doi:10.20944/preprints202108.0471.v1

Article

Identifying the main risk factors for CVD prediction using machine learning algorithms

Luis Rolando Guarneros-Nolasco[1], Nancy Aracely Cruz Ramos[2], Giner Alor-Hernández[3], Lisbeth Rodríguez Mazahua[4], and José Luis Sánchez-Cervantes [5,*]

[1,2,3,4] División de Estudios de Posgrado e Investigación, Tecnológico Nacional de México / I.T. Orizaba, Av. Oriente 9 No. 852 Col. Emiliano Zapata, C.P. 94320, Orizaba, Veracruz, México; luisguarneros@gmail.com; nancy.cramos5@gmail.com; galorh@orizaba.tecnm.mx; lrodriguezm@orizaba.tecnm.mx

[5] CONACYT-Tecnológico Nacional de México / I.T. Orizaba, Av. Oriente 9 No. 852 Col. Emiliano Zapata, C.P. 94320, Orizaba, Veracruz, México.; jlsanchez@conacyt.mx

* Correspondence: jlsanchez@conacyt.mx; Tel. +52 229 781 3796

Abstract: CVDs are a leading cause of death globally. In CVDs, the heart is unable to deliver enough blood to other body regions. Since effective and accurate diagnosis of CVDs is essential for CVD prevention and treatment, machine learning (ML) techniques can be effectively and reliably used to discern patients suffering from a CVD from those who do not suffer from any heart condition. Namely, machine learning algorithms (MLAs) play a key role in the diagnosis of CVDs through predictive models that allow us to identify the main risks factors influencing CVD development. In this study, we analyze the performance of ten MLAs on two datasets for CVD prediction and two for CVD diagnosis. Algorithm performance is analyzed on top-two and top-four dataset attributes/features with respect to five performance metrics –accuracy, precision, recall, f1-score, and roc-auc – using the train-test split technique and k-fold cross-validation. Our study identifies the top two and four attributes from each CVD diagnosis/prediction dataset. As our main findings, the ten MLAs exhibited appropriate diagnosis and predictive performance; hence, they can be successfully implemented for improving current CVD diagnosis efforts and help patients around the world, especially in regions where medical staff is lacking.

Keywords: Big data; Health prevention; Machine learning; Medical data

1. Introduction

In 2019, the World Health Organization (WHO) predicted that 17.5 million people would die from cardiovascular diseases (CVDs), thus accounting for 30% of deaths worldwide. CVDs are the leading cause of death globally, as more people die each year from CVD-related diseases than from anything else. Of all CVDs, an estimated 7.4 million are attributed to coronary heart disease, while 6.7 million are attributed to stroke, hypertension, coronary artery disease, rheumatic heart disease, and heart failure, among others. CVDs affect low- and middle-income nations the most. In fact, it is estimated that by 2030, nearly 23.6 million people will die from CVDs, as it is expected to remain the leading cause of death in the world's poorest countries [1].

CVDs include several types of heart conditions. The most common of them all, coronary heart disease, may cause heart attacks that kill more than 370,000 people each year. Heart failure is another CVD leading to morbidity and mortality and one of the earliest manifestations of CVD. In recent years, the World Heart Federation has defined multiple risk factors affecting the incidence and occurrence of heart failure, such as arterial hypertension, diabetes, smoking, defective heart valves, damaged heart muscles, and obesity [2].

Since "classical" CVD risk factors, such as hypertension, have been successfully treated with medication, the balance between risk factors depending on age and sex and their distribution across the general population may change over time. Moreover, new and relatively less-known risk factors may emerge. As regards CVD diagnosis, timing and accuracy are key, yet not always ensured. Although early and accurate CVD detection helps medical staff determine appropriate and effective treatments to increase the chances of survival of patients, many developing countries and low-income regions lack specialists to perform such diagnostic tests. Moreover, when CVD diagnoses are inaccurate and medical procedures are performed incorrectly, they may jeopardize patient health.

In the last years, multiple organizations and researchers have built large databases of electronic health records (EHR). Along with timely and accurate diagnoses, such databases contribute to current efforts to improve CVD patient life quality in the long-term and provide researchers the opportunity to identify potential CVD risk factors among age- and sex-specific patient groups in the general population. From this perspective, computational sciences support the healthcare sector with valuable CVD predictions through computer-aided detection methods.

Among modern methods for computer-aided detection, machine learning (ML) is an emerging technology for clinical data analysis and prediction generation in the context of early detection of diseases. In this work, we analyze the performance of ten machine learning algorithms (MLAs), such as linear regression, decision trees, support vector machine, and k-nearest neighbor, among others, using four datasets with clinical data of patients diagnosed with heart disease. Our goal is to identify the main risk attributes that have impact on the development of CVDs in the general population. The remainder of this paper is structured as follows: section 2 discusses current research on MLA applied in clinical datasets, MLA performance metrics, and clinical datasets available in repositories for the data science community. Subsequently, section 3 presents the evaluation model conducted to identify the main CVD risk factors from dataset attributes. Then, section 4 presents and discusses the results from the case study. Finally, in section 5, we propose our conclusions and highlight our suggestions for upcoming works.

2. Related work

This section reviews research using public datasets for CVD diagnosis and prevention. The reviewed research articles are classified into two main trends: CVD prediction and CVD diagnosis.

2.1. CVD prediction

Pandey et al. [3] designed a model for predicting heart disease to assist medical professionals in predicting heart disease status. The model exploits the Cleveland Heart Disease dataset and uses the J38 decision tree for classifying heart disease based on a series of clinical attributes. The model results highlighted fasting blood sugar as the most important heart disease attribute. Samuel et al. [4] proposed an integrated decision support process (combining ANN and Fuzzy_AHP) for heart failure prediction. The researchers analyzed the performance of said process using three performance metrics and concluded that it could be employed to accurately predict the risks of suffering from heart failure in clinical settings. In Amin et al. [5], the researchers sought to identify key attributes and data mining procedures that could improve the accuracy of CVD prediction. To this end, the researchers developed a series of predictive models using different combinations of features and seven classification methods: k-NN, decision tree, Naïve Bayes, Logistic Regression, support vector machine (SVM), Neural Network, and Vote. The results showed that the best-performing model achieved an accuracy of 87.4% in terms of heart disease prediction. Mienye et al. [6] proposed a two-stage model that effectively predicts heart diseases. First, the researchers trained an improved sparse autoencoder (SAE), which is an unsupervised neural network that serves to study the best description of the training data. Then, they employed an artificial neural network (ANN) for predicting patient health status based on the learned records. The experimental results obtained with the proposed method increased the performance of the ANN classifier. In turn, Chicco & Jurman [7] applied a series of ML classifiers to both predict patient survival and identify the characteristics associated with the most relevant heart failure risk factors. Similarly, the researchers developed an alternative feature classification study using traditional biostatistical tests and compared the results with those obtained by the MLAs. They concluded in their analysis that serum creatinine and ejection fraction were the most significant attributes for predicting heart failure. Ayon et al. [8] studied seven ML models for coronary heart disease prediction using the Statlog and Cleveland datasets. From their comparative studies the researchers found that the highest accuracy (98.15%) on the Statlog dataset was obtained with Deep Neural Network, whereas SVM showed the best performance on the Cleveland dataset (97.36%). Mohan et al. [9] introduced an ML-based heart disease prediction model (HRFLM) combining Random Forest features and a linear method. The system operates with diverse feature configurations and various classification techniques. According to the test results, the model performed effectively, with an accuracy level of 88.7%. In Shah et al. [10], the researchers relied on ML techniques for effectively predicting heart disease using a small number of features and running a few tests. The researchers used 14 essential attributes from the Cleveland dataset and conducted a series of performance tests on four MLAs. Their results showed that the highest accuracy in terms of heart disease prediction was achieved with K-Nearest Neighbor. Dwivedi [11] tested six ML techniques for heart disease prediction. They reported the highest accuracy (85%) with logistic regression on the Statlog dataset. From a similar perspective, Belavagi and Muniyal [12] used historical medical data to predict coronary heart disease with the South African Heart Disease dataset using three MLAs to discover correlations in the data to improve coronary heart disease prediction rate. The results showed that the Nayve Bayes algorithm was promising for heart disease detection. Finally, researchers Deepika and Seema [13] used effective mechanisms for chronic disease prediction by mining health data. They used four MLAs to perform diabetes and heart disease diagnosis and presented the comparative revision of the diverse classifiers to measure their performance based on accuracy. According to the results, the highest accuracy was achieved by SVM (95.556%) on the heart disease dataset and by Naive Bayes (73.588%) on the diabetes dataset.

2.2. CVD diagnosis

Tiwaskar et al. [14] conducted a study to compare statistical, ML, and data mining methods in terms of their ability to assist in predicting heart failure risks. The researchers compared the performance of statistical evaluation, Decision Trees, Random Forest, and convolutional neural network, and they obtained prediction accuracy results of 85%, 80.1%, 85.38%, and 93%, respectively. Similarly, Nahar et al. [15] analyzed those health factors that contribute to heart disease in both genders. To this end, they relied on rule mining, a computational intelligence approach. As main results, the researchers found that factors such as asymptomatic chest pain and the existence of exercise-induced angina pectoris pointed to the probable presence of heart disease in both men and women. From a slightly different perspective, Ahmad et al. [16] conducted a survival analysis of heart failure patients admitted to two hospitals in Pakistan and used Cox regression to model mortality. The researchers found age, renal dysfunction, blood pressure, ejection fraction, and anemia as significant risk factors for mortality among patients suffering from heart failure. In Detrano et al. [17], patients were classified according to whether or not they suffered from heart disease using cardiac catheterization to test a new discriminant function model to estimate probabilities of occurrence of coronary heart disease. If one or two coronary arteries in a patient showed more than 50% of narrowing, said patient was considered to suffer from heart disease. Shimpi et al. [18] proposed an ML-based model for cardiac arrhythmia detection and classification. The model compares different MLAs – Random Forest, SVM, and Logistic Regression – and chooses the most accurate, i.e. SVM. Similarly, Niazi et al. [19] introduced a model for cardiac arrhythmia diagnosis using KNN and SVM as classification algorithms using 20-fold for cross-validation. The average accuracy achieved was 73.85% by KNN and 68.8% by SVC. Fida et al. [20] proposed a classifier ensemble method for improving heart disease diagnosis using the Cleveland, Statlog, and South African Hearth datasets. Namely, homogeneous ensemble was applied for heart disease classification. Then, the results were optimized using a genetic algorithm. To evaluate the data, the researchers used a 10-fold cross-validation, whereas the performance of the method was evaluated using the metrics of classifier accuracy, sensitivity, and specificity to test the feasibility of the method. The genetic algorithm proved to be an effective technique for optimizing and finding quality solutions, since the proposed method achieved a maximum accuracy of 98.63%. Singh & Singh [21] designed a cardiac arrhythmia diagnosis system that can identify the 30 best attributes using three filter-based feature selection methods on three different ML methods (linear SVM, Random Forest, and JRip) applied on the cardiac arrhythmia dataset. The system achieved its highest level of accuracy (85.58%) with Random Forest. Soman & Bobbie [22] applied three ML methods – OneR, Naive Bayes, and J48 – to classify arrhythmias from ECG recordings and found that OneR and Naive Bayes exhibited the most constant accuracy rate. Researchers Kodati et al. [23] used different varieties of unsupervised clustering algorithms to determine their accuracy in terms of cardiac disease search and diagnosis. The algorithms were applied on the Cleveland dataset. The study results highlighted k-means as the most appropriate algorithm for cardiac disease diagnosis.

MLAs are similarly applied for the prediction and diagnosis of chronic degenerative diseases. For instance, Haq et al. [24] proposed an ML-based diabetes diagnostic system that uses a filtering method centered on Decision Tree to select the most significant dataset attributes. The model proved to perform remarkably thanks to the different configurations of the chosen attributes. Similarly, the researchers found that plasma glucose concentrations, diabetes pedigree function, and blood mass index were the most prominent features in the dataset for diabetes prediction. Ghosh & Waheed [25] evaluated the most popular classification algorithms in terms of accuracy, precision, sensitivity, and specificity using a dataset of liver patients. Similarly, in their findings, the researchers highlighted attributes such as age, sex, SGOT, SGPT, SGPT, SGPT, ALP, total bilirubin, direct bilirubin, total protein, and albumin as crucial in deciding liver status.

Authors Mishra et al. [26] conducted a comparative study of the impact of wrapper and filter selection methods on classification performance across various chronic disease datasets. Similarly, the researchers proposed an integrated hybrid method for variable evaluation in which they associated a new alternative of K-Means cluster analysis, called Integrated Supervised K-Means, with the Correlation Feature Selection (CFS) and Best First Search (BFS) methods, thus achieving a classification accuracy of 96.85%. Danjuma [27] evaluated the performance of ML classification systems applied on the clinical prognosis of postoperative life probability among lung cancer patients. They used a k=10 cross-validation to calculate the performance accuracy of the classifiers and found that the Perceptron algorithm exhibited the best accuracy performance (82.3%). Researchers Li & Chen [28] studied the relationship between breast cancer and some factors as a means to reduce death probability of breast cancer. To this end, they used five classification systems for the classification of two breast-cancer-related datasets: the Breast Cancer Coimbra Dataset (BCCD) and the Wisconsin Breast Cancer Database (WBCD). According to the results, Random Forest performed best on the AUC metric.

According to our review of the literature, the eight most common MLAs applied in CVD detection and diagnosis include Decision Tree, Random Forest, k-Nearest Neighbors, Logistic Regression, SVM, ANN Perceptron, Gradient Boosting, and AdaBoost. On the other hand, current initiatives for detecting and diagnosing chronic degenerative diseases (i.e. diabetes, breast cancer, and lung cancer) rely mostly on algorithms K-Nearest Neighbors, SVM, AdaBoost, Random Forest, Decision Tree, Neural Network, and Logistic Regression. Additionally, we found that existing initiatives for CVD prediction and diagnosis fail to recognize all the main attributes of CVD, since the applied algorithms perform on few public datasets. For instance, Pandey et al. [3] determined only 13 key attributes from the Cleveland dataset, whereas Nahar et al. [15] and Amin et al. [5] only found two and nine attributes, respectively, also from the Cleveland dataset. In turn, on the Faisalabad dataset, Ahmad et al. [16] managed to identify five cardiac disease attributes, whereas Chicco & Jurman [7] identified only two heart disease attributes. In this sense, perhaps the greatest contribution of our study is the fact that analyze the performance of ten MLAs on four existing datasets to identify the main CVD risk factors for CVD prediction and detection. This will enable preventive CVD diagnosis to include an appropriate monitoring of the identified risk factors.

3. Materials and Methods

The subsequent sections briefly discuss our research methods and the results from the analysis of the ten MLAs on four datasets.

3.1. Datasets

We identified four main clinical datasets: the Cleveland dataset, the Framingham Heart study dataset, the Faisalabad Institute dataset, and the South African Hearth dataset. Each of them contains data on heart disease clinical instances. The Cleveland health disease dataset is an open-access dataset stored in the online repository of the University of California, Irvine (UCI). It is frequently used to perform search analyses of heart failure risk in patients, as it contains 303 patient records with no missing values. Also, the Cleveland database contains 76 attributes, 13 of which are considered as key. As Janosi et al. [29] point out, current experimental studies relying on the Cleveland dataset attempt to distinguish heart failure presence from heart failure absence. The Framingham Heart study dataset is an ongoing cohort study project being conducted in Framingham, Massachusetts. It is publicly accessible on the Kaggle website [30] and comprises 15 columns and around 4,200 rows of data. Each row presents a person's behavioral, demographic, and medical (history and current) data, while each column is a potential risk factor. The Faisalabad Institute dataset is based on 13 attributes and one class with records of 299 heart failure patients (105 women and 194 men) at Faisalabad Institute of Cardiology and the Allied hospital in Faisalabad, Pakistan. The dataset is hosted on the Kaggle website for public consultation. Finally, the South African Hearth dataset consists of 462 records of patient data and contains 13 attributes to predict mortality from heart disease. The dataset is publicly accessible from the KEEL (Knowledge Extraction based on Evolutionary Learning) website [31].

Table 1. Review of Heart Disease Datasets

Dataset	Number of attributes	Number of Records	Prediction / Diagnosis
Faisalabad Institute	13 + Class	299	Prediction/Diagnosis
Framingham	15 + Class	3,658	Prediction
Cleveland	13 + Class	303	Prediction/ Diagnosis
South African Hearth dataset	9 + Class	462	Prediction

3.2. Machine Learning Classifiers

Our study revolves around the binary prediction and identification of main CVD risk factors. We used ten different classifying procedures from the diverse areas of ML. The classifiers comprise a linear statistical approach (linear Regression [32]), three tree-based methods (Random Forest [33], XGBRF [34], and decision tree [35]), one SVM [36], one instance-based learning model [37], and four ensemble boosting methods (Gradient Boosting [38], LightGBM [39], CatBoost [40], and AdaBoost [41]). We measured the performance of each classifier algorithm independently. Then, we properly recorded all the results for further analysis.

3.2. Methodology

Several authors propose methodologies such as [5, 8, 9, 10, 12, and 20]. Our proposal is based on some of these models. We followed a six-staged methodology to evaluate the performance of the ten MLAs on the clinical datasets and thus identify the main CVD risk factors. The six stages are as follows: 1) Load data dataset, 2) Pre-process data, 3) Select attributes, 4) Run ML models, 5) Apply evaluation metrics, and 6) Process MLA/classifier performance results.

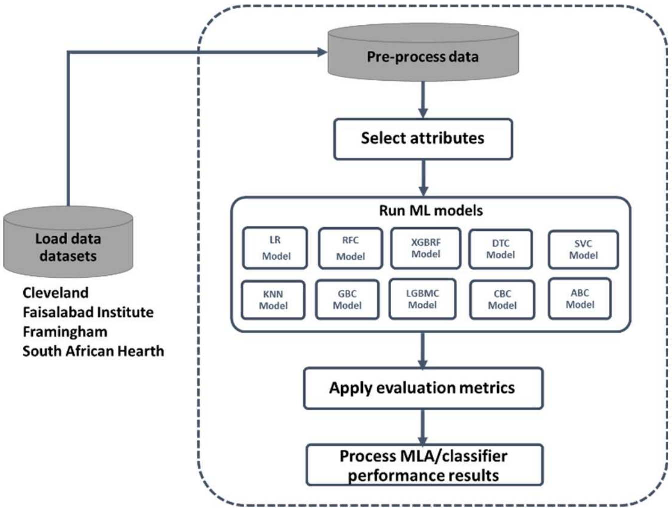


Figure 1. Methodology for the evaluation of CVD datasets.

Each methodology stage can be described as follows:

1. **Load data dataset**. Select and load data from the dataset containing clinical records of patients with CVDs.
2. **Pre-process dataset.** Review loaded data to understand their content. Then, select the classification variable to obtain the best results.
3. **Select attribute or main risk factors.** Use Random Forest to select the top-two and top-four attributes from each dataset. Split data for training and testing (i.e. 70% for training and 30% for testing), and k=10. Similarly, calculate the best parameters for RandomizedSearchCV for n_estimators, max_attributes, and max_depth. Most of the algorithms have these parameters in common, except for K-nearest neighbor and MLP. Also, parameter ramdom_state was set to 42 in all the evaluations.
4. **Run ML classifiers:** Apply the ten ML classifiers to discern participants with CVD from healthy individuals.

doi:10.20944/preprints202108.0471.v1

5. **Apply evaluation metrics**. Analyze MLA classification performance with respect to five criteria: accuracy, precision, recall, f1-score, and area under the curve (ROC-AUC).
6. **Process performance results.** Gather and compare performance values from the ten MLAs and record such results for further analysis. Then, choose the best-performing MLA or classifier.

3.3. Validation of the classification method

We analyzed the performance of the ten ML classifiers or MLAs with the help of the train-test split technique and k-fold cross-validation (k=10). Classifier performance was analyzed with respect to five performance evaluation metrics: accuracy, precision, recall, f1-score, and ROC-AUC. The train-test split technique [42] is a simple and agile procedure that is adaptable to large datasets. It can be used to assess MLA performance by splitting a given dataset into training and testing sets. Hence, a given model is trained using the training set, and then the model is applied to the test set. Cross-validation [42] is also used to calculate MLA performance, as it ensures less variance than a single split of the training and test sets. Cross-validation means segmenting the dataset into k-parts, e.g. k=3, k=5 and k=10. After performing the cross-validation, k different performance scores are obtained, which can be synthesized through a mean and standard deviation. The result is a better approximation of the algorithm's performance on the new data. This technique is usually more reliable than the train-test split method, since algorithms are trained and evaluated several times on different data. The choice of k should allow the test partition size to be large enough to construct a reasonable sample; hence, k values of 3, 5, and 10 are common.

4. Results and discussion

This section discusses the results from the several performance analyses of the ten ML classification models in terms of their ability to identify the top-two and top-four main attributes from publicly available datasets of CVD patient records. As previously mentioned, we conducted the classifier performance evaluations, first by applying the train-test split method (70%-30%), and second with k-fold cross-validation (k=10). During the evaluations, we recorded five performance measures: accuracy, precision, recall, f1-score, and ROC-AUC.

4.1 Results of MLA classifier performance

4.1.1 Attribute selection in medical diagnosis datasets

4.1.1.1 The Cleveland dataset

We applied Random Forest on the Cleveland dataset to identify and select the four most important CVD attributes. Table 2 lists the 13 key attributes of the dataset, from which the top four were retrieved. Additionally, Figure 2a graphically shows the ranking of these attributes.

Table 2. Selected attributes from the Cleveland dataset with Random Forest

Attribute name	Attribute description	Score
cp	Chest pain type	13.55
thalach	Maximum heart rate	12.52

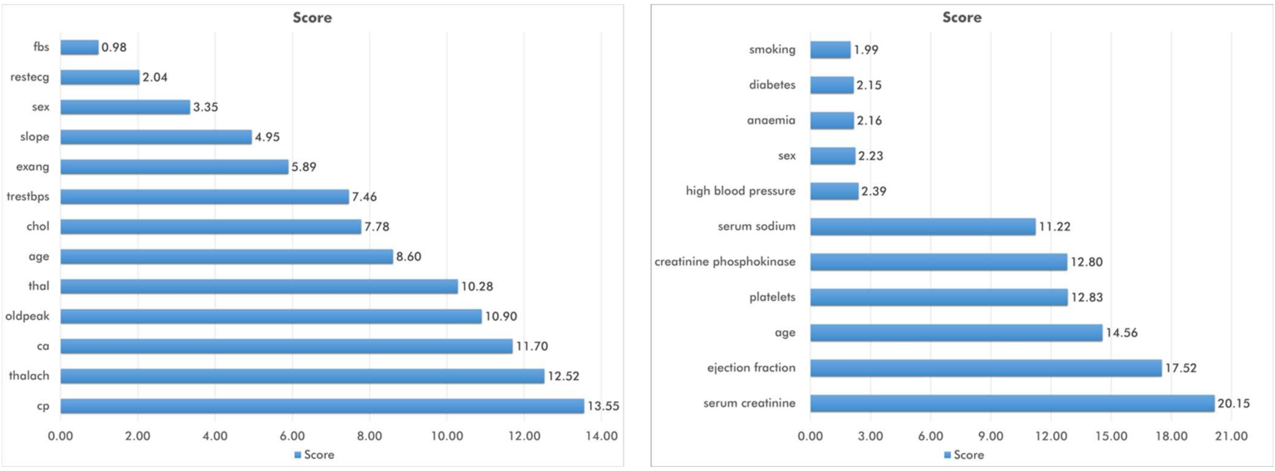
ca	Number of vessels colored by fluoroscopy	11.70
oldpeak	Exercise-induced ST depression	10.90
thal	Thallium scan	10.28
age	Age in years	8.60
chol	Serum cholesterol	7.78
trestbps	Blood pressure at rest	7.46
exang	Exercise-induced angina	5.89
slope	Slope of the peak in exercise-induced ST depression	4.95
gender	Gender	3.35
restecg	ECG results at rest	2.04
fbs	Fasting blood sugar	0.98

4.1.1.2 Faisalabad dataset

Random Forest yielded 11 main CVD attributes on the Faisalabad dataset. Table 3 lists such attributes in ranked order, whereas Figure 2b depicts a graph of said ranking. As in the previous case, the top-two and top-four attributes were used in the classifier performance analyses.

Table 3. Selected attributes from the Faisalabad dataset with Random Forest

Attribute name	Attribute description	Score
Serum creatinine	Level of blood creatinine	20.15
Ejection fraction	Percentage of blood leaving the heart at each heart beat	17.52
Age	Patient age	14.56
Platelets	Count of platelets in blood	12.83
Creatinine phosphokinase	Level of the CPK enzyme in blood	12.80
Serum sodium	Level of sodium in blood	11.22
High blood pressure	Whether a patient has hypertension	2.39
Gender	Whether a patient is a woman or a man	2.23
Anemia	Decrease of red blood cells or hemoglobin	2.16
Diabetes	Whether a patient has diabetes	2.15
Smoking	Whether a patient smokes	1.99



(**a**) (**b**)

Figure 2. Ranking of attributes selected by Random Forest for CVD prediction on (**a**) Cleveland dataset and (**b**) Faisalabad dataset

4.1.2 *Attribute selection in medical prediction datasets*

4.1.2.1 Framingham dataset

On the Framingham dataset, Random Forest ranked the most important CVD attributes as listed on Table 4. Additionally, Figure 3a graphically shows the ranking of said attributes.

Table 4. Selected attributes from the Framingham dataset with Random Forest

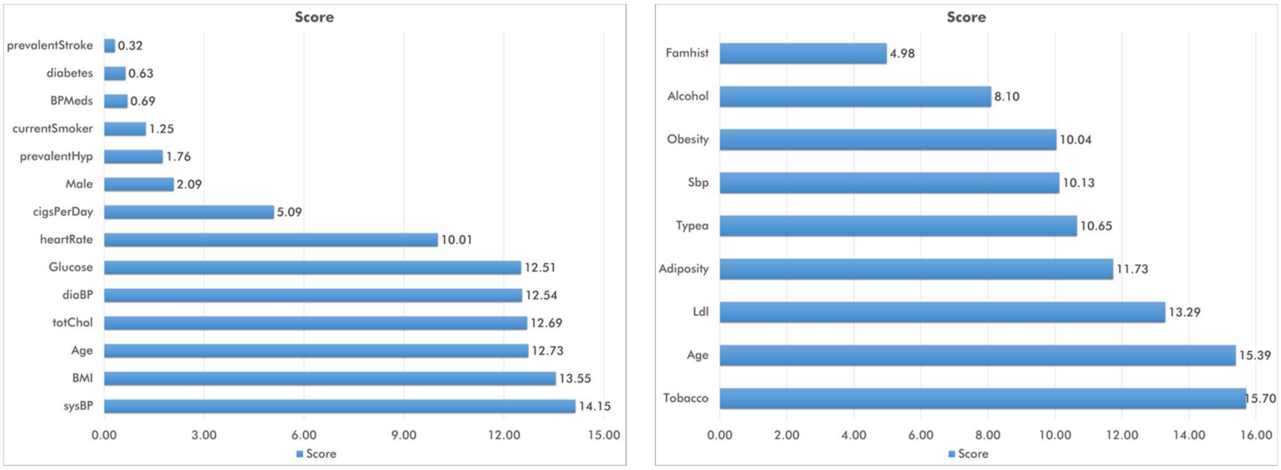
Attribute name	Attribute description	Score
sysBP	Systolic blood pressure	14.15
BMI	Body mass index	13.55
Age	Patient age at exam time	12.73
totChol	Total cholesterol	12.69
diaBP	Diastolic blood pressure	12.54
Glucose	Serum glucose	12.51
heartRate	Heart rate	10.01
cigsPerDay	Number of cigarettes smoked per day	5.09
Male	Male patient	2.09
prevalentHyp	Hypertension over 24-year follow-up	1.76
currentSmoker	Currently smoking cigarettes	1.25
BPMeds	Patient taking anti-hypertensive medications	0.69
diabetes	Diabetic patient	0.63
prevalentStroke	Stroke over 24-year follow-up	0.32

4.1.2.2 South African Hearth dataset

The nine key attributes from the South African Hearth dataset were ranked by Random Forest as listed in Table 5. Additionally, Figure 3b graphically shows the 9 ranking of such attributes. As in the three previous databases, the top two and top four attributes were used to run the classifier performance analyses.

Table 5. Main attributes on the South African Hearth dataset identified with Random Forest

Feature	Feature description	Score
Tobacco	Cumulative tobacco	15.70
Age	Age at onset	15.39
Ldl	Low density lipoprotein cholesterol	13.29
Adiposity	Adiposity	11.73
Typea	Type-A behavior	10.65
Sbp	Systolic blood pressure	10.13
Obesity	Obesity	10.04
Alcohol	Current alcohol consumption	8.10
Famhist	Family history of heart disease	4.98



(**a**) (**b**)

Figure 3. Ranking of attributes selected by Random Forest for CVD prediction on (**a**) Framingham dataset and (**b**) South African Hearth dataset

4.2 Results of the train-test split technique for classifier performance on two and four attributes

We relied on the Cleveland and Faisalabad datasets to analyze the performance of the classifiers on datasets for CVD diagnosis.

4.2.1 Datasets for CVD diagnosis

We analyzed the performance of the ten ML classifiers on both the top two and the top four dataset attributes using the train-test data split technique (70%-30%). The analysis results are discussed below.

4.2.1.1 Classifier performance on top-two CVD attributes

We tested the performance of the ten ML classifiers on the top-two attributes from the Cleveland and Faisalabad datasets. Selected Cleveland attributes comprised cp

doi:10.20944/preprints202108.0471.v1

(score=13.55) and thalach (score=12.52), and Faisalabad attributes referred to serum creatinine (score=20.15) and ejection fraction (score=17.52). Table 6 lists the results from the analysis.

Table 6. Train-test set performance analysis of classifiers on top-two attributes from the Cleveland and Faisalabad datasets

Dataset	Predictive Model	Performance evaluation metrics				
		% Accuracy	% Precision	% Recall	% f1-score	% roc_auc
Cleveland	AdaBoost Classifier	71.43	74.00	74.00	74.00	71.15
	CatBoost Classifier	**81.32**	83.67	82.00	**82.83**	**81.24**
	Decision Tree Classifier	79.12	78.18	**86.00**	81.90	78.37
	GradientBoosting Classifier	61.54	64.71	66.00	65.35	61.05
	KNeighbors Classifier	70.33	76.74	66.00	70.97	70.80
	LGBM Classifier	71.43	71.43	80.00	75.47	70.49
	Logistic Regression	78.02	**84.09**	74.00	78.72	78.46
	Random Forest Classifier	68.13	68.42	78.00	72.90	67.05
	Support Vector Classification	79.12	82.98	78.00	80.41	79.24
	XGBRF Classifier	**81.32**	83.67	82.00	**82.83**	**81.24**
Faisalabad	AdaBoost Classifier	64.44	60.00	40.54	48.39	60.84
	CatBoost Classifier	68.89	71.43	40.54	51.72	64.61
	Decision Tree Classifier	**74.44**	73.33	59.46	**65.67**	**72.18**
	GradientBoosting Classifier	67.78	61.76	56.76	59.15	66.11
	KNeighbors Classifier	72.22	**83.33**	40.54	54.55	67.44
	LGBM Classifier	68.89	62.86	59.46	61.11	67.47
	Logistic Regression	63.33	64.29	24.32	35.29	57.45
	Random Forest Classifier	70.00	63.89	**62.16**	63.01	68.82
	Support Vector Classification	58.89	50.00	13.51	21.28	52.04
	XGBRF Classifier	71.11	66.67	59.46	62.86	69.35

As can be observed from Table 6, CatBoost and XGBRF classifiers showed the best results in terms of accuracy performance (81.32%). As for the Faisalabad attributes, Decision Tree exhibited the highest accuracy (74.44%). Conversely, the lowest-performing classifiers with respect to accuracy included GradientBoosting Classifier (61.54%) on the Cleveland dataset and Support Vector Classification (58.89%) on the Faisalabad dataset. As regards precision, Logistic Regression and KNeighbors proved to be the best-performing classifiers on the Cleveland dataset and the Faisalabad dataset, respectively, with precision scores of 84.09% and 83.33%, respectively. The lowest-performing classifiers in terms of precision were once again GradientBoosting Classifier (64.71%) on the Cleveland dataset and Support Vector Classification (50.0%) on the Faisalabad dataset. In conclusion, on the Cleveland dataset, CatBoost Classifier exhibited good performance in terms of accuracy, f1-score, and roc-auc, whereas Logistic Regression performed best in terms of precision. As regards the Faisalabad dataset, Decision Tree Classifier performed best in accuracy, f1-score, and roc-auc, Decision Tree Classifier exhibited best performance in terms of precision, and Random Forest Classifier performed best in terms of recall.

4.2.1.2 Classifier performance on top-four attributes

At this stage, we tested the performance of the ML classifiers on the four best-ranked attributes from both the Cleveland dataset and the Faisalabad dataset. Cleveland attributes included cp (score= 13.55), thalach (score=12.52), ca (score=11.70), and oldpeak (score= 10.90). On the other hand, Faisalabad attributes comprised serum creatinine (score=20.15), ejection fraction (score=17.52), age (score=14.56), and platelets (score=12.83). Figure 4 graphically introduces the results of the analysis.

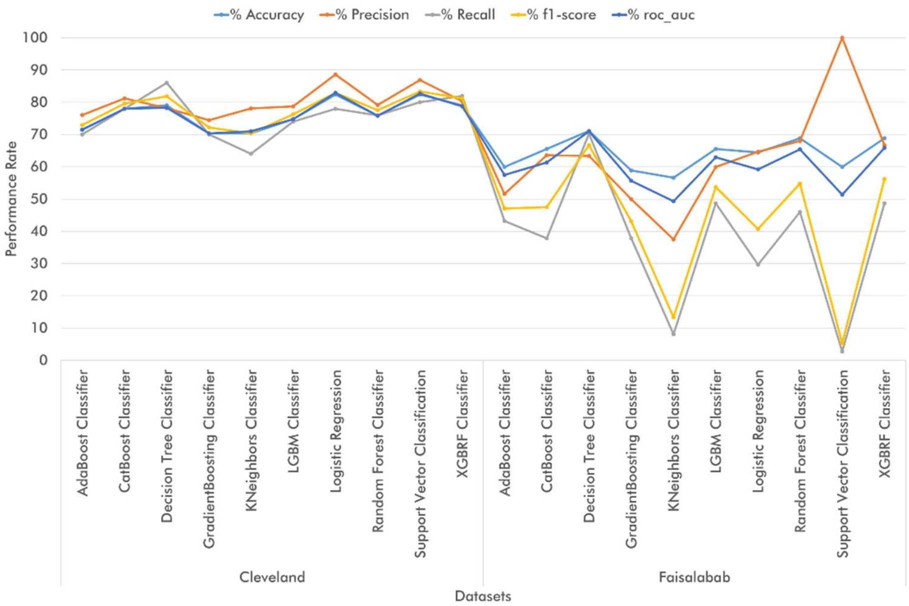


Figure 4. Train-test dataset performance of ten LM classifiers on the top-four attributes from the Cleveland and Faisalabad attributes

As depicted in Figure 4, the highest accuracy on the Cleveland dataset was achieved with Logistic Regression and Support Vector Classification (82.42%), whereas Decision Tree classifier outperformed on the Faisalabad dataset (71.11%). As for precision, the best-performing classifiers included Logistic Regression (88.64%) on the Cleveland dataset and Support Vector Classification (100.00%) on the Faisalabad dataset. Conversely, the lowest accuracy was yielded by both GradientBoosting Classifier and KNeighbors Classifier (70.33%) on the Cleveland dataset and KNeighbors Classification (56.67%) on the Faisalabad dataset. The lowest-performing algorithms in terms of precision were GradientBoosting Classifier (74.47%) on the Cleveland dataset and KNeighbors Classifier on the Faisalabad dataset (37.50%). Overall, the classifiers exhibited better performance on the Cleveland dataset across the five metrics, whereas on the Faisalabad dataset the classifiers exhibited favorable behavior only in terms of accuracy and roc-auc. In conclusion, evaluating four attributes instead of two significantly improves classifier performance in accuracy and precision metrics.

4.2.2 *Datasets for CVD prediction*

We evaluated the performance of ten ML classifiers on the top two and four attributes from the Framingham and the South African Hearth datasets. The data was split in 70% for algorithm training and 30% for algorithm testing. The results are introduced and discussed below.

4.2.2.1 Classifier performance on top-two attributes

doi:10.20944/preprints202108.0471.v1

Framingham attributes sysBP (score=14.15) and BMI (score=13.55) and South African Hearth attributes Tobacco (score=15.70) and Age (score=15.39) were used at this stage. Table 7 lists the results from the analysis of classifier performance.

Table 7. Train-test set performance evaluation of ML classifiers on the top-two attributes from the Framingham and the South African Hearth datasets

Dataset	Predictive Model	Performance evaluation metrics				
		% Accuracy	% Precision	% Recall	% f1-score	% roc_auc
Framingham	AdaBoost Classifier	66.91	62.02	58.91	60.43	65.91
	CatBoost Classifier	73.60	71.90	63.09	67.21	72.29
	Decision Tree Classifier	64.16	58.68	55.57	57.08	63.09
	GradientBoosting Classifier	75.81	73.68	67.83	70.63	74.81
	KNeighbors Classifier	72.04	68.38	64.76	66.52	71.14
	LGBM Classifier	77.84	74.61	73.26	73.93	72.27
	Logistic Regression	65.65	63.88	45.82	53.37	63.18
	Random Forest Classifier	75.75	72.61	69.78	71.16	75.00
	Support Vector Classification	64.99	65.64	38.58	48.60	61.71
	XGBRF Classifier	75.63	73.13	68.25	70.61	74.71
South African Hearth	AdaBoost Classifier	65.47	51.52	34.69	41.46	58.46
	CatBoost Classifier	71.22	63.64	42.86	51.22	64.76
	Decision Tree Classifier	71.94	63.16	48.98	55.17	66.71
	GradientBoosting Classifier	63.31	47.62	40.82	43.96	58.19
	KNeighbors Classifier	65.47	52.00	26.53	35.14	56.60
	LGBM Classifier	67.63	54.76	46.94	50.55	62.91
	Logistic Regression	73.38	71.43	40.82	51.95	65.96
	Random Forest Classifier	67.63	55.26	42.86	48.28	61.98
	Support Vector Classification	71.22	80.00	24.49	37.50	60.58
	XGBRF Classifier	72.66	68.97	40.82	51.28	65.41

As shown in Table 7, LGBM classifier achieved the best performance on the Framingham dataset in terms of accuracy, precision, recall, and f1-score. On the South African Hearth dataset, Decision Tree classifier outperformed the other algorithms in terms of recall, f1-score, and roc-auc. Conversely, Decision Tree Classifier proved to be the lowest-performing algorithm on the Framingham dataset, with performance scores of 64.16% in accuracy and 58.68% in precision. As regards the South African Hearth dataset, GradientBoosting Classifier exhibited the lowest scores with an accuracy performance of 63.31% and a precision performance of 47.62%. Overall, in top-two attribute classifications, classifiers exhibit good performance in both accuracy and precision.

4.2.2.2 Classifier performance on top-four attributes

In the four-attribute classification analysis, Framingham attributes included sysBP (score= 14.15), BMI (score=13.55), Age (score=12.73), and totChol (score=12.69), whereas South African Hearth dataset attributes included Tobacco (score=15.70), Age

doi:10.20944/preprints202108.0471.v1

(score=15.39), Ldl (score=13.29), and Adiposity (score= 11.73). Figure 5 depicts the results from the analysis.

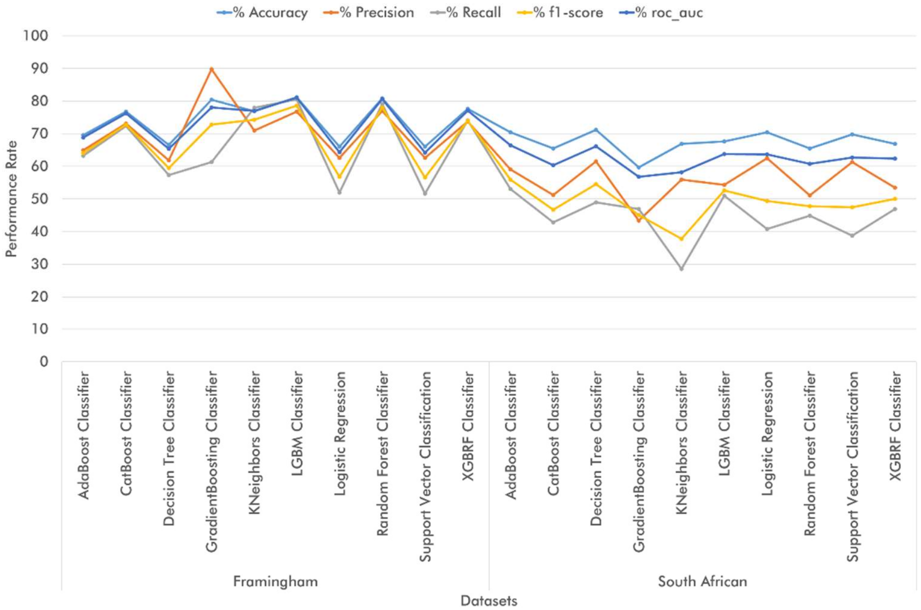


Figure 5. Train-test set performance of ten ML classifiers on the top-four attributes from the Framingham and South African Hearth datasets.

According to Figure 5, in the four-attribute classification, LGBM classifiers and Decision Tree Classifier exhibited the best performance in terms of accuracy (81.18% and 71.22%, respectively) on the Framingham and the South African Hearth datasets, respectively. As for precision, GradientBoosting classifier (89.80%) and Logistic Regression (62.50%) proved to be the best-performing algorithms on the Framingham and the South African Hearth datasets, respectively. Conversely, the most underperforming algorithms in terms on accuracy included Support Vector Classification (66.01%) on the Framingham dataset and GradientBoosting Classifier (59.71%) on the South African Hearth dataset. Decision Tree Classifier and GradientBoosting Classifier exhibited the lowest precision performance on the Framingham and the South African Hearth datasets, respectively, with values of 61.86% and 43.40% each. We concluded in this step that the classifiers performed better on the Framingham dataset across the five metrics, whereas in the South African Hearth dataset, favorable classifier behavior was observed only in terms of accuracy and roc-auc metrics. Also, we concluded that selecting four attributes does not considerably increase classifier performance in terms of accuracy and precision.

4.3 *Results of k-fold cross-validation for classifier performance on top two and four attributes*

As previously mentioned, we also relied on 10-fold cross-validation to validate the performance of the ML classifiers on the top two and four attributes of each dataset. The results of the cross-validation analyses are discussed below.

4.3.1 Medical diagnostic datasets

In this section, we discuss our results on the performance analysis of the LM classifiers when using k-fold cross-validation. The classifiers were applied on the top two and four attributes on the Cleveland and Faisalabad datasets.

doi:10.20944/preprints202108.0471.v1

4.3.1.1 Classifier performance on top-two attributes

The selected attributes from the Cleveland dataset included cp (score=13.55) and thalach (score=12.52), whereas serum creatinine (score=20.15) and ejection fraction (score=17.52) were chosen from the Faisalabad dataset. Table 8 lists the obtained results on the performance of the ten classifiers.

Table 8. k-fold cross-validation performance analysis of ML classifiers on the top-two attributes from the Cleveland and Faisalabad datasets

Dataset	Predictive Model	Performance evaluation metrics				
		% Accuracy	**% Precision**	**% Recall**	**% f1-score**	**% roc_auc**
Cleveland	AdaBoost Classifier	69.96	73.39	71.43	72.12	73.31
	CatBoost Classifier	74.25	76.96	75.62	76.14	79.36
	Decision Tree Classifier	76.22	78.00	**79.85**	78.52	77.87
	GradientBoosting Classifier	68.31	72.90	67.87	69.98	69.12
	KNeighbors Classifier	69.65	74.48	67.10	70.11	74.32
	LGBM Classifier	70.31	72.13	75.07	73.33	71.02
	Logistic Regression	**77.22**	**79.95**	78.64	**78.96**	79.76
	Random Forest Classifier	69.31	71.75	73.38	72.21	72.92
	Support Vector Classification	76.22	78.69	76.76	77.54	**80.25**
	XGBRF Classifier	75.23	77.02	78.05	77.40	78.44
Faisalabad	AdaBoost Classifier	70.60	60.46	40.44	46.22	68.88
	CatBoost Classifier	**76.28**	68.24	52.11	**58.55**	79.23
	Decision Tree Classifier	74.93	64.06	55.44	58.20	78.19
	GradientBoosting Classifier	71.91	60.11	46.78	51.48	76.78
	KNeighbors Classifier	74.26	62.70	49.22	53.76	77.87
	LGBM Classifier	73.23	60.14	54.33	56.17	77.70
	Logistic Regression	74.59	**76.67**	34.33	45.57	76.63
	Random Forest Classifier	72.24	56.67	**57.11**	56.08	80.42
	Support Vector Classification	71.57	71.33	24.11	34.22	76.33
	XGBRF Classifier	75.61	64.92	54.22	58.04	**81.18**

In the two attribute classification with cross-validation, Logistic Regression achieved the greatest performance in accuracy, precision, and f1-score on the Cleveland dataset, whereas CatBoost yielded the best performance in terms of accuracy and f1-score on the Faisalabad dataset. On the other hand, the lowest-performing algorithms on the Cleveland dataset included GradientBoosting in terms of accuracy, f1-score, and roc_auc, and Random Forest Classifier (71.75%) in terms of precision. On the Faisalabad dataset, AdaBoost Classifier yielded the lowest results in accuracy and roc_auc, and Random Forest Classifier (56.67%) exhibited the poorest precision performance. We concluded from this step that the k-fold cross-validation approach increases classifier performance in precision and roc-auc metrics in a two-attribute classification.

4.3.1.2 Classifier performance on top-four attributes

For the top-four attribute classification analysis, the selected Cleveland attributes included cp (score=13.55), thalach (score=12.52), ca (score=11.70), and oldpeak (score=10.90).

The selected Faisalabad dataset attributes comprised serum creatinine (score= 20.15), ejection fraction (score=17.52), age (score=14.56), and platelets (score=12.83). Figure 6 depicts a graphic representation of the analysis results.

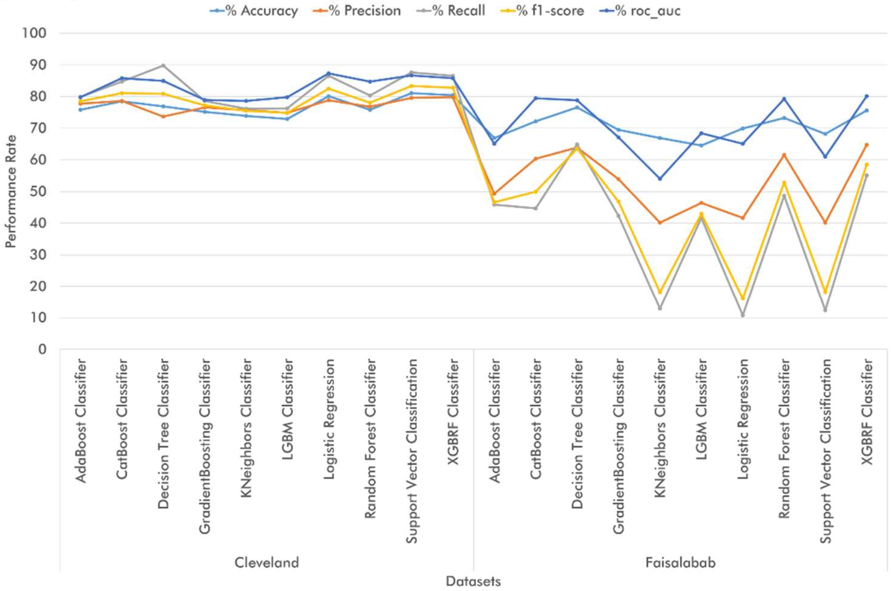


Figure 6. K-fold cross validation performance of ten ML classifiers on the top-four features of the Cleveland and the Faisalabad datasets

As can be observed from Figure 6, classifiers Support Vector Classification and Decision Tree yielded the best accuracy results on the Cleveland and Faisalabad datasets, respectively, with values of 81.16% and 76.59% each. The highest precision was achieved by XGBRF classifier on both datasets with values of 79.77% and 64.74%, respectively. On the other hand, the algorithm achieving the lowest performance in accuracy was LGBM Classifier with values of 72.94% on the Cleveland dataset and 64.59 % on the Faisalabad dataset. In terms of precision, Decision Tree Classifier proved to be the lowest-performing algorithm (73.73%) on the Cleveland dataset, whereas KNeighbors Classifier and Support Vector Classification yielded the lowest results (40.17%) on the Faisalabad dataset. Overall, the classifiers exhibited better performance on the Cleveland dataset than in the Faisalabad dataset, with an adequate behavior above 75%. On the Faisalabad dataset, the classifiers showed adequate performance only in accuracy and roc-auc and poor performance in terms of recall and f1-score. We concluded in this step that k-fold cross-validation increases classifier performance in the four-attribute classification analysis in accuracy and roc-auc metrics.

4.3.2 Medical prediction datasets

In this section, we discuss our results on the performance analysis of the LM classifiers when using k-fold cross-validation. The classifiers were applied on the top two and four attributes on the Framingham and the South African Hearth datasets.

4.3.2.1 Classifier performance on top-two attributes

Selected Framingham attributes included sysBP (score=14.15) and BMI (score=13.55), whereas selected South African Hearth attributes comprised Tobacco (score=15.70) and Age (score=15.39). Table 9 introduces the results of the classifier performance analysis using cross-validation.

Table 9. K-fold cross-validation analysis of ten ML classifiers on top-two Framingham and South African dataset attributes

Dataset	Predictive Model	Performance evaluation metrics				
		% Accuracy	% Precision	% Recall	% f1-score	% roc_auc
Framingham	AdaBoost Classifier	66.67	64.28	56.21	59.87	73.40
	CatBoost Classifier	72.60	71.86	62.66	66.69	80.67
	Decision Tree Classifier	63.57	61.67	50.20	54.93	64.92
	GradientBoosting Classifier	76.06	**75.17**	68.19	71.18	82.64
	KNeighbors Classifier	71.47	68.63	65.89	67.19	79.71
	LGBM Classifier	**77.42**	72.24	**74.52**	**74.03**	**84.74**
	Logistic Regression	63.15	63.08	41.57	50.07	66.13
	Random Forest Classifier	75.84	72.94	72.10	72.34	84.17
	Support Vector Classification	62.80	64.76	35.81	46.07	65.58
	XGBRF Classifier	75.66	73.83	69.84	71.63	83.54
South African Hearth	AdaBoost Classifier	65.58	49.19	40.62	44.05	68.23
	CatBoost Classifier	66.68	53.86	40.62	45.67	69.26
	Decision Tree Classifier	**72.51**	62.16	**52.50**	**55.20**	71.61
	GradientBoosting Classifier	61.26	43.52	42.50	42.50	62.86
	KNeighbors Classifier	66.23	54.28	31.87	39.29	65.89
	LGBM Classifier	63.64	46.49	46.25	45.72	66.22
	Logistic Regression	70.55	63.56	38.75	47.20	**74.33**
	Random Forest Classifier	61.70	43.46	37.50	39.77	64.09
	Support Vector Classification	68.83	**66.31**	23.12	33.64	72.29
	XGBRF Classifier	65.60	50.42	40.62	43.89	68.80

According to Table 9, in the top-two attribute classification, LGBM classifier yielded the best results for accuracy, recall, f-1 score, and roc_auc, whereas Decision Tree achieved the best performance on the South African Hearth dataset in terms of accuracy, recall, and f1-score. As regards precision, GradientBoosting outperformed the other nine classifiers on the Framingham dataset with a value of 75.17%, whereas Support Vector Classification achieved the best precision on the South African Hearth dataset with a value of 66.31%. On the other hand, on the Framingham dataset, Support Vector Classification was the lowest-performing algorithm in terms of accuracy, recall, and f1-score, while Decision Tree Classifier underperformed in terms of precision (61.67%). On the South African Hearth dataset, GradientBoosting Classifier was the lowest-performing algorithm in accuracy and roc-auc, whereas Random Forest Classifier exhibited the lowest precision performance (43.46%). We also observed that classifiers performed better on the Framingham dataset than on the South African Hearth dataset across the five metrics, although there was improved behavior on the South African Hearth dataset if compared to the previous analyses.

4.3.2.2 Classifier performance on top-four attributes

For the top-four attribute performance analysis, the selected attributes included sysBP (score= 14.15), BMI (score=13.55), Age (score=12.73), and totChol (score=12.69) for the Framingham dataset. On the other hand, Tobacco (score=15.70), Age (score=15.39), Ldl (score=13.29), and Adiposity (score=11.73) were selected on the South African Hearth dataset. Figure 7 shows the results from the analysis.

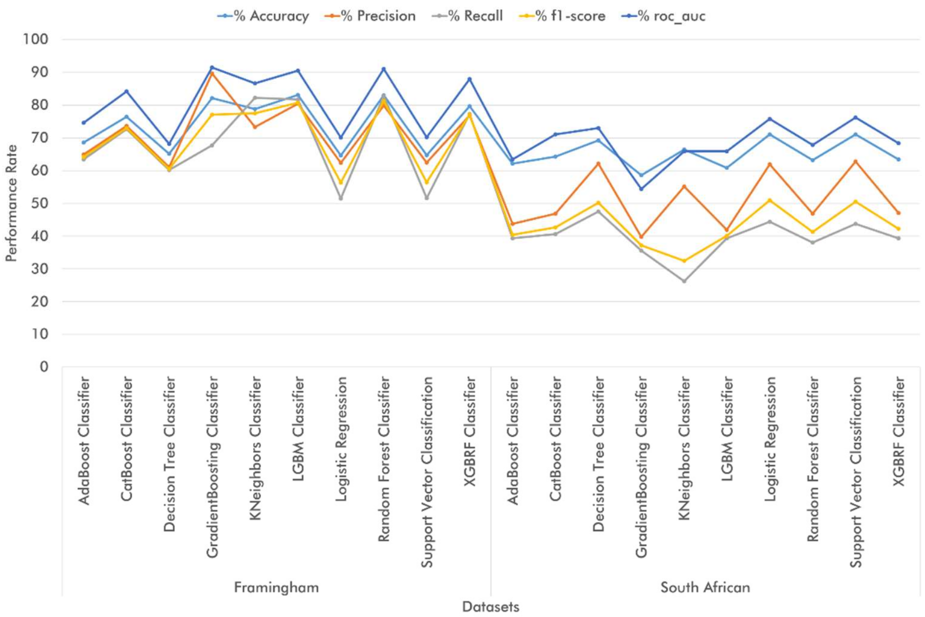


Figure 7. k-fold cross-validation performance of ten ML classifiers on top-four attributes from the Framingham and South African Hearth datasets

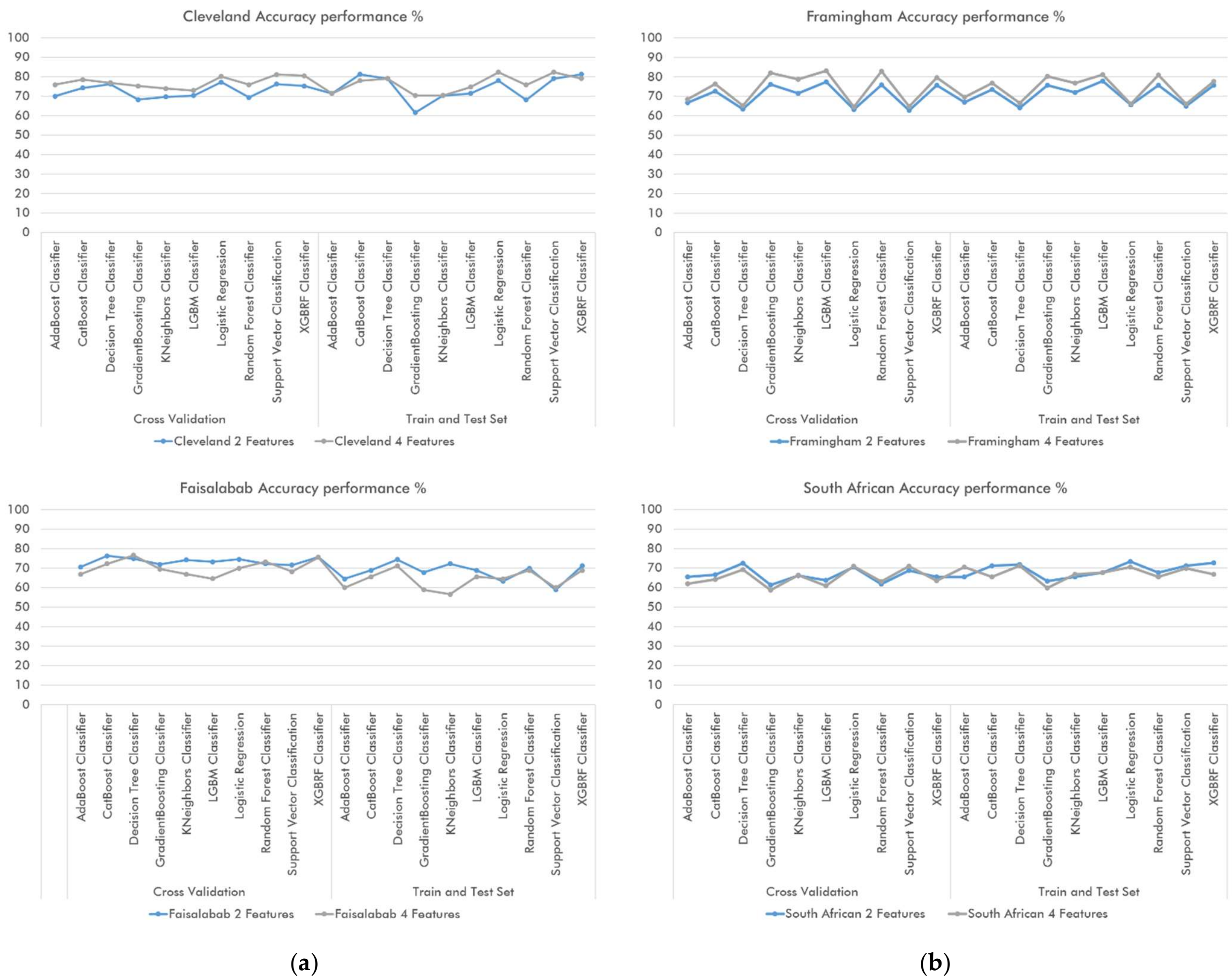
As depicted in Figure 7, the highest accuracy was achieved by LGBM classifier (83.10%) on the Framingham dataset, and by both Support Vector Classification and Logistic Regression (70.99% respectively) on the South African Hearth dataset. Regarding precision, the highest performance was exhibited by GradientBoosting classifier (89.60%) on the Framingham dataset and Support Vector Classification (62.79%) on the South African Hearth dataset. Conversely, the lowest accuracy performance was exhibited by Logistic Regression (64.61%) on the Framingham dataset and GradientBoosting Classifier (58.64%) on the South African Hearth dataset. The lowest precision was recorded by Decision Tree Classifier (61.05%) on the Framingham dataset and GradientBoosting Classifier (39.77%) on the South African Hearth dataset. Overall, the Framingham dataset allowed for better classifier performance across the five metrics, while the South African Hearth dataset exhibits better performance than in previous analyses.

4.4 Most important dataset attributes

The importance of this research lies in finding the best precision and accuracy results from the ten ML classifiers to identify the top-two and top-four attributes for CVD detection and prevention. At this stage, we compared the results obtained from all the previous performance analyses. When comparing the accuracy metrics, we found that in both the two-attribute and the four attribute classifications, the ML classifiers performed adequately on all the CVD diagnostic and prediction datasets using both performance analysis techniques: train-and-test split and k-fold cross-validation. Specifically, when working with medical diagnosis datasets, the ten classifiers performed better when applied on the top-four attributes of the Cleveland dataset and the top-two attributes of the Faisalabad dataset. Conversely, when working with medical prediction datasets, we observed overall better classifier performance on the top-four attributes from the Framingham dataset and the top-two attributes from the South African Hearth dataset.

doi:10.20944/preprints202108.0471.v1

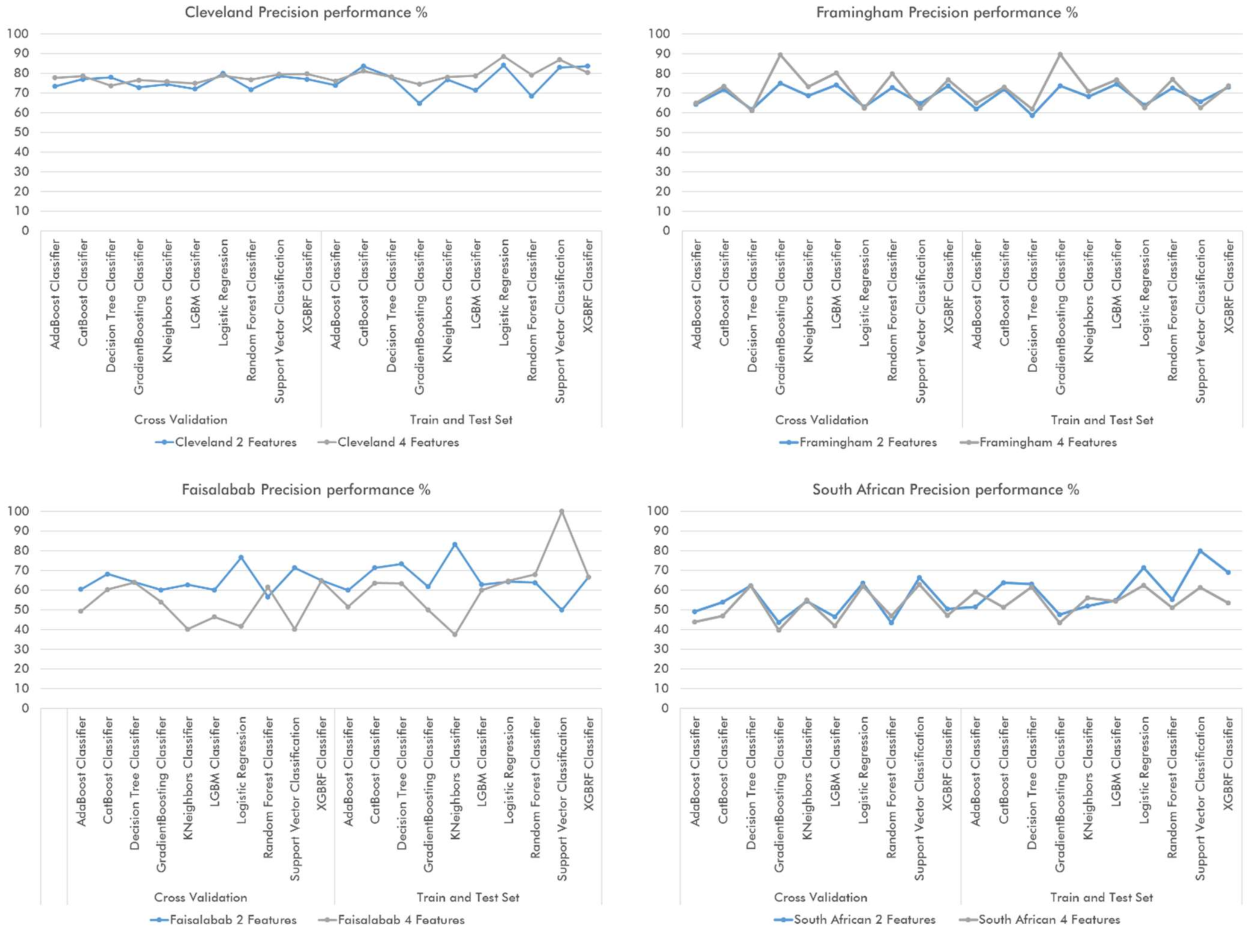
As for the best validation technique, we found that it is feasible to rely on both train-and-test set validation and k-fold cross validation to obtain adequate classifier performance on the Cleveland, South African Hearth, and Framingham datasets. However, on the Faisalabad dataset, ML classifiers performed better when using k-fold cross-validation than during the train-and-test split tests.



(a) (b)

Figure 8. Comparison of ML classifier accuracy performance using k-fold cross-validation and train-test split on top-two and top-four attributes from (**a**) Medical diagnostic datasets; (**b**) Medical prediction datasets

As regards precision metrics, Figure 9 shows that adequate classifier performance was achieved in all top-two and top-four attribute classifications on the Cleveland, Framingham, and South African Hearth datasets. Additionally, we found that when working with medical diagnosis datasets, the ML classifiers performed better in terms of precision on the Cleveland dataset during the top-four attribute classifications and on the Faisalabad dataset during the top-two attribute classifications. On the other hand, when dealing with medical prediction datasets, we achieved better classifier performance results on the Framingham dataset (top-four attribute classification) and the South African dataset (top-two classification). As for the best-evaluated technique, train-and-test set validation worked best on the Cleveland dataset, whereas on the Faisalabad and Framingham datasets, some algorithms performed better when using the train-and-test set technique, while some others worked best when using k-fold cross-validation. Regarding the South African Hearth dataset, it is feasible to use both train-test split and k-fold cross-validation, since the ML classifiers exhibited adequate performance with both techniques.



(**a**) (**b**)

Figure 9. Comparison of ML classifier precision performance using k-fold cross-validation and train-test split on top-two and top-four attributes from (**a**) Medical diagnostic datasets; (**b**) Medical prediction datasets

As a result of the previous analysis, we managed to identify the main attributes for CVD diagnosis across the four datasets. On the Cleveland database, such attributes include cp (Chest Pain Type), thalach (maximal heart rate), ca (number of vessels colored by fluoroscopy), and oldpeak (exercise relative to rest). In the top-two attribute classification, CatBoost Classifier and XGBRF Classifier achieved the best accuracy (81.32%), Logistic Regression yielded the best precision performance (84.09%), Decision Tree Classifier outperformed in terms of recall (86.00%), and CatBoost Classifier and XGBRF Classifier achieved the best performance results in terms of f1-score and roc-auc, respectively (82.83% and 81.24%, respectively). On the other hand, when using k-fold cross-validation, Logistic Regression exhibited the best performance in accuracy (77.22%), precision (79.95%), and f1-score (78.96%), whereas Decision Tree Classifier showed the best results in terms of recall (79.85%), and Support Vector Classification yielded the best performance in roc-auc (80.25%).

As regards the top-four classification of Cleveland attributes using train-test split, Logistic Regression and Support Vector Classification yielded the highest accuracy (82.42%), whereas Logistic Regression alone outperformed the other algorithms in terms of precision (88.64%). On the other hand, the best-performing classifiers in recall, f1-score, and roc-auc were Decision Tree Classifier (86.00%), Support Vector Classification (83.33%), and Logistic Regression (82.9%), respectively. Finally, when using k-fold cross-validation, Support Vector Classification exhibited the highest classification accuracy (81.16%), XGBRF Classifier yielded the best results in terms of precision (79.77%), Decision Tree Classifier was the best-performing algorithm in recall (89.78%), Support Vector Classification showed the best results in f1-score (83.33%), and Logistic Regression was the best-performing algorithm in roc-auc (87.34%).

In the Faisalabad dataset, the main attributes identified included serum creatinine, ejection fraction, patient age, and platelets. In the top-two attribute classification using the test-train split technique, the best-performing classifiers were as follows: Decision Tree Classifier in accuracy (74.44%), f1-score (65.67%), and roc-auc (72.18%), KNeighbors Classifier in precision (72.22%), and Random Forest Classifier in recall (62.16%). Conversely, when relying on k-fold cross-validation, CatBoost Classifier exhibited the best results in accuracy and f1-score (76.28% and 58.55%,respectively), Logistic Regression yielded the highest precision (76.67%), Random Forest Classifier outperformed the other classifiers in terms of recall (57.11%), and XGBRF Classifier showed the best performance in roc-auc (81.18%). In the top-four classification of Faisalabad attributes using the train-test split technique, Decision Tree Classifier proved to be the best-performing algorithm as regards accuracy (71.11%), recall (70.27%), f1-score (66.67%), and roc-auc (70.98%), whereas Support Vector Classification exhibited the highest precision performance (100.00%). On the other hand, during k-fold cross validation, the best classification performance was exhibited by Decision Tree Classifier in terms of accuracy (76.59%), recall (64.89%), and f1-score (63.59%), and by XGBRF Classifier in terms of precision (64.74%) and roc-auc (80.17%).

As regards the two CVD medical prediction datasets, the main attributes identified in the Framingham dataset included sysBP (systolic blood pressure), BMI (Body Mass Index), age (age at exam time), and totChol (total cholesterol). In the top-two attribute classification using the train-test split technique, LGBM Classifier proved to be the best-performing classifier across the five metrics: accuracy (77.84%), precision (74.61%), recall (73.26%), f1-score (73.93%), and roc-auc (77.27%). On the other hand, when using k-fold cross-validation, LGBM Classifier exhibited the best performance in accuracy (77.42%), recall (74.52%), f1-score (74.03%), and roc-auc (84.74%), whereas GradientBoosting Classifier showed the best precision results (75.17%). As regards the top-four attribute classification with the train-test split technique, LGBM Classifier outperformed the other classifiers in terms of accuracy (81.18%), recall (80.50%), f1-score (78.59%), and roc-auc (81.10%), while the highest precision was achieved by GradientBoosting Classifier (89.80%). On the other hand, when using k-fold cross-validation, the best performing classifiers included LGBM Classifier in terms of accuracy (83.10%), GradientBoosting Classifier in terms of precision (89.60%) and roc-auc (91.41%), and Random Forest Classifier in recall and f1-score (82.74% and 81.21%, respectively).

In the South African Hearth dataset, the main attributes included tobacco (cumulative tobacco), age (age at the exam), LDL (low-density lipoprotein cholesterol), and adiposity. In the top-two attribute classification using train-test split, the best-performing classifiers proved to be Logistic Regression in accuracy (73.38%), Support Vector Classification in precision (80.00%), and Decision Tree Classifier in terms of recall (48.98%), f1-score (55.17%), and roc-auc (66.71%). Conversely, when relying on k-fold cross-validation, Decision Tree Classifier yielded the highest accuracy (71.22%), Logistic Regression outperformed the other classifiers in precision (62.50%), and AdaBoost Classifier exhibited

doi:10.20944/preprints202108.0471.v1

the best performance in recall (53.06%), f1-score (55.91%), and roc-auc (66.53%). Regarding the top-four attribute classifications, the best-performing classifiers with the train-test split technique were Decision Tree Classifier in accuracy (71.22%), Logistic Regression in precision (62.50%), and AdaBoost Classifier in recall (53.06%), f1-score (55.91%), and roc-auc (66.53%). On the other hand, when relying on k-fold cross-validation, both Logistic Regression and Support Vector Classification achieved the highest accuracy (70.99%), whereas Support Vector Classification itself exhibited the best performance in terms of precision (62.79%). Decision Tree yielded the best results in recall (47.50%), Logistic Regression in f1-score (50.98%), and Support Vector Classification in roc-auc (76.17%).

From this discussion of the results, we concluded that the ten studied ML classifiers performed adequately in the classification of top-two and top-four dataset attributes. Hence, efforts in predicting and/or diagnosing CVD with said features will yield the expected results.

Table 10. Main risk factors for CVD diagnosis and prediction

Method	**Dataset**	**Best Rated Features**	**Feature Description**
Diagnosis	Cleveland	cp	Chest Pain Type
		thalach	Maximum heart rate
		ca	Number of vessels colored by fluoroscopy
		oldpeak	Exercise induced ST segment depression
	Faisalabad	Serum creatinine	Level of creatinine in the blood
		Ejection fraction	Percentage of blood leaving the heart at each hear beat
		Age	Patient age
		Platelets	Platelets in the blood
Prediction	Framingham	sysBP	Systolic blood pressure
		BMI	Body Mass Index
		Age	Age at exam time
		totChol	Total Cholesterol
	South African Hearth	Tobacco	Cumulative tobacco
		Age	Age at onset
		Ldl	Low-density lipoprotein cholesterol
		Adiposity	Adiposity

Of the variables identified, age in the Faisalabad, Framingham, and South African Hearth datasets is an important risk factor for any CVD. As regards heart rate (found in the Cleveland dataset as thalach), normal ranges of pulse per minute (bpm) should be monitored. On the other hand, blood pressure is known to trigger all types of CVDs. It refers to the force exerted against the walls of the arteries as the heart pumps blood to the body. In this sense, systolic pressure range, found in the Framingham dataset, should be properly monitored, especially among patients suffering from hypertension. Levels of blood cholesterol in the body are measured with cholesterol tests, which determine the amount of each type of cholesterol and certain fats in the body. LDL cholesterol, or bad cholesterol, (attribute from the South African Hearth dataset) is a major CVD risk factor, since it causes plaque buildup in the arteries, thus reducing blood flow. Similarly, total blood cholesterol levels – attribute found in the Framingham dataset – must be monitored in all CVD diagnosis and detection efforts.

Regarding Cleveland dataset attributes, coronary angiography (ca) is a special procedure that uses contrast dyes and X-rays to see how blood flows in the arteries in the heart, thus showing whether many of the coronary arteries are blocked or narrowed due to fatty plaques and how serious it may be. Coronary angiography thus allows monitoring the development of CVDs such as heart disease, arterial disease, and coronary artery disease. As for cp, ECGs (i.e. graphical representation of the electrical forces working on the heart) allow monitoring the cardiac cycle of pumping and filling in a known pattern of changing electrical pulses that accurately reflect the action of the heart. ECGs are performed by collecting the pulses through electrodes attached to the surface of the body. Hence, ECGs help identify CVDs such as heart failure, arrhythmia, heart disease, and arterial or coronary artery disease. Finally, exercise induced ST segment depression (oldpeak) can be monitored via stress tests (i.e. ergometry) to examine how the heart functions during physical activity to prevent the development of CVDs, such as heart failure, heart disease, arterial and coronary artery disease. These attributes are the most important for correct CVD prediction and diagnosis. Similarly, we identified other important attributes, such as tobacco and blood platelet count. On the one hand, nicotine in the body must be monitored among both smokers and non-smokers by modifying patient lifestyle, whereas high blood platelet counts may be an indicator of CVD. Finally, as discussed by Davide Chicco, et al., other key attributes for CVD detection and diagnosis include ejection fraction (i.e. percentage of blood leaving the heart at each heart beat) and serum creatinine (i.e. level of blood creatinine), whose abnormal levels are usually observed among diabetic patients, kidney disease sufferers, and patients with high blood pressure.

5. Conclusions and future directions

MLAs play a key role in healthcare services by analyzing medical data for disease diagnosis. CVDs are a critical medical problem for healthcare professionals and researchers. To approach this issue, we have conducted a dataset study with clinical data of CVDs to identify the main risk factors that influence CVD development using MLAs. First, we relied on Random Forest to identify and select the top four attributes in each dataset to improve the training and testing of the algorithms. Then, we analyzed the classification performance of the predictive models on four datasets and using the train-test split technique and k-fold validation. Finally, we compared the obtained results. Performance metrics comprised accuracy, precision, recall, f1-score, and roc-auc, whereas the analyzed datasets included the Cleveland and the Faisalabad datasets – for CVD diagnosis - and the Framingham and South African datasets – for CVD prediction.

We compared the performance of the ten algorithms in two-attribute and four attribute classifications. We found adequate and consistent algorithm performance in the top-two attribute classifications when using both train-test split and k-fold cross-validation techniques. Our results demonstrate that, in most of the datasets, age, heart rate, and blood pressure are the most significant CVD attributes, followed by weight, cholesterol, tobacco, serum creatinine, ejection fraction, chest pain type, number of vessels, platelet count, and adiposity. All these attributes stood out in the prediction performance analysis and thus have an impact on CVD detection. We conclude that the main attributes are suitable for follow-up in the preventive diagnosis of CVD and for timely and accurate treatment when necessary. Moreover, follow-up strategies can significantly improve the chances of long-term survival of patients with CVDs, thus saving millions of lives.

As regards our suggestions for future work, we recommend replicating our study in other medical databases to contribute to current prevention and diagnosis efforts of other diseases, such as diabetes and breast cancer. Similarly the risk factors identified in this study can be used in the development of mobile applications for heart disease monitoring in which patient clinical data are automatically recorded and further analyzed by healthcare professionals for a correct diagnosis. Finally, an attractive proposal would be to build a large database with the main attributes detected from various sources: clinical datasets, wearable devices, mobile applications, and medical records. This outcome could be achieved by relying on big data techniques and will contribute to current efforts to improve our quality of life.

Acknowledgments: This work was sponsored by the Council for Scientific Research and Technological Development in Veracruz (COVEICYDET). We thank COVEICYDET for supporting this work through project name Prevention and Early Detection of Cardiovascular Diseases (Arrhythmias & Tachycardias) using Machine Learning Techniques, Big Data, and the Internet of Things (identifier number 12 1806). L.G.N. acknowledges CONACyT for supporting their postdoctoral stays at Tecnológico Nacional de México. J.L.S.C. thanks CONACYT for the research position granted under the Cátedras-CONACYT program.

Conflicts of Interest: The authors declare no potential conflicts of interest with respect to the publication of this research.

References

1. "The top 10 causes of death." https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed May 31, 2021).
2. "What is CVD? - World Heart Federation." https://world-heart-federation.org/what-is-cvd/ (accessed May 31, 2021).
3. A. K. Pandey, P. Pandey, K. L. Jaiswal, and A. K. Sen, "A Heart Disease Prediction Model using Decision Tree," 2013. Accessed: May 27, 2021. [Online]. Available: www.iosrjournals.orgwww.iosrjournals.org.
4. O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," Expert Syst. Appl., vol. 68, pp. 163–172, Feb. 2017, doi: 10.1016/j.eswa.2016.10.020.
5. M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Telemat. Informatics, vol. 36, pp. 82–93, Mar. 2019, doi: 10.1016/j.tele.2018.11.007.
6. I. D. Mienye, Y. Sun, and Z. Wang, "Improved sparse autoencoder based artificial neural network approach for prediction of heart disease," Informatics Med. Unlocked, vol. 18, p. 100307, Jan. 2020, doi: 10.1016/j.imu.2020.100307.
7. D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," BMC Med. Inform. Decis. Mak., vol. 20, no. 1, pp. 1–16, Feb. 2020, doi: 10.1186/s12911-020-1023-5.
8. S. I. Ayon, M. M. Islam, and M. R. Hossain, "Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques," IETE J. Res., 2020, doi: 10.1080/03772063.2020.1713916.
9. S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
10. D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction Using Machine Learning," in Lecture Notes in Electrical Engineering, Nov. 2020, vol. 708, no. 6, pp. 603–609, doi: 10.1007/978-981-15-8685-9_63.
11. A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," Neural Comput. Appl., vol. 29, no. 10, pp. 685–693, May 2018, doi: 10.1007/s00521-016-2604-1.
12. M. C. Belavagi and B. Muniyal, "Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection," in Procedia Computer Science, Jan. 2016, vol. 89, pp. 117–123, doi: 10.1016/j.procs.2016.06.016.
13. K. Deepika and S. Seema, "Predictive analytics to prevent and control chronic diseases," in Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, Apr. 2017, pp. 381–386, doi: 10.1109/ICATCCT.2016.7912028.
14. S. A. Tiwaskar, R. Gosavi, R. Dubey, S. Jadhav, and K. Iyer, "Comparison of Prediction Models for Heart Failure Risk: A Clinical Perspective," Jul. 2018, doi: 10.1109/ICCUBEA.2018.8697509.
15. J. Nahar, T. Imam, K. S. Tickle, and Y. P. P. Chen, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach," Expert Syst. Appl., vol. 40, no. 1, pp. 96–104, Jan. 2013, doi: 10.1016/j.eswa.2012.07.032.
16. T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Survival analysis of heart failure patients: A case study," PLoS One, vol. 12, no. 7, p. e0181001, Jul. 2017, doi: 10.1371/journal.pone.0181001.
17. R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," Am. J. Cardiol., vol. 64, no. 5, pp. 304–310, Aug. 1989, doi: 10.1016/0002-9149(89)90524-9.

18. P. Shimpi, S. Shah, M. Shroff, and A. Godbole, "A machine learning approach for the classification of cardiac arrhythmia," in Proceedings of the International Conference on Computing Methodologies and Communication, ICCMC 2017, Feb. 2018, vol. 2018-Janua, pp. 603–607, doi: 10.1109/ICCMC.2017.8282537.
19. K. A. K. Niazi, S. A. Khan, A. Shaukat, and M. Akhtar, "Identifying best feature subset for cardiac arrhythmia classification," in Proceedings of the 2015 Science and Information Conference, SAI 2015, Sep. 2015, pp. 494–499, doi: 10.1109/SAI.2015.7237188.
20. B. Fida, M. Nazir, N. Naveed, and S. Akram, "Heart disease classification ensemble optimization using Genetic algorithm," in Proceedings of the 14th IEEE International Multitopic Conference 2011, INMIC 2011, 2011, pp. 19–24, doi: 10.1109/INMIC.2011.6151471.
21. N. Singh and P. Singh, "Cardiac arrhythmia classification using machine learning techniques," in Lecture Notes in Electrical Engineering, vol. 478, Springer Verlag, 2019, pp. 469–480.
22. T. Soman and P. O. Bobbie, "Classification of Arrhythmia Using Machine Learning Techniques."
23. S. Kodati, R. Vivekanandam, and G. Ravi, "Comparative analysis of clustering algorithms with heart disease datasets using data mining weka tool," in Advances in Intelligent Systems and Computing, 2019, vol. 900, pp. 111–117, doi: 10.1007/978-981-13-3600-3_11.
24. A. U. Haq et al., "Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data," Sensors (Switzerland), vol. 20, no. 9, p. 2649, May 2020, doi: 10.3390/s20092649.
25. S. R. Ghosh and S. Waheed, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis," 2016. Accessed: May 27, 2021. [Online]. Available: www.ijcsit.com.
26. S. Mishra, P. K. Mallick, H. K. Tripathy, A. K. Bhoi, and A. González-Briones, "Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier," Appl. Sci., vol. 10, no. 22, pp. 1–35, Nov. 2020, doi: 10.3390/app10228137.
27. K. J. Danjuma, "Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients," Apr. 2015, Accessed: May 27, 2021. [Online]. Available: http://arxiv.org/abs/1504.04646.
28. Y. Li and Z. Chen, "Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction Performance Evaluation of Machine Learning Methods for Breast," Cancer Predict. Appl. Comput. Math., vol. 7, no. 4, pp. 212–216, 2018, doi: 10.11648/j.acm.20180704.15.
29. A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "UCI Machine Learning Repository: Heart Disease Data Set," UCI, 2019. https://archive.ics.uci.edu/ml/datasets/heart+disease (accessed Jun. 01, 2021).
30. "FraminghamHeartAttackPrediction | Kaggle," 2021. https://www.kaggle.com/apoorvak141619/framinghamheartattackprediction/data?select=framingham.csv (accessed Jun. 01, 2021).
31. A. F. Ernández, J. L. Uengo, and J. D. Errac, "KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on)," J. Mult. Log. Soft Comput., vol. 17, pp. 255–287, 2011, Accessed: Jun. 01, 2021. [Online]. Available: https://sci2s.ugr.es/keel/dataset.php?cod=184#inicio.
32. R. Iyer, D. W. Hosmer, and S. Lemeshow, Applied Logistic Regression., vol. 40, no. 4. 1991.
33. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
34. K. R. Bhatele and S. S. Bhadauria, "Glioma segmentation and classification system based on proposed texture features extraction method and hybrid ensemble learning," Trait. du Signal, vol. 37, no. 6, pp. 989–1001, Dec. 2020, doi: 10.18280/TS.370611.
35. Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553. Nature Publishing Group, pp. 436–444, May 27, 2015, doi: 10.1038/nature14539.
36. M. J. Sorich, J. O. Miners, R. A. McKinnon, D. A. Winkler, F. R. Burden, and P. A. Smith, "Comparison of Linear and Nonlinear Classification Algorithms for the Prediction of Drug and Chemical Metabolism by Human UDP-Glucuronosyltransferase Isoforms," J. Chem. Inf. Comput. Sci., vol. 43, no. 6, pp. 2019–2024, Nov. 2003, doi: 10.1021/ci034108k.
37. B. Venkata Ramana, M. S. P. Babu, and N. . Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis," Int. J. Database Manag. Syst., vol. 3, no. 2, pp. 101–114, 2011, doi: 10.5121/ijdms.2011.3207.
38. G. Biau, B. Cadre, and L. Rouvière, "Accelerated gradient boosting," Mach. Learn., vol. 108, no. 6, pp. 971–992, Mar. 2019, doi: 10.1007/s10994-019-05787-1.
39. G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," 2017. Accessed: Jun. 01, 2021. [Online]. Available: https://github.com/Microsoft/LightGBM.
40. J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," J. Big Data, vol. 7, no. 1, pp. 1–45, Dec. 2020, doi: 10.1186/s40537-020-00369-8.
41. J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class AdaBoost *," 2009.
42. J. Brownlee, 00 ML Mastery - Understand You Data, Create Accurate Models and Work Projects End-to-End, vol. 91. 2017.