cardiovascular disease using structured clinical data. The results also highlight that simpler and interpretable models can perform competitively when appropriate preprocessing and feature selection methods are applied. The findings suggest that machine learning based approaches can support healthcare professionals in early diagnosis and risk assessment using routinely collected patient information. The proposed framework provides a practical and interpretable decision support approach that may contribute to improved clinical outcomes, timely intervention, and more effective preventive healthcare strategies.**

Index Terms - Cardiovascular disease, Machine learning, Clinical data, Classification, Predictive modeling, Healthcare analytics Component.

INTRODUCTION

Cardiovascular disease (CVD) is among the most common causes of death in the globe and has continued to challenge the health care systems greatly. It encompasses various heart and blood vessels disease like coronary artery disease, heart failure and stroke. Global health reports suggest that millions of people die annually of cardiovascular diseases and a large percentage of them die earlier than they should. The increasing rate of risk factors like hypertension, diabetes, obesity, sedentary and smoking lifestyle further heightens the chances of the occurrence of the diseases. Risk identification in people early is thus necessary to decrease mortality, enhance patient outcomes, and maximize the use of healthcare resources [1]. The sooner the diagnosis, the sooner the preventive interventions, and the possibility to make correct clinical decisions, which can subsequently help improve it.

Early detection of cardiovascular disease is difficult in spite of the fact that advanced technologies are available to aid medical diagnosis. Symptoms are not visible in most instances when the disease is already at its advanced stage. Conventional diagnostic methods used usually include clinical knowledge, laboratory tests, and radiographic methods, which may be time wasting and expensive. Furthermore, the complex medical data needs the involvement of qualified medical staff, and the diagnosis can be made incorrectly as the human factor can limit or incomplete information [2]. There is also the presence of various interacting risk factors that also make it hard to predict the high risk patients accurately and at an early stage. These issues emphasize the necessity to have smart and automated prediction systems that might help clinicians to make correct decisions [3].

The recent development of artificial intelligence and especially machine learning exhibited great potential in helping to solve healthcare prediction problems. The machine learning programs are able to process huge amounts of medical information, discover latent patterns, and produce predictive models which can aid in the diagnosis of a disease. Machine learning techniques are able to observe nonlinear relationships and complex interactions among clinical features unlike traditional statistical methods, which tend to assume linear relationships between variables. This feature renders them especially well-suited to medical applications where the data about the patients are heterogeneous and multidimensional. The systems relying on machine learning are capable of assisting clinicians with information-based findings, shorter diagnosis periods, and increased prediction rates [3]. The heart is the organ in the human body that is considered to be the second most vital organ, after the brain. The heart's bewilderment ultimately leads to body confusion. The world in which we live is experiencing significant changes that have a partial impact on our day-to-day lives. We are now living in the present age, and these changes are having an effect on us. The cardiovascular disease (CVD) that is the primary cause of the deaths recorded on the globe is one of the primary diseases that is among the top five deadliest diseases [4]. The prediction of this disease is very important because it helps us to take the appropriate actions at the appropriate times. The term "cardiovascular diseases" (CVD) refers to the collection of different diseases that affect the heart and vascular system, typically as a result of atherosclerosis. Most CVD is chronic, meaning that it gradually deteriorates over time without causing any symptoms for a long period of time before becoming severe and indicating symptoms to varying degrees [5]. In the early stages of cardiovascular disease, patients don't have any symptoms; however, as the disease grows, patients do have symptoms, and by the time it is sometimes too late to prevent or treat the disease [6, 7]. Thus, despite the challenge, it is essential to immediately identify and predict CVD hypersensitivity in otherwise healthy individuals in order to assess the outcome. Analyzing the substantial CVD-health data that is available in the massive database of hospital records will be important in the early diagnosis of CVD. In addition to this, machine learning algorithms and other approaches that are associated with intelligent systems are beneficial in this field, and the results that they provide are not only accurate but also reliable [8-10].

This research is driven by the fact that there is a necessity to come up with a robust and effective machine learning system to diagnose cardiovascular diseases based on clinical data. By analyzing patient characteristics from a publicly accessible dataset, machine learning models can be taught to more accurately predict the existence of a disease. Comparative analysis of several algorithms can reveal their advantages and disadvantages and help determine which the best prediction algorithm is. This type of framework can aid in the detection of conditions early, aid clinicians in decision making and decrease the uncertainty in diagnosis. Moreover, the potential of real world implementation in low resources healthcare settings is increased by using the generally available clinical data.

The purpose of the proposed study is to research the efficiency of various machine learning methods in the cardiovascular disease diagnosis through clinical attributes. Various classification algorithms would be applied and tested according to various performance parameters in order to have a holistic evaluation. The results help in realizing the role of machine learning in enhancing predictive accuracy and help in clinical decision systems. The given strategy highlights the role of the systematic evaluation, preprocessing of data, and analysis of features in order to increase the level of reliability in prediction.

The key findings of the research can be summarized in the following statements:

An all-encompassing machine learning diagnosis model in cardiovascular disease based on clinical data is derived.

There are several classification algorithms which are applied and the comparative analysis of these algorithms is carried out to find the most efficient predictive model.

Strong performance evaluation is done using a multi metric appraisal strategy that takes into account accuracy, precision, recall, F1 score, and other measures.

The paper can give clues on the feasibility of machine learning methods in assisting clinical decision making in cardiovascular care. It is experimentally shown that machine learning models are effective in enhancing early cardiovascular disease detection.

The study attracts attention to the future of machine learning based methods with potential to improve cardiovascular disease diagnosis and leads to the creation of intelligent healthcare prediction systems.

RELATED WORK

In recent years, machine learning methods have been extensively deployed to predict and diagnose cardiovascular diseases as they can analyze a variety of medical data and find unknown patterns related to the risk of getting a disease. The growing access to clinical datasets and computation facilities has prompted the researcher to analyze various classification algorithms to enhance prediction and aid clinical decision making [3]. A number of studies have shown that machine learning founded methods can be superior to conventional statistical approaches due to their ability to model nonlinear associations between patient features and risk factors. Subsequently, the cardiovascular disease prediction has emerged as a significant study in analytics in healthcare.

Cardiovascular disease prediction has been explored using a number of machine learning algorithms, including Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Naive Bayes, and K Nearest Neighbor [11]. The Logistic Regression is widely applied due to its interpretation and the capacity to provide the estimation of disease occurrence probability. The models based on Decision Tree offer clear rules on decision that can be easily understood during the clinical setting. At high dimensional feature space levels in the proper tuning Support Vector Machines perform well. Random Forest algorithms are a combination of several decision trees to enhance more accuracy on forecasting and less overfitting. In the same fashion, K Nearest Neighbor methods classify patients according to similarity indicators whereas Naive Bayes classifies them in a probabilistic way with significantly low computational complexity. The variety of algorithms used in different studies tells us that it is highly important to compare them to choose the most appropriate one regarding predicting cardiovascular cases. machine learning techniques have developed powerful potential for supporting clinical decision making, contributing to the improvement of clinical guidelines and management algorithms, and establishing clinical practices that are evidence-based for the management of cardiovascular diseases. [12, 13].

Some experiments have shown the successful outcomes of machine learning with cardiovascular data. Other Studies have also suggested using hybrid models, which incorporate various algorithms stacked up or by voting to improve predictive accuracy and performance. Indeterminately, hybrid machine learning models, such as Support Vector Machine, K Nearest Neighbor, and Extreme Gradient Boosting have shown better performance on classification rates than single-model performances.

Also, machine learning techniques have been compared with various datasets that were obtained in hospitals and in open repositories. High accuracy levels of above 90 percent have been reported in some studies that use neural networks, the random forest, and gradient boosting techniques. The chi square testing and feature importance analysis methods of feature selection have been applied to determine the most appropriate attributes that affect the prediction of a disease. Moreover, grid search and other hyperparameter optimization methods like cross validation have been used to improve the performance and reliability of the models. These methods argue in the relevance of data preprocess and parameter optimization in obtaining better diagnostic accuracy.

Predictive models on machine learning are able to enhance the quality of detection on cardiovascular diseases using comparative evaluation and feature engineering. A feature selection framework based on optimization that has been found to maximize the performance of the early prediction of cardiovascular disease and its computation efficiency State-of-the-art machine learning structures that are optimally trained make heart diseases diagnosis much more appropriate in clinical datasets.

In spite of the positive outcomes mentioned in the literature, the current research on cardiovascular disease predictions has multiple limitations. Applying relatively small datasets is one of the key problems that might restrict the generalization of the models and predispose them to overfitting. Most studies are based on few features without systematic selection methods of features, thus could lower the reliability of prediction. Furthermore, other studies are based on the idea of accuracy as the primary measure of evaluation and pay no attention to such critical performance indicators as precision, recall, and model robustness. It is also impossible to determine the most effective algorithm to use in clinical applications since its comparative analysis across several algorithms is not comprehensive. In addition to this, datasets and experimental conditions are varied making it difficult to come up with direct comparisons between the studies. The cardiovascular prediction models built through machine learning may be enhanced with superior diagnostic performance in case they are used along with effective selection of features and comparative evaluation of algorithms.

The other critical drawback is feasibility of proposed models in real world healthcare setting. The computational tools needed to run some of these models are involved, or they need characteristics that might not be easy to access in clinically common practice [1]. Moreover, interpretability is the issue that healthcare professionals must be concerned with since the reasoning of clinical decisions must be clear most of the times. As such, there is the necessity of research articles that test various machine learning algorithms with available clinical data and strike the right balance between accuracy and interpretability.

The literature-based identified research gap suggests the need to conduct a systematic comparative study of the machine learning methods of cardiovascular disease diagnosis based on clinical features. A more refined assessment model incorporating a variety of performance identifiers and standard-processing steps can be more confident when it comes to making several findings about the performance of a model. Also, to determine which algorithm is most appropriate in clinical implementation, it is necessary to consider the predictive performance, operational efficiency, and applicability level of the algorithm critically. The following gaps can help to develop strong decision support systems to detect early cardiovascular diseases [14].

A summary of some of the studies on predicting cardiovascular diseases using predictive machine learning approaches such as the algorithm, dataset, performance, and limitations are presented in Table I.

TABLE I.
COMPARISON OF EXISTING STUDIES

Ref.	Methods Used	Dataset	Accuracy (%)	Limitations
[15]	SVM, KNN, ANN, Random Forest with optimization	UCI dataset	98.54	Limited data diversity
[16]	Naive Bayes, SVM, Logistic Regression, AdaBoost	Cleveland dataset	86.6	Feature selection improvement needed
[17]	Random Forest, CART, Naive Bayes	Chinese dataset	AUC 0.787	Need external validation

[18]	Decision Tree, SVM, Logistic Regression	Hospital dataset	91	Limited dataset
[19]	Multiple ML models including ANN	UCI dataset	>93	Need larger datasets
[11]	Logistic Regression, Decision Tree, Random Forest	Cleveland dataset	92.10	Small dataset
[20]	Multiple classification models	UCI dataset	89.01	Need hybrid approaches

Altogether, the existing literature indicates the efficiency of machine learning algorithms in predicting cardiovascular diseases, but the issues of data volume, feature selection, the evaluation consistency, and the comparative analysis still exist. The current paper overcomes these limitations by introducing several machine learning models based on clinical data and experimenting on their performance using extensive metrics to establish the most trustworthy method to identify a cardiovascular disease diagnosis.

MATERIALS AND METHODS

The section presents the description of the dataset, preprocessing, feature selection method, and machine learning algorithms in prediction of the cardiovascular disease. The general treatment will adopt a definite pipeline procedure comprising of data gathering and preparation, training the model, and performance appraisal. The aim is to develop predictive models based on clinical characteristics potentially useful in cardiovascular disease early detection.

Dataset Description

The data in this research was retrieved in the publicly-available UCI Machine Learning Repository which is a collection of cardiovascular-related clinical records on diagnosis. This dataset is very relevant, accessible and well defined clinical attributes have made it be used a lot in medical data mining research. It is a set of patient data obtained in a variety of medical facilities, comprising demographic data, physiological measurements, and diagnostic signs of heart disease. The data is structured in the form of cases involving specific patients and each case has a number of features which depict the clinical conditions and a target variable which is the presence or absence of cardiovascular disease.

Clinical characteristics that were incorporated in the dataset are the variables that are measured regularly and are known to be related to cardiovascular risk evaluation. Such characteristics are the age and sex of the patient, the nature of chest pain, blood pressure when at rest, serum cholesterol levels, as well as the level of sugar in the blood when at rest. More so, the dataset will include the outcome of resting electrocardiographic tests, the highest heart rate during physical activity, and the presence or absence of angina-induced by exercise. The other significant variables are the ST depression value at exercise, slope of the peak exercise ST, the number of large vessels of the blood vessel found with the help of fluoroscopy, and the existence or deficiency of thalassemia. These attributes in combination offer useful physiological and diagnostic data that may impact the assessment of cardiovascular health. The variability of clinical presentations can enable machine learning models to acquire patterns and correlation between various risk factors and the ultimate disease outcome, which facilitates a better prediction and risk detection.

The variable under study is a cardiovascular disease diagnosis usually being made coded as a binary outcome varied upon the presence or absence of disease. In other versions of the data set, more than one or two classes can constitute the various severity levels, although when making predictions, the target is often reduced to a binary classification problem. This formulation allows machine learning algorithms to assign patients to risk groups and clinical decisions.

The relevance of the dataset to the clinical setting is that it reflects the actual measurements of the patients, which are grouped in the context of medical checkups. One can consider blood pressure, cholesterol, electrocardiographic, and other attributes as the standard risk factors of the cardiovascular disease, and they are essential in diagnosis. Accordingly, the predictive models created based on these features can be used in healthcare settings in practice. It is possible to use machine learning algorithms to find patterns related to the likelihood of diseases and help clinicians with early diagnosis and prevention by relying on the structured clinical data.

Table II is an overview of the data characteristics and their clinical relevance.

TABLE II. DATASET ATTRIBUTES AND DESCRIPTION

Attribute	**Description**	**Clinical Significance**
Age	Age of the patient (years)	Major non modifiable risk factor
Sex	Gender of the patient	Risk variation between males and females
Chest Pain Type	Type of chest pain experienced	Indicator of cardiac abnormalities
Resting Blood Pressure	Blood pressure at rest (mm Hg)	Hypertension risk assessment
Serum Cholesterol	Cholesterol level (mg/dl)	Indicator of arterial blockage risk
Fasting Blood Sugar	Blood sugar level after fasting	Diabetes related cardiovascular risk
Resting ECG	Electrocardiographic results	Detection of heart abnormalities
Maximum Heart Rate	Maximum heart rate achieved	Cardiac function assessment
Exercise Induced Angina	Presence of angina during exercise	Indicator of coronary artery disease
ST Depression	ST segment depression value	Myocardial ischemia indicator

Slope of ST Segment	Slope of peak exercise ST segment	Cardiac stress response
Number of Major Vessels	Number of vessels detected by fluoroscopy	Coronary artery blockage indicator
Thalassemia	Blood disorder status	Associated cardiac complications
Target Variable	Presence or absence of cardiovascular disease	Diagnostic outcome

The use of this dataset enables the development of predictive models that are clinically meaningful and applicable for real world cardiovascular disease diagnosis.

Data Preprocessing

Machine learning requires data preprocessing since raw clinical data may include inconsistencies, missing values, and differences across scales which may have a detrimental impact on model performance. Both quality and efficiency of the models training and reliability of the predictions are improved through proper preprocessing. This paper has used a number of preprocessing operations in order to ready the data to be processed using machine learning.

The initial step entailed missing values. Unreported measurements or lack of patient information often results in medical datasets having incomplete data about the patient. The identification of missing values was performed and considered with the help of the corresponding methods of their imputation to make sure that the data set was consistent and could be used to train a model. Data that contained too much missing information were scrutinized so as not to bias the data.

Encoding techniques were used to transform categorical data in the data set, including the type of chest pain, the results of the resting electrocardiographic, and the thalassemia status into numeric forms. The encoding is needed since the majority of machine learning algorithms need numerical input features. The number of categories was categorized using label encoding or one hot encoding techniques, respectively depending on the attributes of the categorical variables to maintain meaningful relationships between the categories, as well as computational efficiency.

Normalization was done to transform numerical variables to the same range. The clinical characteristics like blood pressure, cholesterol level, and heart rate can differ greatly in their magnitude, which potentially affects the learning behavior of the model. The scaling of features was used like min max normalization or standardization to make sure that all the features were equally utilized in training the model. This is especially necessary in sensitive algorithms like Support Vector Machine and K Nearest Neighbor.

To assess the models' performance, the data were subsequently divided into training and test sets. The majority of the time, some data was used for training and the remainder for testing. Because of this division, the models may learn the patterns in the training data and then be evaluated on unseen data to see how well they generalize. Additionally, more reliable performance estimations can be obtained by using cross validation techniques.

Class balancing came to play in case there was a possible imbalance in the disease and non disease cases. It is possible that imbalanced datasets will produce biased models that are biased towards the most common class. Such methods as resampling or artificial data generation are also possible when they are needed to provide equal distribution of classes. All these preprocessing measures enhance the performance and strength of the cardiovascular disease prediction machine learning models.

Feature Selection

The process of feature selection is significant in machine learning since not all the variables of the input used will be equally important in the prediction process. Unrelevant or duplicating features may raise the complexity of the model, decrease accuracy and cause overfitting. The importance of picking out the most informative clinical attributes is especially critical in medical datasets due to the bias to make models more interpretable and all predictions relying on physiologically meaningful indicators. As such, feature selection was used in the study to come up with the most relevant predictors related to cardiovascular disease.

The relationship between the input features and the target variable was evaluated using a correlation based and statistical analysis. The feature importance scores were computed in order to estimate the extent to which each attribute was useful in prediction of the disease. The low-relevant and high-redundancy attributes were scrutinized thoroughly and eliminated as needed so as to reduce noise in the dataset. The process contributed to the enhancement of the computational efficiency and enabled the machine learning models to prioritize the most crucial clinical indicators.

The characteristics that were chosen were largely the key risk factors of cardiovascular factors such as age, nature of chest pain, blood pressure at rest, serum cholesterol level, maximum heart rate reached, angina developed during exercising, the values of ST depression, the number of large vessels, and thalassemia status. The attributes have a clinical significance because they are physiological conditions that are directly linked to heart disease. An example is, the type of chest pain and ST depression which provide the information about cardiac stress and ischemia and, the cholesterol level and blood pressure which provide the information about the long-term risk of cardiovascular.

The predictive models were also able to learn much faster and more effectively and perform better in generalization by reducing the feature space to the most relevant variables. The interpretability of the results is also improved through feature selection, which enables clinicians to learn more about which risk factors are the most useful in predicting the disease.

Machine Learning Models

Various machine learning algorithms that were supervised were adopted in the research in order to test their capability in cardiovascular disease prediction using medical data. Such models were chosen because of populating medical prediction tasks, the possibility to work with structured datasets, and the dissimilarity in the way the learning processes are provided. By comparing several algorithms, it is possible to find the most appropriate model to be used in order to offer the correct diagnosis and decision support.

Logistic Regression is a classification tool that is generally applied to binary prediction. It approximates the chances of occurrence of a disease with the help of input features with the help of a logistic function. Its simplicity and interpretability are the reasons why the model is commonly applied in the healthcare context. Clinicians can gain insights into how the variables affecting prediction can be individual and how they contribute to the overall outcome.

Decision Tree is a rule based algorithm which group data or subdivides data into branches depending on the feature conditions. It builds a hierarchical shape, with every inner node indicating a decision regulation and every terminal node indicating an outcome of a classification. Decision trees are simple to interpret and give decisive decision paths hence suitable in clinical applications.

Random Forest improves prediction accuracy and helps reduce overfitting by combining many decision trees. Each tree is trained on a different random subset of the data, and the final output is decided through majority voting. This method is stable and can capture complex, nonlinear relationships between input features and the target variable.

Support Vector Machine (SVM) is a strong classification technique that finds the optimal boundary, or hyperplane, between classes in a high dimensional feature space. It works well when class boundaries are complex. Kernel functions allow the data to be mapped into higher dimensions, which often improves class separation.

K Nearest Neighbor (KNN) is an instance based method that classifies a sample by comparing it with the closest training samples. The predicted class depends on the majority label among its nearest neighbors in the feature space. Although simple, KNN can perform well when the data is properly scaled and clearly structured.

Naive Bayes is a probabilistic classifier based on Bayes' theorem, with the assumption that features are independent of each other. It is fast and efficient, especially for smaller datasets. Despite its simplicity, it often produces reliable results, including in medical prediction tasks.

Table III presents the main hyperparameters used for configuring the machine learning models in this study.

TABLE III
HYPERPARAMETERS OF MACHINE LEARNING MODELS

Model	**Hyperparameters**
Logistic Regression	Regularization type, Regularization strength (C), Solver
Decision Tree	Maximum depth, Minimum samples split, Criterion
Random Forest	Number of trees, Maximum depth, Minimum samples split
Support Vector Machine	Kernel type, Regularization parameter (C), Gamma
K Nearest Neighbor	Number of neighbors (K), Distance metric
Naive Bayes	Probability distribution parameters
Gradient Boosting (if used)	Number of estimators, Learning rate, Maximum depth

The implementation of several machine learning models will allow carrying out a thorough evaluation and comparison of the options and finding the most correct and valid strategy to diagnose cardiovascular disease.

Prediction Framework

The generalization prediction model produced during this study is based on a systematic pipeline that is supposed to work on clinical information and generate correct cardiovascular disease predictions. The first step is the gathering of the clinical data set, the second step is preprocessing, which makes the data ready to analysis. Preprocessing involves the processing of missing values, the encoding of categorical variables as well as normalizing the numerical features in a manner that is consistent across the dataset. These measures enhance the quality of data and render the input machine learning algorithms. Preprocessing is followed by feature selection step whereby most relevant variables related to cardiovascular disease risk are identified. The choice of significant features aids in enhancing the model efficiency, lowering the complexity, and aiding in understanding the findings better.

These chosen characteristics are then utilized to train several machine learning models, namely; Logistic Regression, Decision Tree, Random Forest, Support Vector machine, K Nearest Neighbor, and the Naive Bayes. All the models are trained on the training data to learn the trends of disease presence and then used to make decisions. In order to identify the most effective model, the models are compared with each other based on the performance measures of accuracy, precision, recall and F1 score among other useful performance measures. The end result of this framework is a predictive model, which may help clinicians to identify the patients who might be at risk of cardiovascular disease in regard to their clinical presentation, which can help to diagnose the patients at an early stage and ensure that they can be treated in time.

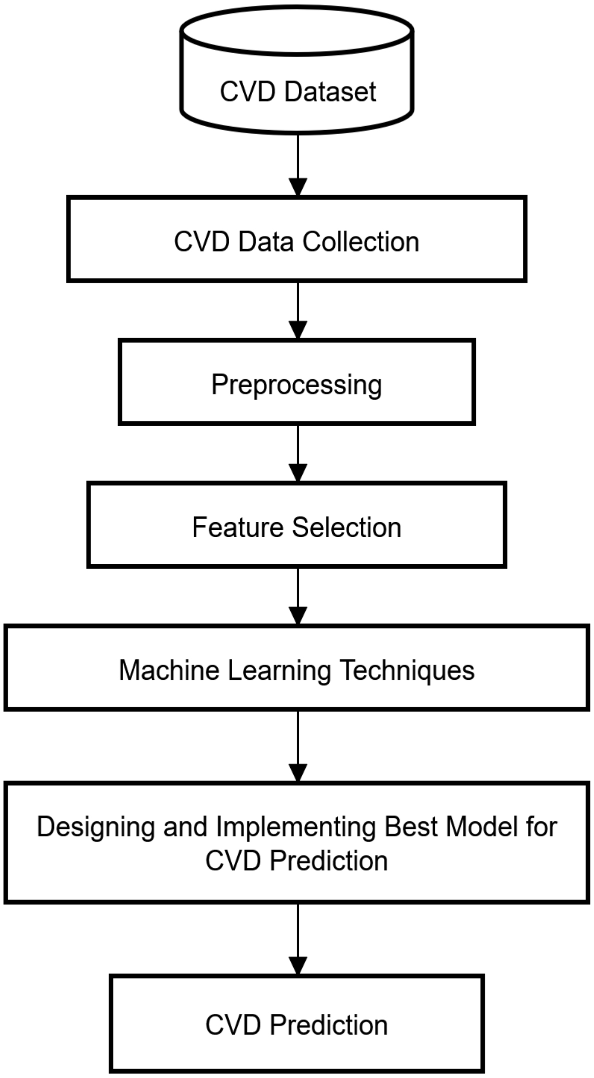
CVD Dataset
CVD Data Collection
Preprocessing
Feature Selection
Machine Learning Techniques
Designing and Implementing Best Model for CVD Prediction
CVD Prediction

FIGURE. 1
PROPOSED WORKFLOW FOR CARDIOVASCULAR DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES.

The general process flow of cardiovascular disease (CVD) prediction comprises several sequential steps, the first of which is data collection, and then data preprocessing and selection of features. The data obtained after the process is then utilized to train machine learning models in order to determine the most effective predictive method. The last step is the implementation of the optimized model to create the CVD prediction outcome. Such organization of the pipeline guarantees systematic treatment of data, better feature representation, and predictive consistency.

RESULTS AND DISCUSSION

This part contains experimental design, performance testing, and comparison of results as well as a discussion of the results achieved on the applied machine learning models. The aim is to evaluate the efficiency of the various algorithms in cardiovascular disease prediction, using clinical data, and to determine the most trustworthy method to be used to assist in diagnosis. There were several measures of performance that were employed in order to offer a complete analysis of the classification performance and the ability to generalize.

Experimental Setup

These experiments were done through a python machine learning environment. Preprocessing, model training, and evaluation were done with standard data science libraries, such as NumPy, Pandas, Scikit-learn, and Matplotlib. These programs offer effective computation help in implementing classification algorithms and performance metrics analysis. The workflow was based on an experimental loading, preprocessing, feature selection, model training, and performance evaluation of a single computational framework.

Experiments were carried out using a standard computing system having a multi core processor, adequate memory capacity as well as a current operating system environment. As the data set in this research is comparatively moderate in terms of its size, the processing power needed was not difficult, and could not demand specific computing hardware, including graphical processing units. This goes to show the feasibility of the proposed framework with real world healthcare settings where sophisticated computing facilities are not always accessible.

In order to assess the extent to which the models can be applied to new data, the dataset was split into training and testing in the form of a holdout validation. A fraction of the data was trained as models and the rest of the samples were held back so that we could use it to test the performance of the models on unknown cases. This division assists in giving a more realistic prediction estimate. Moreover, the cross validation methods can also be used to enhance the reliability and decrease variability of the evaluation outcomes by making sure that alternative data sets are tested during the model evaluation.

A number of evaluation metrics were used to measure classification performance in order to have a complete picture of modeling behavior. These measures were accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC ROC). Accuracy is the percentage of the correctly classified examples and precision and recall are measures of the correctness and sensitivity of the predictions. The F1 score offers a balanced measure to the extent that it brings the measure of precision and recall together. AUC ROC represents the capability of the model to discern between the classes under varying decision thresholds to give an idea of the overall classification performance.

The mathematical expressions of the main evaluation indicators are as follows:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1 Score = (Precision x Recall)/(Precision + Recall) x 2.

In which TP is true positives, TN is true negatives, FP is false positives and FN is false negatives. All these metrics can be used to offer a sound evaluation framework of cardiovascular disease prediction performance.

Performance Results

In order to have an in-depth comparison of the performance of the implemented machine learning models in terms of classification, several evaluation metrics were adopted in order to determine the effectiveness of the machine learning models. The algorithms that were taken into consideration in this paper were the Logistic Regression, Decision Tree, Random Forest, Support Vector machine, K Nearest Neighbor, and the Naive Bayes. All the models were trained on the preprocessed clinical data in a manner that they were able to learn the patterns of cardiovascular disease risk. The models were then tested on the testing subset after training to analyze the ability of the models to predict the presence or absence of cardiovascular disease in an unseen sample. This testing exercise enabled a fair and comparative comparison of the models in the whole range, thus making it feasible to determine variation in predictive ability and general performance. The evaluation also gave a better insight into the generalization power of the models in real-life clinical situations by testing them on different data that was not employed in the training process.

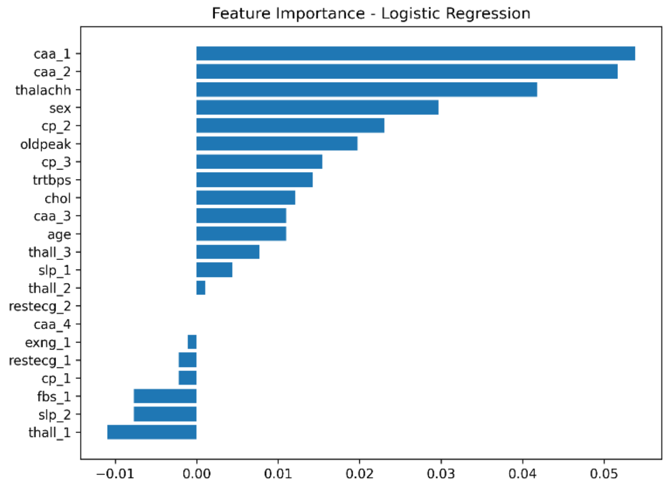

FIGURE 2
FEATURE COEFFICIENT ANALYSIS OF THE LOGISTIC REGRESSION MODEL SHOWING THE RELATIVE CONTRIBUTION OF CLINICAL ATTRIBUTES TO PREDICTION. SHOWING THE RELATIVE CONTRIBUTION OF CLINICAL ATTRIBUTES TO THE PREDICTION OUTCOME.

As illustrated in Figure 2, characteristics associated with indicators of cardiovascular condition are more important in the model as attested by higher scores. The fact that both positive and negative coefficients are observed also indicates the direction of the relationship between the values of features and the membership of the classes, which enhance the ability to interpret the study and contributes to the transparency of the method offered.

Table IV shows the comparative performance outcome of all the models in terms of accuracy, precision, recall and F1 score. The set of these evaluation metrics gives a proper idea of predictive reliability, the effectiveness of classification, and the resilience of the model as a whole. Accuracy is defined as total percentage of correctly predicted cases and precision is defined as percentage of correctly identified positive predictions out of all the predicted positive cases. Recall is the degree of accuracy that the model identifies the actual disease cases, which is an important issue especially in medical diagnosis. F1 score offers a balanced measure, which is an integration of precision and recall and it is a single measure that either indicates the correctness or sensitivity of the model predictions.

TABLE IV
MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1_Score
Logistic Regression	87.91%	87.95%	87.91%	87.92%
SVM	86.81%	87.19%	86.81%	86.83%
KNN	83.52%	83.56%	83.52%	83.53%
Decision Tree	78.02%	78.39%	78.02%	78.05%
Random Forest	83.52%	84.51%	83.52%	83.52%
Gradient Boosting	82.42%	83.16%	82.42%	82.43%
Naive Bayes	80.22%	80.95%	80.22%	80.23%

The Logistic Regression classifier performed best in most of the evaluation measures such as accuracy, F1 score, and AUC compared to other models. The Logistic Regression gave competitive results and the benefit of being interpreted, which is useful in clinical applications. Naive Bayes and Decision Tree demonstrated relatively poorer performance which can be explained by the sensitivity to variations of data and interactions between features.

Figure 3 represents the receiver operating characteristic (ROC) curves of various models at different decision thresholds, which show how various models classify. The values below the curve indicate that the Logistic Regression and Support Vector Machine offer high disease and non-disease classifications. The larger the value of AUC, the higher the level of classification and the higher the sensitivity to detecting diseases.

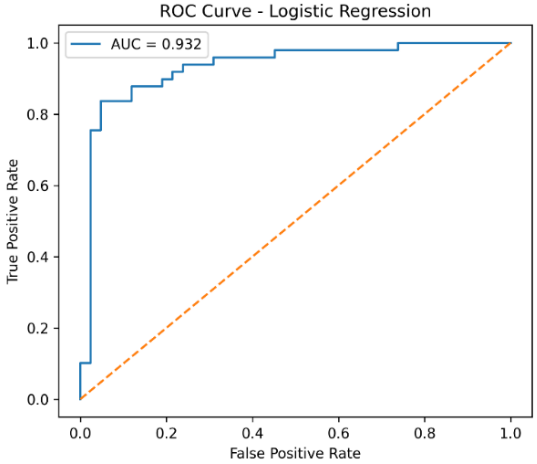

FIGURE 3
RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE OF THE LOGISTIC REGRESSION MODEL EVALUATED ON THE TEST DATASET. THE CURVE ILLUSTRATES THE TRADEOFF BETWEEN TRUE POSITIVE RATE AND FALSE POSITIVE RATE ACROSS DIFFERENT CLASSIFICATION THRESHOLDS.

As shown in Figure 3, the ROC curve is still far above the diagonal reference line of random classification which proves that the model yields effective separation of positive and negative samples. The AUC value of 0.932 is a high predictive power and the capability to classify similar results with varying threshold.

The confusion matrix in Figure 4 providing the results of the model with the best performance shows the allocation of correct and incorrect cases. The chart brings out the fact that the model can well interpret positive and negative cases, and that is critical in medical diagnoses to reduce false negatives and false positives. The correct diagnosis of the disease cases is especially relevant since failure to identify the disease in time can cause serious medical effects.

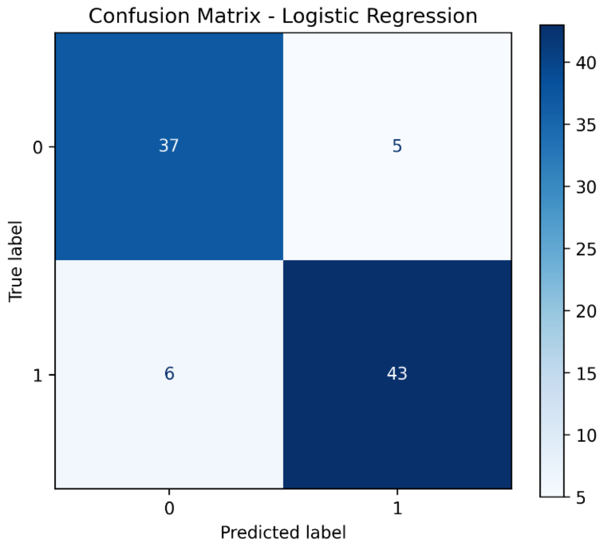

FIGURE 4
CONFUSION MATRIX OF THE LOGISTIC REGRESSION MODEL SHOWING CLASSIFICATION PERFORMANCE ON THE TEST DATASET. THE MATRIX ILLUSTRATES THE DISTRIBUTION OF CORRECTLY AND INCORRECTLY CLASSIFIED SAMPLES ACROSS THE TWO CLASSES.

According to Figure 4, the confusion matrix illustrates that 37/43 samples were classified correctly in classes 0 and 1, respectively, and only a few cases were obtained that were misclassified. The high concentration of values on the diagonal is another indicator that the model is reliably predictive and equally balanced in terms of its classification ability on both sides of the classes.

Table V presents the performance of the best model in terms of classification.

TABLE V.
CLASSIFICATION PERFORMANCE OF THE BEST MODEL

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Class 0	0.8605	0.8810	0.8706	42
Class 1	0.8958	0.8776	0.8866	49
Accuracy	0.8791	0.8791	0.8791	91
Macro Avg	0.8781	0.8793	0.8786	91
Weighted Avg	0.8795	0.8791	0.8792	91

Table V summarizes the classification of the best model. The model had a total accuracy of 87.91 and the accuracy and recall had balanced values of both classes. Class 0 and Class 1 have an F1-score of 0.8706 and 0.8866, respectively, which means that there is no significant distinction in the predictive performance of classes.

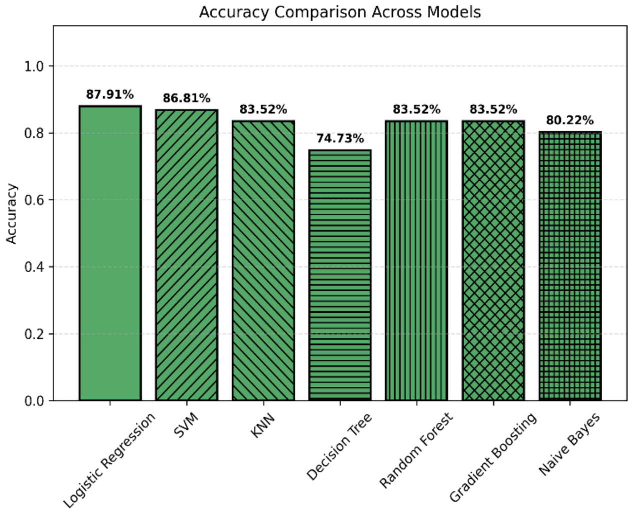

FIGURE 5
ACCURACY COMPARISON OF DIFFERENT MACHINE LEARNING MODELS ON THE TEST DATASET. LOGISTIC REGRESSION ACHIEVED THE HIGHEST CLASSIFICATION ACCURACY AMONG THE EVALUATED MODELS.

Various machine learning algorithms were compared to determine their predictive capability as shown in Figure 5. Logistic Regression was the most accurate of the models evaluated with 87.9 percent accuracy and the second best with 86.8 percent accuracy was Support Vector Machine. Figure 6 shows the comparative performance in terms of F1-score of the tested machine learning models. Logistic Regression had the greatest F1-score, which means that it has an equal level of precision and recall as compared to other classification algorithms.

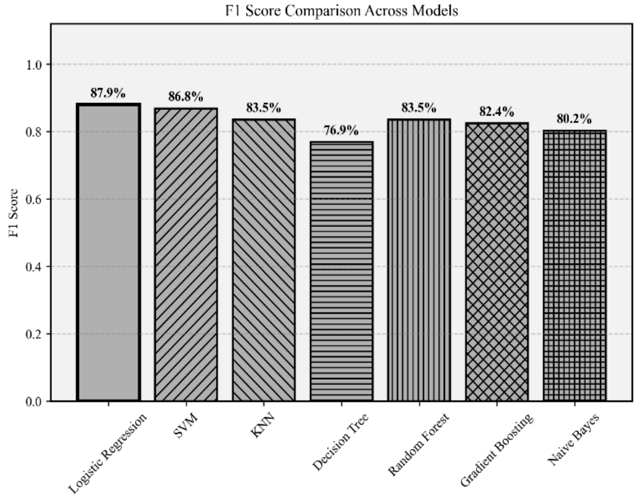

FIGURE 6
F1-SCORE COMPARISON OF DIFFERENT MACHINE LEARNING MODELS EVALUATED ON THE TEST DATASET. LOGISTIC REGRESSION ACHIEVED THE HIGHEST F1-SCORE AMONG THE COMPARED METHODS.

The accuracy of the considered machine learning models is presented in Figure 7. The highest precision model was that of Logistic Regression that showed that it has a low false positive rate as opposed to the other models of classification.

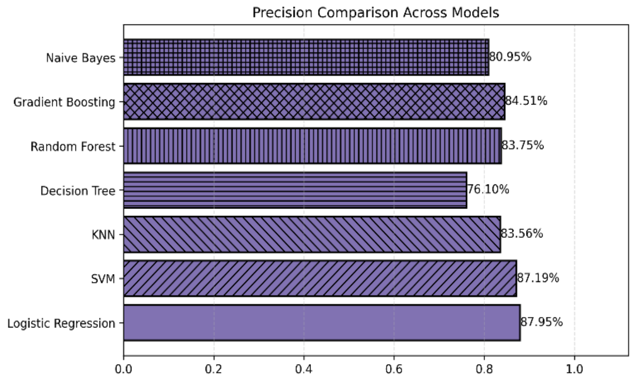

FIGURE 7
PRECISION COMPARISON OF DIFFERENT MACHINE LEARNING MODELS ON THE TEST DATASET. LOGISTIC REGRESSION ACHIEVED THE HIGHEST PRECISION AMONG THE EVALUATED METHODS.

The recall of the considered machine learning models is provided in Figure 8. Logistic Regression proved to have the highest recall which represents that it is good at identifying the positive cases correctly as opposed to the other classification algorithms.

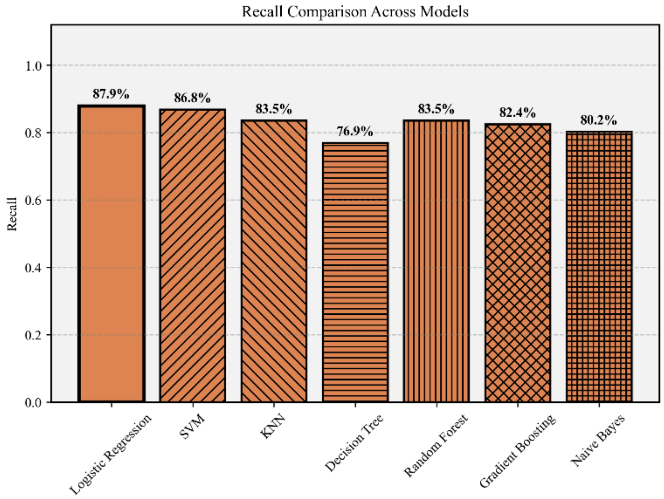

FIGURE 8
RECALL COMPARISON OF DIFFERENT MACHINE LEARNING MODELS ON THE TEST DATASET. LOGISTIC REGRESSION ACHIEVED THE HIGHEST RECALL AMONG THE EVALUATED METHODS.

All in all, experimental findings validate the claim that machine learning algorithms can be applied to predict cardiovascular disease based on clinical evidence. It is also pointed out in the comparative evaluation that proper algorithm and preprocessing policies should be chosen to ensure the best performance in prediction.

Comparative Literature Analysis.

There was a comparative analysis to assess the efficacy of the proposed machine learning structure in comparison to the reported studies concerning cardiovascular disease prediction. The literature has also used the different machine learning algorithms to include the Logistic Regression, Decision Tree, Support Vector Machine and the random forests using clinical data including the UCI heart disease dataset and data collected in hospitals. The performance of these studies is reported with great variability based on the size of dataset, preprocessing techniques, feature selection tactics and evaluation policies. Although some studies have obtained high accuracy rates, most of them work with small datasets or concentrate on one particular measure of performance, which does not allow a full evaluation of model performance.

The outcomes that are achieved in this study indicate that there is a competitive performance in comparison to previous studies. Logistic Regression model demonstrated the best classification accuracy, as well as good F1 score, MCC, and AUC, which shows that it is a reliable predictor. Systematic preprocessing and feature selection along with multi metric evaluation have helped to increase the model robustness and generalization. In comparison with those studies based only on accuracy, other metrics of evaluation, like the MCC and the AUC, give a more detailed perspective on the classification performance, especially in a medical decision making situation, where balanced prediction is crucial.

Table VI provides the comparative outcomes of the results with the chosen existing approaches that are found in the literature. The comparison outlines the variance in algorithms, datasets and results of performance.

TABLE VI
COMPARISON WITH EXISTING STUDIES

Study	Methods Used	Dataset	Accuracy (%)	Remarks
[16]	Naive Bayes, SVM, Logistic Regression	Cleveland dataset	86.6	Feature optimization required
Proposed Study	**Multiple ML models with feature selection**	**UCI dataset**	**87.91%**	**Multi metric evaluation**

Some studies have higher values of accuracy, but most of them are achieved with smaller datasets or specific preprocessing techniques which do not have great prospects of application in real life situations. By contrast, the current research concentrates on a more objective assessment in terms of several algorithms and performance measures, which is a more credible evaluation of predictive ability. The positive gains of this research may be explained by the fact that all preprocessing has been systematized, features have been selected, and the models have been comparatively evaluated.

Also, the suggested framework focuses on the practical relevance as it exploits the affordability of clinical qualities (as opposed to more complicated or costly medical information sources). This practice will enhance the possibilities of implementation in less-resource intensive healthcare settings. On balance, the comparative analysis illustrates that machine learning methods, currently utilized and assessed in the correct way, can deliver the accurate and reliable cardiovascular disease prediction to help with the clinical decision making and early diagnosis.

Discussion

The results of the research indicate that machine learning algorithm could be successfully used to identify cardiovascular disease based on structured clinical data. The high performance especially Logistic Regression, shows that the linear decision boundary developed by the Logistic Regression gives quality classification and variance in prediction is minimized. Clinically, the successful forecasting of the risk of cardiovascular diseases using patient characteristics that are gathered on a regular basis can be helpful in the early detection and preventive management. Such features as the type of chest pain, blood pressure, cholesterol, and electrocardiography indicators are clinically significant and commonly used in practice. Thus, predictive models developed on the basis of these qualities can help medical personnel to detect high risk patients early enough, to intervene promptly and have better treatment results. rowing role of artificial intelligence in cardiology and highlighted how machine learning techniques can assist in early detection and diagnosis of cardiovascular diseases. Their study emphasized that data driven models using clinical attributes can improve diagnostic accuracy and support clinical decision making. The authors also noted that machine learning based prediction systems have the potential to enhance risk assessment and enable more personalized cardiovascular care [21].

The main strength of this paper lies in the fact that multiple machine learning algorithms have been compared in terms of various performance measures in a relatively comprehensive manner. Most of the past studies have largely depended on the accuracy, which is not necessarily the complete scope of classification reliability, particularly in the field of medicine where false negatives and false positives can be of paramount importance. The precision, recall, F1 score, and AUC are included to offer a more fair measure of the model performance. The systematic preprocessing and feature selection is another strength because it increases the data quality and makes the model learning more efficient. The framework also illustrates that there is no need to use complicated and computationally intensive models to achieve an effective prediction and, therefore, has been accessible to practical healthcare implementation.

Although these are the strengths, some limitations should be taken into account. The data that was employed in this research, despite its popularity, has quite a moderate number of samples, which can be a disadvantage in terms of model extrapolation to various populations. The clinical datasets used in various sites or health care systems may have differences in features of the patients and this may affect prediction of the performance. Besides this, machine learning models rely on the quality and completeness of input data, and missing or inaccurate clinical measurements may influence reliability. A second weakness is associated with the interpretability of machine learning models, but simpler models like Logistic Regression or Decision Trees are better interpretable. Enhancement of interpretability is also relevant towards earning trust among medical practitioners.

Practically speaking, the proposed framework has a great potential of integration in clinical decision support systems. As the models are based on the routinely gathered clinical characteristics, they can be introduced without any extra diagnostic tools or sophisticated infrastructure. The computation needs are not too high, which allows it to be introduced into healthcare areas with few resources. The systems may help doctors to conduct initial risk assessment, lessen the diagnostic time and enhance the accuracy of decisions. But in the real world, the implementation would need validation with larger and broader datasets, and the cooperation between the data scientists and the medical professionals to make it more reliable and ethically acceptable.

On the whole, the research confirms that machine learning-based predictive models may be useful instruments in the diagnosis of cardiovascular disease to facilitate early diagnosis and make better decisions in healthcare with the help of clinical knowledge.

CONCLUSION AND FUTURE WORK

This paper examined the performance of machine learning methods in diagnosing cardiovascular diseases in the structured clinical data. The deployment and testing of several classification algorithms were done in terms of their predictive capability in the identification of patients at risk of cardiovascular disease. The results of the experiments confirm that machine learning systems can provide the output of the existence of relationships among clinical characteristics and disease outcomes, as well as positive results in prediction. Preprocessing and feature selection were systematic and this was the reason why the model efficiency and accuracy was better in classification. The multifaceted evaluation measures also made it possible to conduct the general evaluation of the model performance not only in the aspect of accuracy but also in the aspect of balanced assessment of the reliability of prediction.

The best overall performance of all the evaluated algorithms is the Logistic Regression model that showed high performance in most of the measures, such as accuracy, F1 score, and AUC. The Support Vector machine along with the Logistic regression also proved to be competitive meaning that various approaches to machine learning would also be able to give good predictions when optimized adequately. The results indicate the significance of comparative analysis as the most appropriate algorithm in the cardiovascular disease prediction tasks.

The innovations of the research are the creation of a machine learning-based prediction system with the use of clinical features, the comparative analysis of the variety of classification models, and the implementation of a multi metric performance evaluation mechanism. This study also illustrates the usefulness of machine learning models in clinical decision making support without necessarily using sophisticated or costly diagnostic instruments. The proposed approach will offer greater chances of implementation at the real world healthcare settings since it targets patient measurements that are readily available.

The positive outcomes notwithstanding, there are some limitations to take into consideration. The sample data in this study is of moderate size and not necessarily representative of diverse groups of patients in different regions. When used on more heterogeneous or larger datasets, model performance can change.

There are different directions to be applied to the future research to enhance the prediction of cardiovascular diseases. For the generalization and reliability of the model, the application of larger multi center datasets would be helpful. It is possible to explore other optimization and feature selection procedures to achieve better predictive performance. In addition, it will be needed to produce decipherable machine learning models and available clinical decision support systems to be practically implemented in clinical settings. These advances would contribute to the early diagnosis of cardiovascular diseases and improved outcomes in the patients.

REFERENCES

[1] E. W. Steyerberg, *Clinical prediction models: A practical approach to development, validation, and updating*. Springer, 2019.

[2] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, and T. Kitai, "Artificial intelligence in precision cardiovascular medicine," *Journal of the American College of Cardiology,* vol. 69, no. 21, pp. 2657-2664, 2017.

[3] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine,* vol. 380, no. 14, pp. 1347-1358, 2019.

[4] A. Javeed, S. S. Rizvi, S. Zhou, R. Riaz, S. U. Khan, and S. J. Kwon, "Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification," *Mobile Information Systems,* vol. 2020, no. 1, p. 8843115, 2020.

[5] R. H. Eckel *et al.*, "2013 AHA/ACC guideline on lifestyle management to reduce cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines," *Journal of the American college of cardiology,* vol. 63, no. 25 Part B, pp. 2960-2984, 2014.

[6] P. M. Kumar, S. Lokesh, R. Varatharajan, G. C. Babu, and P. Parthasarathy, "Cloud and IoT based disease prediction and diagnosis system for healthcare using Fuzzy neural classifier," *Future Generation Computer Systems,* vol. 86, pp. 527-534, 2018.

[7] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE access,* vol. 7, pp. 81542-81554, 2019.

[8] H. A. Esfahani and M. Ghazanfari, "Cardiovascular disease detection using a new ensemble classifier," in *2017 IEEE 4th international conference on knowledge-based engineering and innovation (KBEI)*, 2017, pp. 1011-1014: IEEE.

[9] M. Gandhi and S. Singh, "Cardiovascular disease detection using a new ensemble classifier," in *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), Greater Noida, India*, 2015, pp. 520-525.

[10] C. Krittanawong *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Scientific reports,* vol. 10, no. 1, p. 16057, 2020.

[11] M. A. Hossen *et al.*, "Supervised machine learning ‐ based cardiovascular disease analysis and prediction," *Mathematical Problems in Engineering,* vol. 2021, no. 1, p. 1792201, 2021.

[12] Z. Malki, E. Atlam, G. Dagnew, A. R. Alzighaibi, E. Ghada, and I. Gad, "Bidirectional residual LSTM-based human activity recognition," *Computer and Information Science,* vol. 13, no. 3, p. 40, 2020.

[13] M. Farsi *et al.*, "Parallel genetic algorithms for optimizing the SARIMA model for better forecasting of the NCDC weather data," *Alexandria Engineering Journal,* vol. 60, no. 1, pp. 1299-1316, 2021.

[14] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine,* vol. 25, no. 1, pp. 44-56, 2019.

[15] M. S. Nawaz, B. Shoaib, and M. A. Ashraf, "Intelligent cardiovascular disease prediction empowered with gradient descent optimization," *Heliyon,* vol. 7, no. 5, 2021.

[16] P. Rani, R. Kumar, N. M. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *Journal of Reliable Intelligent Environments,* vol. 7, no. 3, pp. 263-275, 2021.

[17] L. Yang *et al.*, "Study of cardiovascular disease prediction model based on random forest in eastern China," *Scientific reports,* vol. 10, no. 1, p. 5245, 2020.

[18] M. N. R. Chowdhury, E. Ahmed, M. A. D. Siddik, and A. U. Zaman, "Heart disease prognosis using machine learning classification techniques," in *2021 6th International Conference for Convergence in Technology (I2CT),* 2021, pp. 1-6: IEEE.

[19] A. Akella and S. Akella, "Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution," *Future science OA,* vol. 7, no. 6, p. FSO698, 2021.

[20] S. Ouf and A. ElSeddawy, "A proposed paradigm for intelligent heart disease prediction system using data mining techniques," *Journal of Southwest Jiaotong University,* vol. 56, no. 4, pp. 220-240, 2021.

[21] T. Nakamura and T. J. J. o. C. Sasano, "Artificial intelligence and cardiology: Current status and perspective," vol. 79, no. 3, pp. 326-333, 2022.

AUTHOR INFORMATION

MUHAMMAD IQBAL has received the MSc (Information Technology) degree from University of Sargodha, Pakistan and MS (Computer Science) from Gomal University D I KHAN,KPK, Pakistan. He is currently pursuing the PhD degree in Computer Science from Gomal University D I KHAN, KPK, Pakistan. His Research interest include Software Engineering, Artificial Intellegence, Machine Learning, Deep Learning and IOTs. He is currently working in School Education Department Punjab.

MUHAMMAD ZUBAIR ASGHAR is currently an Assistant Professor with the Institute of Computing and Information Technology (ICIT), Gomal University, Dera Ismail Khan, Pakistan, and approved Ph.D. supervisor recognized by Higher Education Commission (HEC), Pakistan. His Ph.D. research interests include game-ai, first-person shooter games, third-person shooter games, computational intelligence, computational linguistics, machine learning, text mining, opinion mining, sentiment analysis, and big data solutions for social networks. He has published more than 40 publications in journals of international reputation (JCR and ISI indexed). He has more than 15 years of University teaching and laboratory experience in artificial intelligence and intelligent systems design. He is a Guest Editor of special issues in the journal of Social Computing in Health Informatics and Business Intelligence. He is also a Reviewer of many impact factor journals. He is an Associate Editor of IEEE ACCESS and Plos One.

SAHAR BATOOL has received the MSc (Information Technology) degree from University of Sargodha, Pakistan and MS/MPhil (Computer Science) from Institute of Southern Punjab, Multan, Pakistan. Her Research interest include Machine Learning and Deep Learning. He is currently working in School Education Department Punjab.