Implementing Machine Learning to predict the 10-year risk of Cardiovascular Disease

Simranjeet Singh Dahia
School of Computer Science
University of Adelaide, Australia
simran.dahia0084@gmail.com

Claudia Szabo
School of Computer Science
University of Adelaide, Australia
claudia.szabo@adelaide.edu.au

ABSTRACT

Cardiovascular disease (CVD) is the leading cause of death globally, demanding accurate risk prediction models for early intervention and prevention. This project aimed to develop a Machine Learning (ML) model for predicting the 10-year risk of CVD. A comprehensive review of existing literature was conducted, discussing the methods, algorithms, and data sources used in different studies, to evaluate the performance of various models. The review highlighted the potential of ML for improving CVD risk assessment, and the challenges and limitations of current research.

The UCI Heart dataset served as the training data for various ML models, including Logistic Regression (LR), Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and K-Nearest Neighbors (KNN). To optimize model performance, Cross Validation (CV), normalization techniques, and hyperparameter tuning were employed. We report the results, comparing them with traditional models.

The implications of this research extend to improved preventive strategies and interventions, potentially alleviating the burden of CVD on individuals and healthcare systems by more targeted interventions, and the optimization of healthcare resources.

KEYWORDS

Machine Learning (ML), Predictive analysis, Cardiovascular disease (CVD)

1 INTRODUCTION

Cardiovascular disease (CVD) is the leading cause of death globally, accounting for 17.9 million deaths in 2019 [1]. Traditional CVD risk prediction models, such as the American College of Cardiology/American Heart Association (ACC/AHA) risk model [2], Reynolds Risk Score (RRS) [3] and the Framingham Risk Score [4], have shown limited accuracy in predicting long-term CVD risk [5]. Though these have become increasingly common in clinical practice over the last decades, they tend to be limited in scope, lack personalization, and lack transparency, due to the complex interplay of various factors.

More recently, several studies have shown the potential of Machine learning (ML) algorithms in improving CVD risk prediction with more robust, more precise, and personalized prediction models [5, 6, 7]. These models have exhibited superior validation performance compared to traditional models by incorporating a wider range of risk factors and their interactions [5]. ML-based models can help identify high-risk individuals and motivate them to change their behaviors for preventive medicine purposes [6].

In this study, we aimed to develop an ML-based model to predict 10-year CVD risk. To begin with, a comprehensive review of the existing literature on traditional tools and ML models for CVD risk prediction was conducted, examining the methodologies, algorithms, and datasets used in previous studies to evaluate the performance of different models. The review highlighted the potential of ML approaches and gaps in the existing research.

Our study contributes to the growing body of research on ML-based CVD risk prediction and has the potential to improve clinical decision-making and patient outcomes [8].

2 RELATED WORKS

In this section, we explore existing research on CVD risk prediction, including traditional models and the emerging use of ML techniques.

2.1 CVD risk prediction using traditional models

Traditional models for CVD risk prediction have been developed over the last decades, including the FRS [4], RRS [3], QRISK [9], ASCVD 2013 Risk Calculator

https://doi.org/10.32388/1SVUCI

from AHA/ACC Pooled Cohort Equations [2], and ASSIGN Score [10].

These models are widely used in clinics, but according to a scoping review by Mohammed Abd ElFattah, et al., there are limitations such as limited factors, overestimation or underestimation, and inapplicability to different ethnicities [11]. The FRS, developed predominantly for the white US population, has been found to have limited accuracy in predicting CVD risk in non-white populations, such as African Americans and Hispanics, highlighting the need for more accurate risk assessment tools [12, 13, 14]. A systematic review assessing the use and validity of prediction models to estimate the risk of CVD in Latin America and among Hispanic populations in the United States found that FRS overestimated CVD risk for Hispanics [13]. The QRISK, a CVD risk predictive model developed over UK population data, tends to overestimate CVD risk, particularly in older patients and those with a history of CVD [15, 16]. However, a prospective open cohort study found that QRISK is better calibrated to the UK population than Framingham and has better discrimination [15]. The AHA/ACC Pooled Cohort Equations [2] and the ASSIGN Score [10] were developed using specific populations, such as the US and Scottish populations, respectively, which limits their accuracy in other populations [11, 17]. Another study comparing AHA/ACC Pooled Cohort Equations [2] estimated 10-year CVD risk in Black versus White individuals with identical risk factor profiles using pooled cohort equations found differences in estimates [18]. A systematic review assessing the use and validity of prediction models to estimate the risk of CVD in Latin America and among Hispanic populations in the United States found that the FRS overestimated CVD risk for Hispanics with an AUC of 0.69 [19]. The RRS [3], derived in a cohort of 25,000 healthy US women, has limitations, such as being derived from a cohort of only healthy US women [20], which may limit its generalizability to other populations [21]. Additionally, the RRS has been shown to overestimate the risk of CVD in some populations, such as African Americans [22].

A systematic review of 13 CVD risk prediction models highlighted the limitations of traditional models and emphasized the need for personalized risk assessment and inclusion of additional factors, such as genetic and environmental factors, to improve individual risk prediction [23]. Alternative approaches, such as using ML methods, have been proposed to overcome these challenges by incorporating a wide range of data, handling missing data, and identifying novel predictors [8, 24]. Another study compared well-established risk prediction algorithms based on conventional CVD risk factors with a ML based prediction model focused on CVD events [25]. The performance metrics of these tools can be found in Table 1.

Tool	AUC-ROC
FRS	0.6-0.8
RRS	0.80
AHA/ACC Pooled Cohort Equations	0.72
ASSIGN Score	0.75
QRISK	0.70-0.85

TABLE 1. AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE (AUC-ROC) FOR TRADITIONAL MODELS. FRS; FRAMINGHAM RISK SCORE, RRS; REYNOLDS RISK SCORE [30]

In conclusion, traditional models for CVD risk prediction have limitations, including limited factors, overestimation or underestimation, and inapplicability to different ethnicities. Alternative approaches, such as using ML methods, have been proposed to overcome these challenges.

2.2 CVD risk prediction using ML

In the last decade, several studies have explored the use of ML models for predicting CVD risk, offering potential improvements over traditional models [24, 25, 26]. These ML models leverage large datasets from sources such as Electronic Medical Records (EMRs), clinical databases, and population-based surveys. Various ML algorithms, including Logistic Regression (LR), Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), K-Nearest Neighbor (KNN), AdaBoost (AdB), XGBoost (XGB), and Neural Networks, have been employed [24, 25, 27].

In terms of performance, different studies have reported varying results, with the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) values ranging from 0.7 to 0.9 [25, 26, 27]. The most influential predictors identified by these models include age, sex, blood pressure, cholesterol levels, smoking status, and diabetes status [24, 25, 26, 27]. Additionally, some studies have incorporated novel predictors such as genetic data and socioeconomic

status, enabling the development of personalized and more accurate predictive models compared to traditional approaches [24]. For instance, a study by Ward, et al. found that the development of improved risk prediction may require richer data or incorporation of novel variables, such as genetic information across different racial/ethnic groups [28]. The study by Alaa, et al. developed an ML-based model called AutoPrognosis, which incorporated 473 available variables, including genetic data, to predict CVD risk [24]. These studies demonstrate the potential benefits of incorporating novel predictors in ML-based models for CVD risk prediction.

A comprehensive review by Subramani, et al. focused on ML and Deep Learning (DL) models for CVD risk prediction, encompassing the use of Electronic Health Records (EHRs) and imaging data. This review demonstrated that ML and DL models have the potential to enhance CVD risk prediction accuracy by incorporating a broad range of risk factors, including genetic, lifestyle, and environmental factors [29]. These models can also analyze complex relationships between risk factors to provide more personalized risk estimates. However, their complexity may limit interpretability, and challenges related to data availability and standardization exist.

Several notable studies have showcased the effectiveness of ML models for CVD risk prediction, with their results in Table 2.

Although these studies highlight the potential benefits of ML in CVD risk prediction, they also have certain limitations. Some studies were limited to specific populations or datasets, limiting the generalizability of their findings [25, 26, 29]. Additionally, the interpretability of some ML models, particularly DL models, may pose challenges, which can hinder their clinical utility. The accuracy and usefulness of ML models are highly dependent on the quality and completeness of the input data, and their performance may not generalize well to different populations or settings [24, 26, 29].

Despite these limitations, ML models hold promise for improving CVD risk prediction by incorporating a wide array of risk factors and analyzing complex interactions among them. Continued research and advancements in ML techniques may lead to more accurate and clinically useful predictive models for CVD risk assessment. In our study, we attempt to address these limitations.

Name	Accuracy (%)	Precision	Recall	F1 Score	ROC-AUC
LR	85.25	0.842	0.871	0.85	0.916
DT	80.32	0.812	0.828	0.814	0.864
SVM	90.2	0.906	0.906	0.9	0.916
RF	87	0.875	0.848	0.87	0.916
NBC	89.2	0.876	0.904	0.893	0.917
KNN	90.2	0.906	0.906	0.9	0.912
AdB	90.2	0.938	0.882	0.901	0.912
XGB	89	0.875	0.875	0.88	0.912
SGD	89	0.906	0.879	0.9	0.919
QDA	84.43	0.753	0.925	0.872	0.883
EVCH	92	0.906	0.936	0.921	0.927
EVCS	92	0.906	0.936	0.92	0.927

TABLE 2. PERFORMANCE METRICS OF DIFFERENT ML ALGORITHMS USING VARIOUS OPTIMIZATION METHODS (DHP, GSCV, RSCV, SEE APPENDIX 1)

3 METHODOLOGY

We used Python 3.9.13 on Anaconda Jupyter Notebook for the study, following the Project Plan given in Fig 1.

3.1 Data

The dataset plays the most significant role in ML. In this study, we used the UCI Heart dataset for training and testing of our models. The UCI Heart Dataset contains information on individuals with and without heart disease, consisting of 303 instances and 76 attributes [31]. This dataset has been widely used for classification tasks, such as predicting whether a person has heart disease or not, using ML techniques [27, 32]. The dataset was created by the Hungarian Institute of Cardiology and is available on the UCI Machine Learning Repository [31].

3.2 Feature selection and Data Engineering

For feature selection and engineering, we examined all 76 variables in the dataset. However, we determined that only 14 variables were both relevant and complete for our analysis, including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, and ST depression induced by exercise relative to rest.

The target variable represents the presence of heart disease, defined by an integer value ranging from 0 to 4. It indicates different levels of heart disease presence—class 0 represents the absence of heart disease, while classes 1, 2, 3, and 4 represent varying degrees of heart disease presence. However, we transformed the target variable into a binary form to simplify the target variable into a two-class classification problem. In this binary transformation, class 0 is labelled as 0 to represent the absence of heart disease. Conversely, classes 1, 2, 3, and 4 are labelled as 1, to indicate the presence of heart disease.

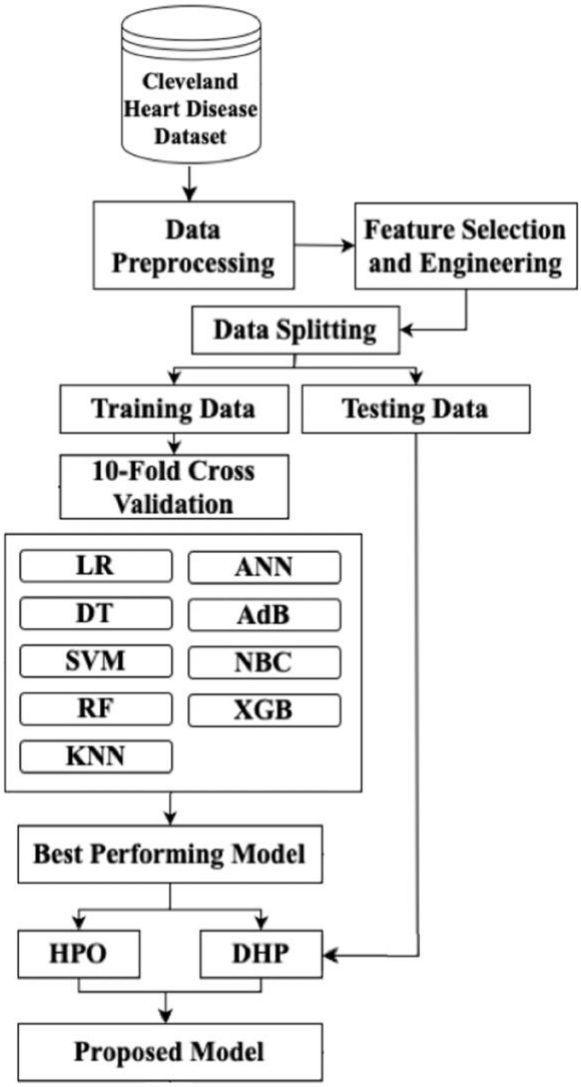

FIG 1. PROJECT WORKFLOW. LR, LOGISTIC REGRESSION; DT, DECISION TREES; RF, RANDOM FORESTS; SVM, SUPPORT VECTOR MACHINES; ANN, ARTIFICIAL NEURAL NETWORKS; ADB, ADABOOST; NBC, NAÏVE BAYES CLASSIFIER; XGB, XGBOOST; HPO, HYPERPARAMETER OPTIMIZATION; DHP, DEFAULT HYPERPARAMATER

To assess the relationships and significance among these variables, we constructed a Correlation Matrix using Pandas library [33], which can be found in Fig 2. The correlation matrix allowed us to examine the pairwise correlations between the selected variables. By evaluating the correlation coefficients, we assessed the strength and direction of the relationship.

We considered variables with high positive or negative correlations (correlation>0.40 or correlation<-0.40) as potentially influential predictors of CVD risk. Variables that exhibited a strong correlation with the outcome variable (CVD risk) were deemed essential for our predictive model. Additionally, we took note of any variables that showed minimal or no correlation with the outcome, as these were deemed less informative and were not included in the subsequent analysis. The highly correlated variables have been summarized in Table 3.

It is worth noting that the feature selection process may be subjective to some extent and dependent on the specific dataset used. Our selection was based on the UCI dataset, and it is possible that different datasets or different expert opinions may yield slightly different sets of relevant variables.

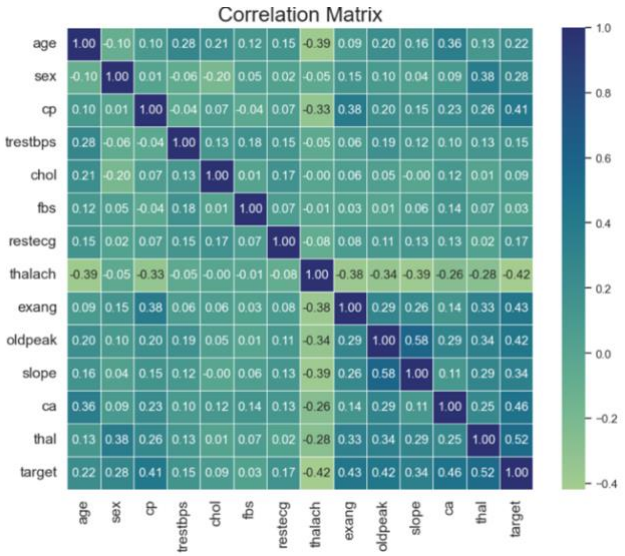

FIG 2. CORRELATION HEATMAP

According to the correlation heat map chest pain (cp), thalach, slope, exang, ca, and thal were highly correlated with the target variable. Table I denotes a brief description of these highly correlated features.

We then applied the ExtraTreesClassifier from Scikit-learn to the dataset to reveal the features with the highest importance; the top four features were: thalach,

ca, thal, and oldpeak. This analysis involved constructing an ensemble of DTs, each trained on random subsets of the data and features. By examining the variability in the trees' predictions, the algorithm determined the relevance of each feature in contributing to the classification task [34]. The resulting feature importance values provide insights into the crucial factors influencing the prediction of CVD risk. The results can be seen in Fig 3.

Attribute	Type	Description
Chest Pain (cp)	Discrete	Chest Pain type: a. Typical Angina b. Atypical Angina c. Non-anginal Pain d. Asymptomatic
thalach	Continuous	Maximum Heart Rate Achieved
slope	Discrete	slope: the slope of the peak exercise ST segment: a. upsloping b. flat c. downsloping
exang	Discrete	exercise induced angina: 1 = yes 0 = no
ca	Discrete	number of major vessels (0-3) colored by flourosopy
thal	Discrete	thal: 3 = normal 6 = fixed defect 7 = reversable defect
oldpeak	Continuous	ST depression induced by exercise relative to rest

TABLE 3. HIGHLY CORRELATED FEATURES

After this, split the data into a training set and testing set. We experimented with various ratios of the train-test split, and it was observed that 80% of the training set and 20% of the testing set of total data was the most efficient as it depicted low bias and low variance for ML algorithms we implemented [35].

We also performed ten-fold Cross Validation (CV) over the training data. Default hyperparameter (DHP) and Hyperparameter Optimization (HPO) were carried out so that more enhanced results can be achieved in terms of the performance metrics. Hence, quantitative, and qualitative analyses are presented so that the most efficient model can be proposed finally. The overall workflow diagram is depicted in Fig 1.

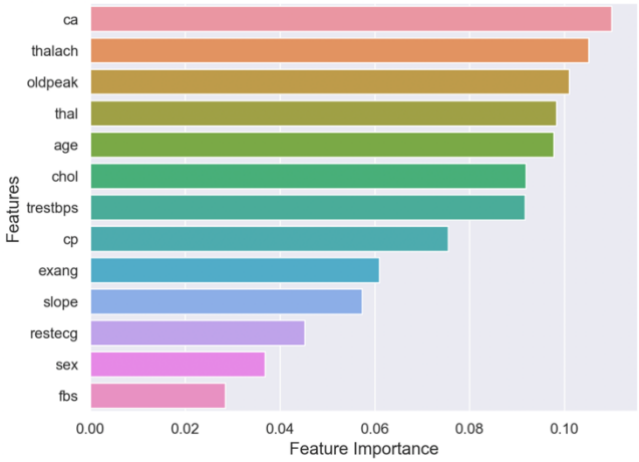

FIG 3. FEATURE IMPORTANCE

3.3 Machine Learning

In this section, we explain the ML algorithms we used in this study. The algorithms we used include Adaboost (AdB) [36, 37], Decision Trees (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forests (RF), Support Vector Machines (SVM), and XGBoost (XGB).

3.3.1 Adaptive Boosting (AdB; Adaboost)

AdB is an ensemble learning method that creates a strong classifier from several weak classifiers. It is a boosting algorithm that uses an iterative approach to learn from the mistakes of weak classifiers and turn them into strong ones. During the training phase, the distribution weight of the sample is increased as the error rate increases, and oppositely the new distribution weight is reduced as the error rate decreases. Then samples are continually trained with the unknown distribution weights. The aim is to have strong feedback by reducing the next machine's error and reaching better accuracy rates in the end. AdB is best used to boost the performance of decision trees on binary classification problems. AdB was originally called AdaBoost.M1 by the authors, Freund and Schapire [37].

3.3.2 Decision Trees (DT)

DT is a powerful tool for both classification and regression tasks. It is a non-parametric supervised

learning algorithm that can be used for solving regression and classification problems, unlike other supervised learning algorithms. DTs are highly interpretable and provide a foundation for more complex algorithms, e.g., Random Forest (RF). DTs classify the examples by recursively splitting the training data into subsets based on the values of the attributes until a stopping criterion is met. DTs are able to generate understandable rules, perform classification without requiring much computation, and are able to handle both continuous and categorical data [38].

3.3.3 K-Nearest Neighbors (KNN)

The KNN algorithm is a simple, non-parametric, supervised learning classifier used for classification and regression problems. KNN is a lazy learning algorithm that uses the entire dataset in its training phase and stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). Whenever a prediction is required for an unseen data instance, it searches through the entire training dataset for k-most similar instances and the data with the most similar instance is finally returned as the prediction. KNN is mainly used in statistical estimation and pattern recognition [39].

3.3.4 Logistic Regression (LR)

LR is a statistical method used for binary classification problems, where the outcome of a dependent variable is predicted based on previous observations. It is a simple and efficient method for binary and linear classification problems, used to calculate the probability of a binary event occurring, and to deal with issues of classification [40].

3.3.5 Random Forests (RF)

RF is an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of DTs at training time. RF is a collection of DTs that work together to improve the accuracy and stability of the model. For classification tasks, the output of the RF is the class selected by most trees, while for regression tasks, the mean or average prediction of the individual trees is the output. RF corrects for DTs' tendency to overfit to their training set, reducing variance at the expense of a small increase in bias and some loss of interpretability [41].

3.3.6 Support Vector Machines (SVM)

SVM is a supervised ML algorithm used for both classification and regression problems. The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. SVM kernel is a function that takes low-dimensional input space and transforms it into higher-dimensional space, making it useful in non-linear separation problems. SVM chooses the extreme points/vectors that help in creating the hyperplane, which are called support vectors. SVM is one of the most popular supervised learning algorithms used for classification and regression problems [42].

3.3.7 XGBoost (XGB)

XGB, short for eXtreme Gradient Boosting, is an optimized distributed gradient boosting library designed for efficient and scalable training of ML models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. The algorithm implements gradient boosting, a powerful ensemble learning technique, to build DT models. XGB is highly customizable and allows for fine-tuning of various model parameters to performance optimization [43].

We used the training set to train these models one-by-one, while the test set served as an independent dataset to evaluate their performance and assess their generalization ability. Each algorithm underwent a training process where it learned from the training data by identifying patterns and relationships between the input features and the target variable, which was the presence or absence of CVD in this case.

3.4 Performance Metrics

After training of the ML models, we evaluated their performance using standard evaluation metrics to assess their predictive capabilities, including Accuracy, Precision, Recall/Sensitivity, F1 score, and the Area Under the Receiver Operating Characteristic curve (AUC-ROC). These metrics are explained below:

TP = True Positives; the measure of positive predictions made by the ML model which are correct.

TN = True Negatives; the measure of negative predictions made by the ML model which are correct.

FP = False Positives; the measure of positive predictions made by the ML model which are incorrect.

FN = False Negatives; the measure of negative predictions made by the ML model which are incorrect.

- Accuracy: Accuracy measures the proportion of correctly predicted instances to the total number of input samples [44]. Mathematically,

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

- Precision: Precision measures the ratio of TP predictions to the total predicted positives measures [44]. Mathematically,

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

- Recall/Sensitivity: Recall, also known as Sensitivity or the True Positive Rate, is a measure of a model's ability to correctly identify all relevant instances [44]. Mathematically,

$$Recall = \frac{TP}{TP + FN}$$

Both Recall and Precision are measures of a model's performance in a binary classification problem, and they are based on relevance. Recall is used to minimize FNs and is important in domains such as medical, where missing a positive case has a much bigger cost than predicting FNs.

- F1 Score: F1 score is an evaluation metric that combines Precision and Recall scores of a model to assess its predictive skill. It is calculated as the harmonic mean of Precision and Recall scores, providing a balance between the two metrics. F1 score is used to evaluate the performance of a model in binary and multi-class classification problems and is designed to work well on imbalanced data. Mathematically,

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

- AUC-ROC: The AUC-ROC score measures the Area under the Receiver Operating Characteristic Curve. It ranges from 0 to 1, with higher values indicating better model performance. The AUC is an effective and combined measure of Sensitivity and Specificity that describes the inherent validity of diagnostic tests, especially in medical diagnostic test evaluation. The higher the ROC, the better the model is at distinguishing between patients with the disease and no disease [45]. The ROC curve provides a graphical representation of a classifier's performance, and the AUC provides an aggregate measure of performance across all possible classification thresholds (See Fig 4). The AUC is an overall summary of diagnostic accuracy, and it equals 0.5 when the ROC curve corresponds to random chance and 1.0 for perfect accuracy [46].

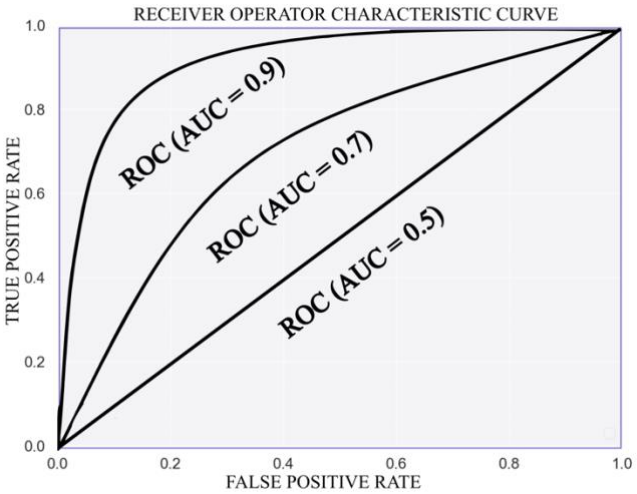

FIG 4. RECEIVER OPERATING CHARACTERISTIC CURVE

By calculating these evaluation metrics for each model, we gained insights into their strengths and weaknesses in predicting CVD risk. This information allowed us to compare the performance of different models and identify the best models.

3.5 Hyperparameter Optimization

To optimize the performance of the ML models, we employed a Hyperparameter Tuning process. Hyperparameters are adjustable parameters that determine the behavior and performance of the models. Grid Search Cross Validation (GSCV) and Randomized Search Cross Validation (RSCV) techniques were utilized to explore different combinations of hyperparameters and identify the optimal configuration that maximizes the model's performance.

3.5.1 Grid Search Cross Validation (GSCV)

GSCV defines a grid of possible hyperparameter values and systematically searching through all combinations to find the best set of hyperparameters. GSCV exhaustively evaluates the models' performance for each combination using Cross Validation (CV),

which helps in assessing the generalization capability of the models. The best set of hyperparameters is chosen based on the highest CV Accuracy score, the optimal value for the hyperparameters, and the best model that has the best hyperparameter [47].

3.5.2 Randomized Search Cross Validation (RSCV)

RSCV involves defining a range of possible hyperparameter values and randomly sampling combinations to find the best set of hyperparameters. RSCV evaluates the models' performance for a random subset of hyperparameters using CV, which helps in assessing the generalization capability of the models. We can control the number of iterations in RSCV to balance the trade-off between computation time and search space exploration [48].

RSCV is more efficient than GSCV, especially when the search space is large. However, it may not guarantee finding the optimal set of hyperparameters, but it can discover new combinations that GSCV may miss [49].

By applying these techniques, we fine-tuned the hyperparameters of the models and optimized their performance.

4 RESULTS

In this section, we present the results of the various ML algorithms we evaluated in our study. The evaluation was performed on an Apple MacBook Pro with an M2 Pro chip, which features a 10-core CPU, 16-core GPU, 16-core Neural Engine, and 16GB unified memory. Table 4 and Fig 5 summarizes the best results obtained for each algorithm, including Accuracy, Precision, Recall, F1 Score, and AUC-ROC. Additionally, we provide the results from the Grid Search Cross Validation (GSCV) and Random Search Cross Validation (RSCV) techniques in Tables 5 and 6, respectively.

The AdB and RF models achieved an Accuracy of 0.8689, Precision of 0.8766, Recall of 0.8689, F1 Score of 0.869, and an AUC-ROC of 0.8734. This indicates that the AdB and RF models performed well in accurately predicting CVD and had a high discrimination ability.

The DT model exhibited lower performance compared to other models, with an Accuracy of 0.6885, Precision of 0.7047, Recall of 0.6885, F1 Score of 0.6874, and an AUC-ROC of 0.6959. The lower scores suggest that the DT model had limitations in accurately predicting CVD risk and had a lower discriminatory power.

The KNN model achieved an Accuracy of 0.7869, Precision of 0.7914, Recall of 0.7869, F1 Score of 0.7872, and an AUC-ROC of 0.7895. These results indicate moderate performance.

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
AdB	0.8689	0.8766	0.8689	0.869	0.8734
DT	0.6885	0.7047	0.6885	0.6874	0.6959
KNN	0.7869	0.7914	0.7869	0.7872	0.7895
LR	0.8525	0.864	0.8525	0.8525	0.8582
RF	0.8689	0.8766	0.8689	0.869	0.8734
SVM	0.8525	0.864	0.8525	0.8525	0.8582
XGB	0.8525	0.8532	0.8525	0.8526	0.8528

TABLE 4. BEST PERFORMANCE METRICS OF DIFFERENT ML ALGORITHMS

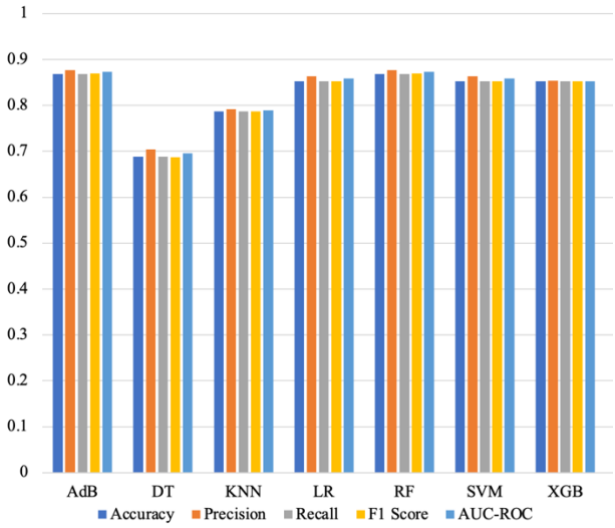

FIG 5. BEST PERFORMANCE METRICS OF DIFFERENT ML ALGORITHMS

The LR, SVM, and XGB models demonstrated similar performance with Accuracy ranging from 0.8525 to 0.8689 and Precision, Recall, F1 Score, and AUC-ROC values around these accuracy levels. These models achieved good predictive accuracy and discrimination ability, making them reliable choices for predicting CVD.

Overall, the results suggest that the AdB, LR, RF, SVM, and XGB models performed well, with high Accuracy, Precision, Recall, F1 Score, and AUC-ROC values. The DTs and KNNs models showed comparatively lower performance in our study.

5 DISCUSSION

In this study, we evaluated the performance of various ML algorithms for CVD risk prediction. Our results showed that the AdB, RF, LR, SVM, and XGB models achieved higher Accuracy, Precision, Recall, F1 Score, and AUC-ROC compared to the DT and KNN models. The AdB and RF models demonstrated the best performance, with an AUC-ROC of 0.8734. These models exhibited high predictive accuracy and discrimination ability, indicating their potential to be used for the prediction of 10-year risk CVD.

On the other hand, the DT model exhibited lower performance compared to the other models, with an AUC-ROC of 0.6959. The lower scores suggest limitations in the DT model's ability to accurately predict CVD risk and its lower discriminatory power.

Comparing our results to traditional CVD risk prediction models, such as the FRS, RRS, AHA/ACC Pooled Cohort Equations, ASSIGN Score, and QRISK, our ML-based models showed competitive or superior performance. The AUC-ROC values achieved by our top-performing models ranged from 0.7895 to 0.8734, while traditional models typically exhibited AUC-ROC values ranging from 0.6 to 0.85. Fig 6 illustrates the comparison of AUC-ROC values between ML models and the average value of traditional models.

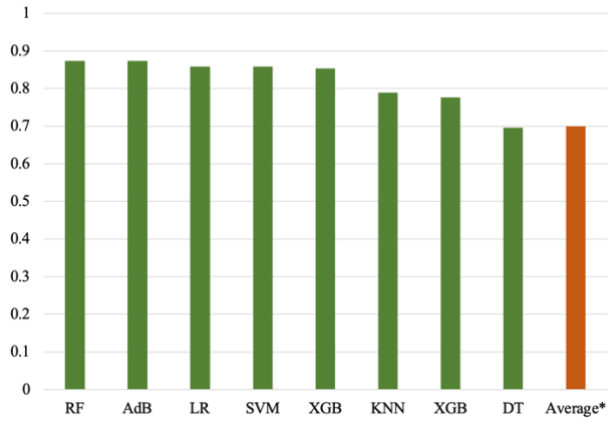

FIG 6. AUC-ROC VALUES OF DIFFERENT ML ALGORITHMS COMPARED TO THE AVERAGE AUC-ROC VALUE OF TRADITIONAL MODELS*

Our findings support previous studies highlighting the limitations of traditional models for CVD risk prediction. ML-based predictive models offer an alternative approach by leveraging a wide range of data, handling missing data effectively, and identifying novel predictors. By incorporating a more comprehensive set of risk factors, ML models can enhance the accuracy and personalization of CVD risk prediction.

However, it is important to acknowledge the limitations of our study. Firstly, the performance of the ML models was evaluated using a specific dataset, and the generalizability of the results to other populations or settings should be further investigated. Additionally, the selection of ML algorithms and hyperparameter tuning may influence the results, and alternative algorithms or parameter settings could yield different outcomes.

ML models heavily rely on the quality and representativeness of the training data. Biases, errors, or missing information in the data can potentially impact the accuracy and generalizability of the models. Therefore, ensuring high-quality, diverse, and well-curated datasets is crucial to maximize the potential of ML models in CVD risk prediction.

While our study has provided valuable insights into the application of ML for CVD risk prediction, there are several avenues for future research and improvement. First, expanding the training dataset by incorporating a larger and more diverse population would enhance the generalizability of the models. Additionally, integrating additional data sources such as genetic information, environmental factors, and lifestyle behaviors could further improve the accuracy and precision of the predictions.

Another potential area for future research is the development of ensemble models that combine the strengths of multiple ML algorithms. Ensemble methods, such as stacking or boosting, have the potential to leverage the individual strengths of different algorithms and improve overall performance.

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
AdB	0.8689	0.8711	0.8689	0.8691	0.8707
DT	0.6885	0.7047	0.6885	0.6874	0.6959
KNN	0.7705	0.7729	0.7705	0.7709	0.7716
LR	0.8525	0.8640	0.8525	0.8525	0.8582
RF	0.8689	0.8766	0.8689	0.8690	0.8734
SVM	0.8361	0.8436	0.8361	0.8362	0.8404
XGB	0.7705	0.7848	0.7705	0.7701	0.7771

TABLE 5. PERFORMANCE METRICS OF DIFFERENT ML ALGORITHMS USING GRID SEARCH CROSS VALIDATION (GSCV)

Moreover, exploring explainability techniques to better understand the underlying factors contributing to

the predictions of ML models can provide valuable insights for clinical decision-making.

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
AdB	0.8689	0.8766	0.8689	0.8690	0.8734
DT	0.6885	0.7136	0.6885	0.6855	0.6986
KNN	0.7869	0.7914	0.7869	0.7872	0.7895
LR	0.8525	0.8640	0.8525	0.8525	0.8582
RF	0.8689	0.8766	0.8689	0.8690	0.8734
SVM	0.8525	0.8640	0.8525	0.8525	0.8582
XGB	0.8525	0.8532	0.8525	0.8526	0.8528

TABLE 6. PERFORMANCE METRICS OF DIFFERENT ML ALGORITHMS USING RANDOM SEARCH CROSS VALIDATION (RSCV)

Furthermore, conducting prospective studies to validate the performance of the ML models in real-world clinical settings would be crucial. This would involve implementing the models in clinical practice and assessing their impact on patient outcomes, such as reducing CVD burden and improving risk management strategies.

These advancements have the potential to revolutionize CVD risk assessment, enable personalized preventive strategies, and ultimately contribute to reducing the burden of CVD.

6 CONCLUSIONS

Our study provides compelling evidence that ML models surpass traditional models in CVD risk prediction, as indicated by higher AUC-ROC values. The superior performance of ML models highlights their potential to revolutionize CVD risk assessment and management.

However, the interpretability of ML models can be challenging. While these models can effectively identify patterns and relationships in complex data, understanding the underlying mechanisms and factors influencing the predictions may not be straightforward. Transparency and interpretability of ML algorithms should be prioritized to build trust and facilitate the adoption of these models in clinical practice.

Despite these limitations, ML models hold immense potential in replacing traditional models for CVD risk prediction. By incorporating a wide range of risk factors, including both conventional and novel predictors, ML models offer a more comprehensive and personalized assessment of an individual's CVD risk. This approach enables clinicians to tailor preventive interventions and treatment strategies according to each patient's specific risk profile, ultimately improving patient outcomes.

Furthermore, ML models, having the ability to continuously learn and adapt as new data becomes available, ensure that risk prediction models stay up to date with evolving knowledge and can better capture the changing landscape of CVD risk factors and trends. By harnessing the power of ML in CVD risk prediction, we have the potential to reduce the burden of CVD on individuals, healthcare systems, and the society. Early identification of individuals at high risk of developing CVD allows for timely interventions and lifestyle modifications, leading to prevention, better management, and improved overall health outcomes.

7 REFERENCES

1. Dritsas, E. and Trigka, M. (2023) ‘Efficient data-driven machine learning models for Cardiovascular Diseases Risk Prediction’, *Sensors*, 23(3), p. 1161.doi:10.3390/s23031161.
2. Goff, D.C. *et al.* (2014) ‘2013 ACC/AHA guideline on the assessment of cardiovascular risk’, *Circulation*, 129(25_suppl_2). doi:10.1161/01.cir.0000437741.48606.98.
3. Cook, N. R., Paynter, N. P., Eaton, C. B., Manson, J. E., Martin, L. W., Robinson, J. G., Rossouw, J. E., Wassertheil-Smoller, S., & Ridker, P. M. (2012). Comparison of the Framingham and Reynolds Risk scores for global cardiovascular risk prediction in the multiethnic Women's Health Initiative. Circulation, 125(14), 1748–S11. https://doi.org/10.1161/CIRCULATIONAHA.111.075929.
4. Jahangiry, L., Farhangi, M. A., & Rezaei, F. (2017). Framingham risk score for estimation of 10-years of cardiovascular diseases risk in patients with metabolic syndrome. Journal of health, population, and nutrition, 36(1), 36. https://doi.org/10.1186/s41043-017-0114-0
5. Kim, J. O. R., Jeong, Y. S., Kim, J. H., Lee, J. W., Park, D., & Kim, H. S. (2021). Machine Learning-Based Cardiovascular Disease Prediction Model: A Cohort Study on the Korean National Health Insurance Service Health Screening Database. *Diagnostics (Basel, Switzerland)*, *11*(6), 943. https://doi.org/10.3390/diagnostics11060943
6. Ordikhani M, Saniee Abadeh M, Prugger C, Hassannejad R, Mohammadifard N, et al. (2022) An evolutionary machine learning algorithm for cardiovascular disease risk prediction. PLOS ONE 17(7):

e0271723. https://doi.org/10.1371/journal.pone.0271723

7. Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. *Open medicine (Warsaw, Poland)*, *17*(1), 1100–1113. https://doi.org/10.1515/med-2022-0508
8. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M (2019) Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLOS ONE 14(5): e0213653. https://doi.org/10.1371/journal.pone.0213653
9. Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M., & Brindle, P. (2007). Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ (Clinical research ed.), 335(7611), 136. https://doi.org/10.1136/bmj.39261.471806.55
10. Collins, G. S., & Altman, D. G. (2009). An independent external validation and evaluation of Qrisk Cardiovascular Risk Prediction: A Prospective Open Cohort Study. *BMJ*, *339*(jul07 2). https://doi.org/10.1136/bmj.b2584
11. Badawy, M. A. E. M. D., Naing, L., Johar, S., Ong, S., Rahman, H. A., Tengah, D. S. N. A. P., Chong, C. L., & Tuah, N. A. A. (2022). Evaluation of cardiovascular diseases risk calculators for CVDs prevention and management: scoping review. BMC public health, 22(1), 1742. https://doi.org/10.1186/s12889-022-13944-w
12. Cortes-Bergoderi M, Thomas RJ, Albuquerque FN, Batsis JA, Burdiat G, Perez-Terzic C, Trejo-Gutierrez J, Lopez-Jimenez F. Validity of cardiovascular risk prediction models in Latin America and among Hispanics in the United States of America: a systematic review. Rev Panam Salud Publica. 2012 Aug;32(2):131-9. doi: 10.1590/s1020-49892012000800007. PMID: 23099874.
13. Clark, C. J., Alonso, A., Spencer, R. A., Pencina, M., Williams, K., & Everson-Rose, S. A. (2014). Predicted long-term cardiovascular risk among young adults in the national longitudinal study of adolescent health. American journal of public health, 104(12), e108–e115. https://doi.org/10.2105/AJPH.2014.302148
14. Bosomworth N. J. (2011). Practical use of the Framingham risk score in primary prevention: Canadian perspective. Canadian family physician Medecin de famille canadien, 57(4), 417–423.
15. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. Heart. 2008 Jan;94(1):34-9. doi: 10.1136/hrt.2007.134890. Epub 2007 Oct 4. PMID: 17916661.
16. Livingstone, S.J., Guthrie, B., Donnan, P.T. *et al.* Predictive performance of a competing risk cardiovascular prediction tool CRISK compared to QRISK3 in older people and those with comorbidity: population cohort study. *BMC Med* **20**, 152 (2022). https://doi.org/10.1186/s12916-022-02349-6
17. Colantonio, L. D., Richman, J. S., Carson, A. P., Lloyd-Jones, D. M., Howard, G., Deng, L., Howard, V. J., Safford, M. M., Muntner, P., & Goff, D. C. (2017). Performance of the atherosclerotic cardiovascular disease pooled cohort risk equations by social deprivation status. *Journal of the American Heart Association*, *6*(3). https://doi.org/10.1161/jaha.117.005676
18. Vasan, R. S., & van den Heuvel, E. (2022). Differences in estimates for 10-year risk of cardiovascular disease in black versus white individuals with identical risk factor profiles using pooled cohort equations: An in silico cohort study. *The Lancet Digital Health*, *4*(1). https://doi.org/10.1016/s2589-7500(21)00236-3
19. Cortes-Bergoderi M, Thomas RJ, Albuquerque FN, Batsis JA, Burdiat G, Perez-Terzic C, Trejo-Gutierrez J, Lopez-Jimenez F. Validity of cardiovascular risk prediction models in Latin America and among Hispanics in the United States of America: a systematic review. Rev Panam Salud Publica. 2012 Aug;32(2):131-9. doi: 10.1590/s1020-49892012000800007. PMID: 23099874.
20. cigna. (n.d.). *Atherosclerotic cardiovascular disease risk assessment ... - CIGNA*. Atherosclerotic Cardiovascular Disease Risk Assessment: Emerging Laboratory Evaluations. https://static.cigna.com/assets/chcp/pdf/coveragePolicies/medical/mm_0137_coveragepositioncriteria_cardiac_disease_risk_laboratory_studies.pdf
21. Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. JAMA. 2007 Feb 14;297(6):611-9. doi: 10.1001/jama.297.6.611.

Erratum in: JAMA. 2007 Apr 4;297(13):1433. PMID: 17299196.

22. DeFilippis A, Blaha M, Ndumele C, et al. The Association of Framingham and Reynolds Risk Scores With Incidence and Progression of Coronary Artery Calcification in MESA (Multi-Ethnic Study of Atherosclerosis). J Am Coll Cardiol. 2011 Nov, 58 (20) 2076–2083.https://doi.org/10.1016/j.jacc.2011.08.022
23. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, Lassale CM, Siontis GC, Chiocchia V, Roberts C, Schlüssel MM, Gerry S, Black JA, Heus P, van der Schouw YT, Peelen LM, Moons KG. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016 May 16;353:i2416. doi: 10.1136/bmj.i2416. PMID: 27184143; PMCID: PMC4868251.
24. Alaa, Ahmed & Bolton, Thomas & Angelantonio, Emanuele & Rudd, James & Schaar, Mihaela. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLOS ONE. 14. e0213653. 10.1371/journal.pone.0213653.
25. Chiarito, M., Luceri, L., Oliva, A., Stefanini, G., & Condorelli, G. (2022). Artificial Intelligence and Cardiovascular Risk Prediction: All That Glitters is not Gold. *European cardiology*, *17*, e29. https://doi.org/10.15420/ecr.2022.11
26. Krittanawong, C., Virk, H.U.H., Bangalore, S. *et al.* Machine learning prediction in cardiovascular diseases: a meta-analysis. *Sci Rep* 10, 16057 (2020). https://doi.org/10.1038/s41598-020-72685-1
27. Asif, Md. Asfi & Nishat, Mirza & Faisal, Fahim & Dip, Rezuanur & Udoy, Mahmudul & Shikder, Md & Ahsan, Ragib. (2021). Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease. Engineering Letters. 29. 731-741.
28. Ward, A., Sarraju, A., Chung, S. *et al.* Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. *npj Digit. Med.* **3**, 125 (2020). https://doi.org/10.1038/s41746-020-00331-1.
29. Subramani, S., Varshney, N., Anand, M. V., Soudagar, M. E., Al-keridis, L. A., Upadhyay, T. K., Alshammari, N., Saeed, M., Subramanian, K., Anbarasu, K., & Rohini, K. (2023). Cardiovascular diseases prediction by machine learning incorporation with Deep Learning. *Frontiers in Medicine*, *10*. https://doi.org/10.3389/fmed.2023.1150933
30. Cook N. R. (2010). Methods for evaluating novel biomarkers - a new paradigm. *International journal of clinical practice*, *64*(13), 1723–1727. https://doi.org/10.1111/j.1742-1241.2010.02469.x
31. Janosi,Andras, Steinbrunn,William, Pfisterer,Matthias, and Detrano,Robert. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.
32. Al Ahdal, A., Rakhra, M., Rajendran, R. R., Arslan, F., Khder, M. A., Patel, B., Rajagopal, B. R., & Jain, R. (2023). Monitoring Cardiovascular Problems in Heart Patients Using Machine Learning. *Journal of healthcare engineering*, *2023*, 9738123. https://doi.org/10.1155/2023/9738123
33. *Pandas*. pandas. (n.d.). https://pandas.pydata.org/
34. sklearn. (n.d.). *Sklearn.ensemble.extratreesclassifier*. scikit. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html
35. Joseph, V. R., Optimal ratio for data splitting, *Stat. Anal. Data Min.: ASA Data Sci. J.*. 15 (2022), 531– 538. https://doi.org/10.1002/sam.11583
36. Sevinç E. (2022). An empowered AdaBoost algorithm implementation: A COVID-19 dataset study. *Computers & industrial engineering*, *165*, 107912. https://doi.org/10.1016/j.cie.2021.107912
37. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504
38. *What is a decision tree*. IBM. (n.d.). https://www.ibm.com/topics/decision-trees
39. *What is the K-nearest neighbors algorithm?*. IBM. (n.d.-b). https://www.ibm.com/topics/knn
40. ScienceDirect. (n.d.). *Logistic regression*. Logistic Regression - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/computer-science/logistic-regression
41. *What is the K-nearest neighbors algorithm?*. IBM. (n.d.-c). https://www.ibm.com/topics/knn
42. Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, 137–142. https://doi.org/10.1007/bfb0026683
43. *XGBoost documentation*. XGBoost Documentation - xgboost 1.7.5 documentation. (n.d.). https://xgboost.readthedocs.io/en/stable/
44. Mishra, A. (2020, May 28). *Metrics to evaluate your machine learning algorithm*. Medium.

https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234
45. Hajian-Tilaki K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian journal of internal medicine*, *4*(2), 627–635.
46. Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, *115*(5), 654–657. https://doi.org/10.1161/circulationaha.105.594929
47. Mishra, A. (2020, May 28). *Metrics to evaluate your machine learning algorithm*. Medium. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234
48. *Sklearn.model_selection.RANDOMIZEDSEARCHCV*. scikit-learn. (n.d.-b). https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html
49. Bergstra, Bengio. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research 13 (2012) 281-305*.

8 CONFLICTS OF INTEREST

The authors declare no conflict of interest in relation to this research study.

9 FUNDING

This research study did not receive any specific funding from external sources. The authors conducted the study independently without financial support or influence from any organization.

10 CODE

The code used in this study to evaluate the ML models can be found at: https://github.com/simrandahia/CVD-RISK-PREDICTION.git