A Machine Learning Ensemble Classifier for Cardiovascular Disease Taxonomy

Oyetunde P. Oyelude and Rene V. Mayorga

***Abstract*—This Paper presents an application of Machine Learning in cardiology and the role of ensemble classifiers for Cardiovascular Disease (CVD) taxonomy. The dataset from Kaggle on CVD was used. Data was cleaned and 5 feature reduction techniques were investigated. Furthermore, a statistical unbiased ensemble feature reduction is proposed by imposing a unitary weight on intersecting features. Considering only 7 features, the Recurrent Feature Elimination and the proposed unbiased-ensemble feature reduction techniques were effective for reducing variables. Here, 6 feature reduction methods are considered. Hence, from each feature reduction method; the diverse selected features are then fed into a set of 5 independent ML techniques to compose a corresponding classifier. This ML approach in turn considers the 5 resultant classifiers and one additional proposed Ensemble Classifier based on those 5 classifiers. This proposed Ensemble Classifier consisted of: Multi-Layer Perceptron (MLP), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR) and k-Nearest Neighbor (KNN), classifiers. The output of the Machine Learning (ML) Classifiers approach is a classification/taxonomy to determine an individual with cardiovascular disease; or an individual that is free from cardiovascular disease. By considering the effective Recursive Feature Elimination method and the proposed Ensemble Classifier it was demonstrated that the body weight of an individual, systolic and diastolic blood pressure, cholesterol level, glucose level, level of physical activity, and the age are decisive in diagnosing the CVD condition of an individual. It is relevant to mention that a genetic feature was not available from the considered database; therefore, this potentially important factor was not considered in this study.**

***Index Terms*—Cardiovascular disease, classifiers, machine learning, taxonomy.**

I. Introduction

Due to the large data generated by various algorithms or devices, there is a need for storage and analysis of these data. However, not all the data generated are effective for analysis. Hence, the need to systematically trim the data. Feature reduction involves reducing the dimension of the features in a dataset for use in a learning algorithm. This step helps the algorithm reduce the effect of excessive data dimensions which affect the efficiency of machine learning and analysis of the relationships between data or features [1]. Thus, it is often applied to machine learning models. Feature reduction seems helpful, it is essential to choose a reasonable feature reduction method because all techniques do not guarantee a good feature selection when applied on a dataset.

In this Paper, 5 feature reduction methods are studied; and it also proposed an original ensemble feature reduction method. In previous works, researchers classified CVD, using a MLP for prediction purposes on a set of about 303 data points from University of California Irvine (UCI) Machine Learning repository Cleveland's coronary heart illness database [2]. Similarly, [3] other authors developed a Neural Network (NN) based intelligent system for predicting heart diseases using the same dataset based on the UCI repository. Also, these researchers did not reduce the features but mainly developed a predictive model such as Decision Trees, LR, Naïve Bayes Algorithm, RF, SVM, Gradient Boosted Trees, Deep Learning and MLP which turned out to perform better than the previous. The MLP had a better accuracy. Nirschl [4] used a different approach of hematoxylin and eosin stain technique to identify patients with clinical heart failure using Convolutional Neural Networks (CNN) Classifier was used to evaluate 209 patients.

Each of the 6 feature reduction methods is analyzed in conjunction with each one of 5 diverse/independent ML techniques. Here, it is also proposed to analyze each of the 6 considered features reduction methods with an original Ensemble Classifier based on those 5 Machine Learning techniques.

The proposed ML ensemble classifier can be used to leverage on the performance of various individual ML algorithms to make diagnosis on the heart condition of a subject. Therefore, the diagnosis is done based on the feature reduction techniques used on the data and the classifiers based on the ML techniques implemented.

This Paper effort can serve as a second opinion to doctors and further benefit doctors or healthcare workers working in remote locations of the world where they do not have sufficient resources to measure unneeded features to make a diagnosis, and attention paid to the essential features needed to make a diagnosis using machine learning.

II. Materials and Method

A. Data Collection

The CVD dataset used for this study was acquired from Kaggle [5]. Kaggle is a subsidiary of Google and one of the world's largest communities for data science and machine learning. The data was analyzed for outliers, and the classifiers were built using Python open-source software used for data science and machine learning. Furthermore, Microsoft Excel was used in plotting graphs used for result representation.

The dataset was split into 80:20. 80% of the analyzed dataset was used for training the Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF),

Manuscript received July 29, 2021; revised February 14, 2022; accepted February 14, 2022.

The authors are with the Faculty of Engineering and Applied Science, University of Regina, Canada (e-mail: oyetunde.oyelude@gmail.com).

doi: 10.18178/ijmlc.2022.12.6.1113

Multilayer Perceptron (MLP), k-Nearest Neighbor (KNN) and the ensemble classifiers; while 20% was for validation.

The configuration of the computing machine used for data processing and the classifier training is listed in Table I.

TABLE I: CONFIGURATION OF THE MACHINE USED FOR THIS STUDY

Machine Configuration	
Model	MacBook Pro
Processor	2.6 GHz 6-Core Intel Core i7
Storage	16 GB 2400 MHz DDR4
Graphics	Radeon Pro 555X 4 GB, Intel UHD Graphics 630 1536 MB
Memory	250GB

B. Types of Data Collected

The raw data in comma separated format was loaded into Microsoft Excel for review and subsequently loaded into Python with the *.csv* write syntax. The input variable or features are listed below:

*Age***:** This was originally in days but was subsequently converted to years by dividing with 365.25, which is the standard days in a year.

*Gende*r: 1 - women, 2 – men

Height: This feature was measured in centimeter (cm)

Weight: This feature was measured in kilogram (kg)

Ap_hi: Systolic blood pressure in mm Hg

Ap_lo: Diastolic blood pressure in mmHg

Cholesterol: 1- normal, 2 - above normal, 3 - well above normal

Glucose: 1 - normal, 2 - above normal, 3 - well above normal

Smoke: whether the patient smokes or not [0 - False, 1 – True]

Alcohol: whether patient drinks or not [0 - False, 1 – True]

Active: whether a patient is physically active or not [0 - False, 1 – True]

BMI: this calculated feature is the Body Mass index using $\frac{Mass}{Weight}$ kg/m^2 [6]

Cardio: this is the target variable indicating the presence of CVD or not [0 - False, 1 – True]

C. Data Cleaning and Presentation

A total of 70,000 datasets with 13 variables were loaded for analysis before classification. Out of which, one variable called "id number" was automatically dropped because it was redundant. Upon further analysis, more points were dropped using the following criteria in order to remove outliers and possible inherent errors and missing data in the dataset:

- Minimum height: "147 cm" (factoring in the possibility of having little people or children in the dataset, and the minimum age suggested only adults were considered).
- Minimum systolic blood pressure was set as "90 mm Hg" and maximum at "250 mm Hg" because according to the American Heart Society, less than 120 mm Hg is considered a normal blood pressure and higher than 180 mm Hg is considered hypertensive. The two ranges will eliminate any outlier within this dataset [7]
- Minimum diastolic blood pressure was set at "60 mm Hg" and maximum at "200 mm Hg" because according to the American Heart Society, less than 80 mm Hg is considered a normal blood pressure and higher than 120 mm Hg is considered hypertensive. The two ranges will eliminate any outlier within this dataset [7]
- Minimum weight was set to "40 kg". This was considered because of the average weight for a low body mass index. NHLBI [8] published a clinical guideline on the identification, evaluation and treatment of overweight and obesity in adults. The generated report by the computer program, supported the reason for choosing 40 kg.
- Upon application of the filters above to the dataset, about 2143 data points were dropped, and only 67857 data points were left of the training and validation of the classifier. Furthermore, it can be seen that most of the data removed were outliers that can affect performance of the classifier. Some of the outliers may be as a result of data entry error.

D. Standard Scaling

This process was an essential preprocessing method carried out to improve model performance. Typically, raw datasets without any form of processing may tend to make the machine learning model ineffective [9], [10]. This step facilitates the process of ensuring features are close to the standard normally distributed data. In practice, the shape of the distribution is often ignored the data is transformed to the center by removing the mean value of each feature, then scaling it by dividing non-constant features by their standard deviation [9], [10].

It is also essential to state that this process helped to reduce the skewness of the data to enhance the performance of the models. The standard scaling process was executed using a pipeline to reduce data leakage. In Python, the functions can be called by importing the libraries:

- *StandardScaler*
- *sklearn.pipeline.Pipeline*

E. Feature Reduction

The next task after cleaning the data was to reduce the 12 variable inputs to only 7 significant ones. This process was done using the following feature reduction technique,

- Backward Reduction Method
- Feature Importance Method
- Recursive Feature Elimination
- Univariate Feature Selection
- Ridge Regression or Regularization
- Unbiased-Ensemble Method

F. Unbiased – Ensemble Method: Feature Selection Algorithms

The ensemble method was determined by selecting the input variables that was common to all of the identified feature reduction technique without considering the weights. The steps involved in determining unbiased ensemble features are outlined below:

Step 1: Identify the constituent feature reduction methods for making the ensemble feature reduction method. In this case Univariate, FI, RFE, Ridge and Backward Feature Elimination methods.

Step 2: Identify number of features required, in this case

seven input features are required for the diagnosis.

Step 3: Identify and select the features with the highest mode for the first seven features. In this case ap_hi, cholesterol, age (years) has a mode of 5, followed by weight and ap_lo with a mode of 4 and then glucose and BMI with a mode of 3. These accounts for the 7 input features required. However, this selection does not impose more weights on ap_hi, cholesterol and age because they occur more. But instead, equal weights were applied on all selected 7 input features regardless of the value of their mode.

Furthermore, this method is more of a statistical method as it deals with the mode or the input with the highest votes among all the considered feature reduction methods. But it is expected that some features will have the same modes. In this case, the input features with the same modes are considered. The steps described are illustrated in Table II to show how the feature selection of an ensemble method is done.

In cases when selecting multiple input features from the pool of the original methods with the same modes will lead to selecting more than the required input features. Further elimination will have to be done to determine the actual features. A method that involves imposing a bias on the multiple input features can be considered. Thereby, allowing for the flexibility for the selection of the required number of ensemble input features. Nevertheless, while this bias imposition method seemed feasible, it was not implemented in this work because it was not needed.

The same input may have been selected by the unbiased-ensemble and the univariate method; this does not mean that they are the same. This is because the univariate method requires different weight on the features. However, the unbiased ensemble in this study introduced equal weight on all selected features.

The algorithms and their details used in this Paper are provided in the Appendices in [11]. Each Appendix shows the library and parameters that were used to implement the algorithms for the classification and features reduction.

TABLE II: FEATURE SELECTION MAP USED TO DETERMINE THE VARIABLES OF THE PROPOSED FEATURE REDUCTION METHOD

	Univariate	Feature Importance	RFE	Ridge	Backward Elimination	Unbiased Ensemble
gender				✓		
height		✓				
weight	✓	✓	✓	✓		✓
ap_hi	✓	✓	✓	✓	✓	✓
ap_lo	✓	✓	✓	✓		✓
Cholesterol	✓	✓	✓	✓	✓	✓
gluc	✓		✓		✓	✓
Smoke				✓	✓	
alco						
active			✓		✓	
years	✓	✓	✓	✓	✓	✓
BMI	✓	✓			✓	✓

G. Parameters for Subroutines

The parameters for each classifier are inputted into the subroutines for each individual algorithm. In this study, some of these parameters are defined based on the following requirement-

- Task type – classification
- Computing power - memory / processor speed of machine.

Some of the predefined parameters were ignored, because such parameters had no effect on the subroutines of each classifier. The parameters that were selected for use in this study were implemented uniformly for each feature reduction. This removes any bias when the classification task is executed by the software.

Scikit-learn because of its easy interface allows the entries to be defined within the subroutine. In this study, the user-defined entries were manual entered into the subroutines that had to be defined after selecting the classifier library. This entry values can be a challenging process to determine especially when tuning is required. Therefore, most of the entries were determined after considering the computational power, classification requirements and existing studies.

H. Ensemble Classifier

After the data was prepared for classification and selected features considered. A classifier was used to evaluate the performance of the various features in making a correct diagnosis. An ensemble classifier using the hard-voting technique has been selected to do this task. The ensemble classier gives an output class based on the most votes between individual classifiers made up of; KNN, MLP or Neural Networks, SVM, Logistic Regression, and Random Forest Classifiers.

However, this hard-voting ensemble method was done without imposing any weight or bias on any of the individual classifiers. As against a soft-voting method that requires that a weight average probability be imposed on individual classifiers to make a class prediction. Although a soft-voting classifier can be considered for a classification task. Furthermore, in this study a hard-voting ensemble classifier was favored because, in order to select an excellent diagnosing feature, no form of bias must be introduced into the whole process, a feat that may not be possible if a soft-voting ensemble classifier was considered.

In the same vein, an odd number of individual classifiers was considered because of the situation where there might be a tie. When even numbers of individual classifiers are considered and there is a tie, the ensemble classifier selects based on the sort order. And because this is not desirable in this diagnosis model, it is essential to use odd number of classifiers.

The results of the models are presented in the result section. Furthermore, comparisons were also made of the individual classifiers and the ensemble classifier for the overall features selected.

This is the method used in this study. The best-case scenario in this study is the target model with seven input features. This is described because the process is the same for all the other number of input features. Furthermore, there is a need for uniform parameter settings in order to justify the diagnosis made by the machine learning model.

I. Evaluation Metrics

The following metrices were used to evaluate the performance of the various models-

- Accuracy
- F1-Score
- ROC-AUC
- Confusion Matrix

III. Results

The Kaggle dataset used for this study was analyzed for outliers and removed. Out of the original 70,000 datapoints, only 67,857 of those points were suitable for classification. The data was split into a ratio of 80:20 for training and validation respectively. The results obtained were evaluated based on the metrics discussed in the last section and are presented in this section.

The selected input features were evaluated with the classifiers and the performance of the classifiers are shown in a graph and a classification report. The classification report was generated using the Python command line and presented in the tables shown. This report shows the performance of each classifier based on its classification of the test data.

In the next figures. 1 to 7, the differences on the Mean and the Standard Deviation for all classifiers may appear to be small; however, these differences are significant, [11], [14].

A. Baseline Model

The baseline model took into consideration all the twelve features in the dataset: gender, height, weight, ap_hi, ap_lo, cholesterol, smoke, glucose, alcohol, and active, years (age) and BMI as inputs for the classifier. The Python codes are found in [11]. Fig. 1 shows that the ensemble classifier and the RF classifier are good classifiers with a mean accuracy of 0.7341 and 0.734 respectively.

TABLE III: Metrics Report of the Baseline Model

Accuracy:	0.7316			
Confusion Matrix:	[[5346	1531]		
	[2111	4584]]		
Classification report				
	precision	recall	f1- score	support
0.0	0.72	0.78	0.75	6877
1.0	0.75	0.68	0.72	6695
accuracy			0.73	13572
macro avg	0.73	0.73	0.73	13572
weighted avg	0.73	0.73	0.73	13572
ROC AUC Score:	0.7310			
Precision:	0.7496			
F1 Score:	0.7157			
Recall:	0.6847			

The confusion matrix in Table III and the classification report shows that, of the 75% positive samples, the model classified 68.5% correctly. This impacted a good ROC_AUC Score of 0.73 and F1_score of 0.72, which is good for a classifier because the closer to 1 the better.

Furthermore, comparing ROC of this model with the ROC of the Framingham score of 0.724 shows that the ensemble classifier is a good classifier. This then presents a basis upon which other classifications can be compared when the features are reduced.

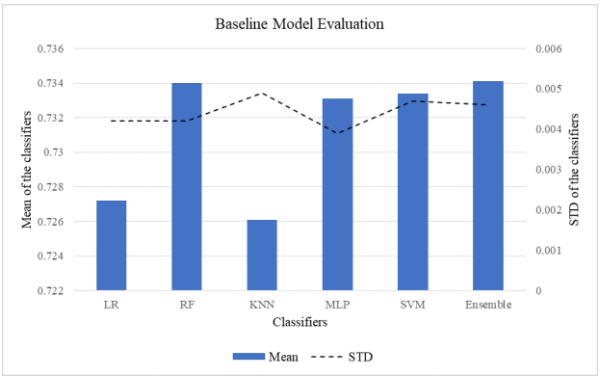


Fig. 1. Graph of mean accuracy and standard deviation of the baseline model.

B. Feature Importance Model

This method considered height, weight, ap_hi, ap_lo, cholesterol, years (age) and BMI as inputs for the classifier. The Python codes are found in [11]. Fig. 2 shows that the Ensemble classifier performs lesser than the SVM classifier, this time with a mean accuracy of 0.7313. The SVM has the best performance with a mean of 0.7322 and with the lowest deviation of 0.0037, when in actual sense the ensemble classifier should perform better. However, certain studies have shown that this is possible [12].

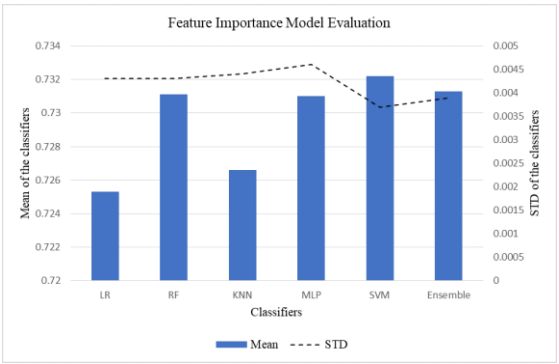


Fig. 2. Graph of mean accuracy and standard deviation of the feature importance model.

TABLE IV: Metrics Report of the Feature Importance Model

Accuracy:	0.7329			
Confusion Matrix:	[[5368	1509]		
	[2116	4579]]		
Classification report				
	precision	recall	f1- score	support
0.0	0.72	0.78	0.75	6877
1.0	0.75	0.68	0.72	6695
accuracy			0.73	13572
macro avg	0.73	0.73	0.73	13572
weighted avg	0.73	0.73	0.73	13572
ROC AUC Score:	0.7323			
Precision:	0.7521			
F1 Score:	0.7164			
Recall:	0.6839			

The confusion matrix in Table IV and the classification report shows some similarity with the baseline model. Out of the 13572 test samples 75% of the samples were positive. However, the model classified 68.4% as the ratio of correct positive results compared to the number of all samples that

should be identified as positive. This result shows that the reduced features improved the performance of the classifier, so it is possible to use height, weight, ap_hi, ap_lo, cholesterol, years (age) and BMI to make a CVD diagnosis.

C. Unbiased Ensemble Model

This method considered weight, ap_hi, ap_lo, cholesterol, glucose level, years (age) and BMI as inputs for the classifier. The Python codes are found in [11]. Fig. 3 shows that the SVM and the RF method are better classifiers ahead of the ensemble classifier as was the case in Fig. 2 where the ensemble classifier did not perform better than the individual classifiers. The ensemble classifier had the second lowest deviation (0.0041) and mean accuracy of 0.7318 when compared to the SVM and RF with a deviation of 0.0038 and 0.0040 and a mean of 0.7323 and 0.7327, respectively.

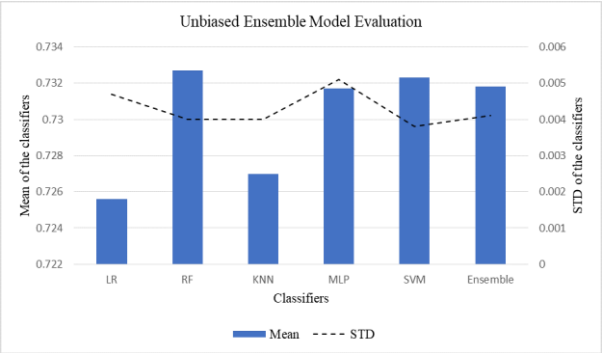


Fig. 3. Graph of mean accuracy and standard deviation of the unbiased ensemble model.

TABLE V: METRICS REPORT OF THE UNBIASED ENSEMBLE MODEL

Accuracy:	0.7335			
Confusion Matrix:	[[5397	1480]		
	[2137	4558]]		
Classification report				
	precision	recall	f1- score	support
0.0	0.72	0.78	0.75	6877
1.0	0.75	0.68	0.72	6695
accuracy			0.73	13572
macro avg	0.74	0.73	0.73	13572
weighted avg	0.74	0.73	0.73	13572
ROC AUC Score:	0.7328			
Precision:	0.7549			
F1 Score:	0.7159			
Recall:	0.6808			

However, Table V shows a ROC_AUC score of 0.7328, which is an increase from the baseline model. This result shows that the unbiased-ensemble is a good feature reduction method and a good diagnosis for cardiovascular disease can be made with, weight, ap_hi, ap_lo, cholesterol, glucose level, years (age) and BMI.

D. Univariate Feature Model

This method considered weight, ap_hi, ap_lo, cholesterol, glucose level, years (age) and BMI as inputs for the classifier. The Python codes are found in [11]. Fig. 4 shows that the RF and ensemble classifiers (deviation of 0.004) have equal mean accuracy of 0.7327 but the RF classifier has a lower deviation of 0.0037.

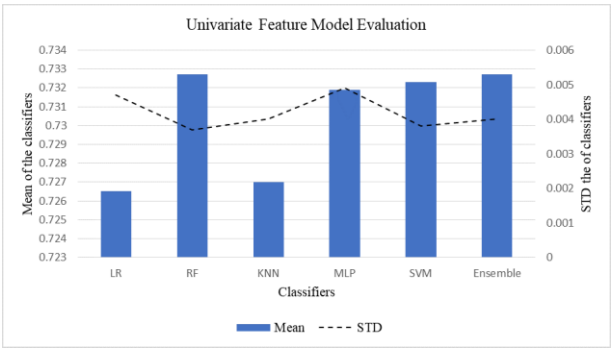


Fig. 4. Graph of mean accuracy and standard deviation of the univariate feature model.

But the confusion matrix in Table VI and the classification report shows classification similarity with the baseline model. The ROC_AUC Score of 0.7324 and precision of 0.7545 showed more improvement from the baseline model and the feature importance method.

TABLE VI: METRICS REPORT OF THE UNIVARIATE FEATURE MODEL

Accuracy:	0.7331			
Confusion Matrix:	[[5395	1482]		
	[2141	4554]]		
Classification report				
	precision	recall	f1- score	support
0.0	0.72	0.78	0.75	6877
1.0	0.75	0.68	0.72	6695
accuracy			0.73	13572
macro avg	0.74	0.73	0.73	13572
weighted avg	0.73	0.73	0.73	13572
ROC AUC Score:	0.7324			
Precision:	0.7544			
F1 Score:	0.7154			
Recall:	0.6802			

Furthermore, this method used the same input features as the unbiased-ensemble method, different weights were assigned on the univariate features but equal weight is applied on the unbiased-ensemble method. This result and the reduced features still show a good performance of an ensemble classifier when weight, ap_hi, ap_lo, cholesterol, glucose level, years (age) and BMI are considered as inputs to make a good diagnosis for cardiovascular disease.

E. Ridge Feature Model

This method considered gender, weight, ap_hi, ap_lo, cholesterol level, smoke, years (age) as inputs for the classifier. The Python codes are found in [11]. Fig. 5 shows that the ensemble classifier is the best classifier with a mean accuracy of 0.7328 and a deviation of 0.0051. The RF had the highest deviation of 0.0056 and a mean accuracy of 0.7326. This figure also shows the low performance of the KNN (0.7284 accuracy) and LR (0.7255 accuracy) models where they have the lowest deviation of 0.004 and 0.0039, respectively.

The confusion matrix in Table VII and the classification report shows that the ROC_AUC Score of 0.7302 and precision of 0.7480 showed less improvement from the baseline model, Unbiased-ensemble, univariate and the

feature importance method. 68.5% of the test data was rightly classified in contrast to the baseline model. This result shows the performance of an ensemble classifier that uses gender, weight, ap_hi, ap_lo, cholesterol level, smoke, years (age) as input variables for CVD diagnosis.

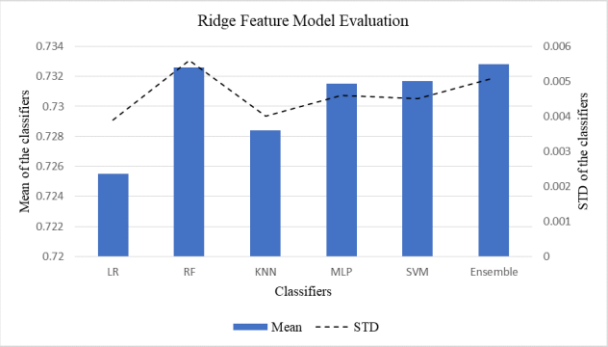


Fig. 5. Graph of mean accuracy and standard deviation of the ridge feature model.

TABLE VII: METRICS REPORT OF THE RIDGE FEATURE MODEL

Accuracy:	0.7308			
Confusion Matrix:	[[5332	1545]		
	[2108	4587]]		
Classification report				
	precision	recall	f1- score	support
0.0	0.72	0.78	0.74	6877
1.0	0.75	0.69	0.72	6695
accuracy			0.73	13572
macro avg	0.73	0.73	0.73	13572
weighted avg	0.73	0.73	0.73	13572
ROC AUC Score:	0.7302			
Precision:	0.7480			
F1 Score:	0.7152			
Recall:	0.6851			

F. Recursive Feature Elimination

This method considered weight, ap_hi, ap_lo, cholesterol level, glucose level, years (age) and activity level, [11]. Fig. 6 shows that the ensemble classifier is the best classifier with a mean accuracy of 0.7341.

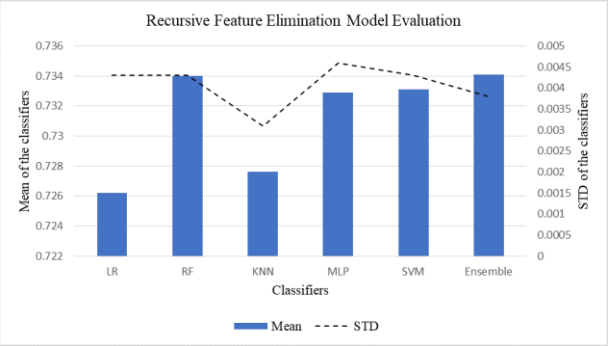


Fig. 6. Graph of mean accuracy and standard deviation of the recursive feature elimination model.

The confusion matrix in Table VIII and the classification report shows that the ROC_AUC Score of 0.7316, accuracy of 0.7322 and f1_score of 0.7169 showed more improvement from the baseline model. Furthermore, 68.7% of the test data was rightly classified in contrast to the baseline model.

TABLE VIII: METRICS REPORT OF THE RECURSIVE FEATURE MODEL

Accuracy:	0.7322			
Confusion Matrix:	[[5336	1541]		
	[2093	4602]]		
Classification report				
	precision	recall	f1- score	support
0.0	0.72	0.78	0.75	6877
1.0	0.75	0.69	0.72	6695
accuracy			0.73	13572
macro avg	0.73	0.73	0.73	13572
weighted avg	0.73	0.73	0.73	13572
ROC AUC Score:	0.7316			
Precision:	0.7491			
F1 Score:	0.7169			
Recall:	0.6874			

This result in Fig. 6 and Table VII shows the significant contribution of, weight, ap_hi, ap_lo, cholesterol level, glucose level, years (age) and activity level to CVD diagnosis.

G. Backward Feature Reduction Model

This method considered ap_hi, cholesterol level, glucose, smoke, active, year (age) and BMI. The Python codes are found in [11]. Fig. 7 shows that the ensemble classifier is the best classifier with a mean accuracy of 0.7328. The RF classifier closely had a mean accuracy of 0.7325.

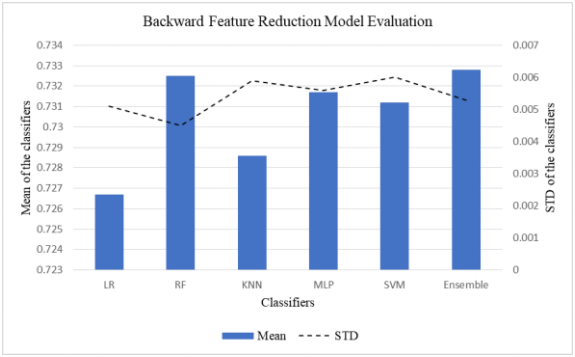


Fig. 7. Graph of mean accuracy and standard deviation of the backward feature reduction model.

TABLE IX: METRICS REPORT OF THE BACKWARD FEATURE MODEL

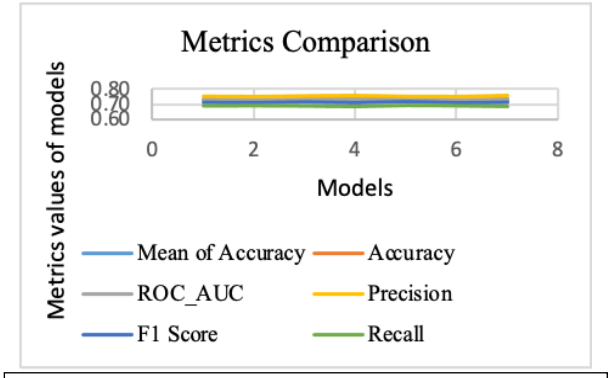
Accuracy:	0.7240			
Confusion Matrix:	[[5195	1682]		
	[2064	4631]]		
Classification report				
	precision	recall	f1- score	support
0.0	0.72	0.76	0.74	6877
1.0	0.73	0.69	0.71	6695
accuracy			0.72	13572
macro avg	0.72	0.72	0.72	13572
weighted avg	0.72	0.72	0.72	13572
ROC AUC Score:	0.7236			
Precision:	0.7336			
F1 Score:	0.7120			
Recall:	0.6917			

The confusion matrix in Table IX and the classification report shows that the ROC_AUC Score of 0.7236, accuracy of 0.7240 and f1_score of 0.7120 showed less Improvement from the baseline model. However, 69.2% of the test data was rightly classified in comparison to the baseline model. This result shows that, cholesterol level, glucose, smoke, active, year (age) and BMI can be considered as input features to make a cardiovascular disease diagnosis.

H. Feature Selection Evaluation Based on Ensemble Classifier Model

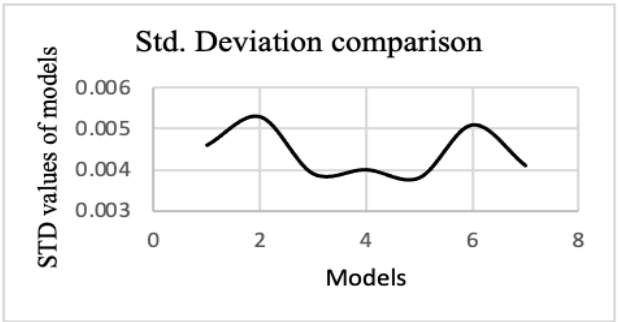
The evaluation of the metrics of different feature selection model using the Ensemble Classifier model is considered using Fig. 8. The classifier with the highest mode is a good choice in selecting which feature improves the classification process.

Fig. 8 shows that the baseline model and the RFE performed better than any other method with 0.7341, in terms of mean of accuracy. The ridge (0.7328) and BFE methods (0.7328) also showed good performance and the FI method performed least with 0.7313.



Keys:1-Baseline 2-Backward Feature Elimination 3-Feature Importance 4-Univariate Features 5-Recurent Feature Elimination 6-Ridge Features 7-Unbiasaed Features

Fig. 8. Comparison of evaluation metrics of each method.



Keys:
1-Baseline 2-Backward Feature Elimination 3-Feature Importance 4-Univariate Features 5-Recurent Feature Elimination 6-Ridge Features 7-Unbiasaed Features

Fig. 9. Standard deviation of each method.

Furthermore, the RFE had the best F1 score and recall values of 0.7169 and 0.6874, respectively when compared to the FI method which had scores of 0.7164 for the F1 score and the BFE method that had a recall value of 0.6854.

However, the unbiased-ensemble method had the best prediction accuracy, precision and ROC_AUC, with a score of 0.7335, 0.7328 and 0.7549 respectively. The univariate method performed second best when the prediction accuracy, precision and ROC_AUC are compared to the unbiased-ensemble method- with scores of 0.7331, 0.7545 and 0.7324, respectively.

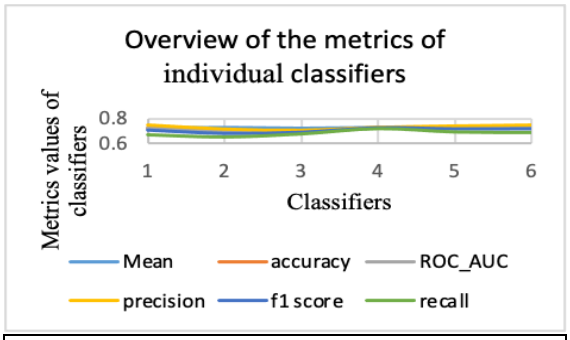
From Fig. 9 where lower values are better methods, the RFE had the lowest deviation of 0.0038 unlike the unbiased-ensemble method that had the fourth lowest deviation of 0.0041. Apart from the standard deviation metrics, the RFE and the unbiased-feature methods had equal modal values of 3 from the earlier metrics evaluated in Fig. 8. This makes the RFE features good inputs in an ensemble classifier for making CVD diagnosis.

I. Individual Classifier Comparison

The results for the evaluation of the individual classifiers are compared with the Ensemble Classifier. Furthermore, the classifiers input for the classifiers was based on the input variables of the recurrent feature elimination method.

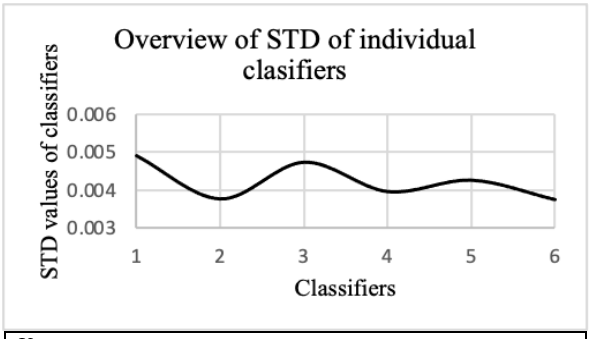
J. Metrics Comparison of Individual Classifiers in the RFE

Fig. 10 shows the performance of all the individual classifiers that were used in the RFE features. It can be clearly seen that the ensemble classifier had the best mean accuracy, prediction accuracy and ROC_AUC with values of; 0.73408, 0.73224 and 0.73164, respectively. The LR features on the other hand had a better precision with a score of 0.74966 compared to 0.74914 by the ensemble classifier.



Keys:
1-LR 2-RF 3-KNN 4-MLP 5-SVM 6-Ensemble

Fig. 10. Overview of the metrics of individual classifiers.



Keys:
1-LR 2-RF 3-KNN 4-MLP 5-SVM 6-Ensemble

Fig. 11. Standard Deviation of methods.

However, the MLP performed better for two metrics, f1 score and recall rate with 0.72465 and 0.72075 respectively. Fig. 11 also showed that the ensemble classifier had the lowest deviation of 0.00376 when compared to the individual classifiers. This confirms that the ensemble classifier was a

better classifier for the RFE input features.

K. Discussions

Fig. 1-Fig. 7 showed the metrics of the ensemble classifier using the baseline inputs and the reduced feature inputs. Furthermore, the conclusions from Fig. 1-Fig. 7 are summarized by the comparisons shown with Fig. 8 and Fig. 9. Upon review of the summary, it was evident that the RFE model was the best performing model when compared to the other. Closely behind was the novel feature reduction method implemented in this study, the unbiased-ensemble or the unbiased method.

The results shown in Fig. 10 and Fig. 11 indicate that the RFE method is a better feature reduction method on routine clinical data for the diagnosis of CVD, because it gave better mean of accuracy, F1 score, rate of recall and the least deviation.

The implication for this study was that correct right diagnosis could be made about CVD from seven routine clinical data. This result has been achieved in some studies by as much as thirteen [13] and in some cases more than twenty features, ranging from a combination of routine to genetic features and time-consuming questionnaires [14]

Researchers have used single classifiers [2], [15] for CVD classification. However, this study was able to achieve good diagnosis because of the use of an ensemble classifier. Furthermore, the input features for the ensemble classier were systematically reduced using recurrent feature elimination method.

However, this result could not measure the level or seriousness of a positive CVD diagnosis.

Further tests will have to be carried out to determine the extent of how CVD will impact a patient. But the methodology confirmed by the results of this study can be the first call by medical personnel before a patient embarks on a tedious and expensive test that may come back negative. This study shows that this expensive test can be avoided by using routine clinical data.

As expected, it was seen in Fig. 10 and Fig. 11 that the Ensemble Classifier had a better mean, lower standard deviation, better accuracy and ROC_AUC. However, LR outperformed the Ensemble Classifier in terms of precision; and the MLP outperformed the Ensemble Classifier in terms of the F1 score; the Ensemble Classifier had second best in both cases. Furthermore, the MLP had a better recall when compared to the Ensemble Classifier.

Nonetheless, if the seven metrics are considered, the Ensemble Classifier outperformed the individual models in four instances, came second best in two instances and third best in only one instance. This shows that the Ensemble Classifier is still a better option than each individual classifier.

IV. Conclusions

Selecting the right features for the diagnosis of CVD can be tedious and critical when there are a large amount of routine clinical data, genetic data, and datasets from high end machines that needs to be reviewed. This Paper study aimed to identify the effective inputs features required to make a correct diagnosis on whether someone has a cardiovascular disease or not. This diagnosis is done based on the feature reduction techniques used on the data and the Machine Learning classifier implemented.

Each of the methods considered for feature reduction were evaluated, and diagnosis were made on the selected input features using a Machine Learning Ensemble Classifier. The ensemble classifier consisted of; Random Forest, Logistic Regression, Support Vector Machines, k-Nearest Neighbor, and Multilayer Perceptron classifiers, to determine the CVD status of an individual. Here, also it was proposed for each of the feature reduction techniques, an Ensemble Classifier for uniformity purpose. The following conclusions are drawn from the study:

- The use of a Recursive Feature Elimination (RFE) method reveals that the following routine clinical data can be decisive in the diagnosis of cardiovascular diseases: body weight of an individual, systolic and diastolic blood pressure, cholesterol level, glucose level, level of physical activity, and the age of an individual. All the seven features listed above are what physicians can advise individuals to stay informed on as it contributes to their risk of developing a cardiovascular disease. With this knowledge, individuals can take careful steps to: monitor their weights, avoid diets or factors that can raise their blood pressures, avoid any substance that can lead to an increase in cholesterol level in their bodies, limit food items that may possibly increase the glucose level in their bodies and engage in some level of physical activities to keep their hearts healthy.
- The proposed unbiased-ensemble method confirms the same features as the RFE except the addition of BMI and exclusion of activity level of an individual. This novel method proved useful because of its ability to impose a unity weight across all features. This ability of the unbiased-ensemble method made it suitable for determining important features that are used in making a cardiovascular disease diagnosis.
- Machine Learning can help in acting as a confirmatory tool for physicians when evaluating an individual for cardiovascular disease. Hence, it reduces complexities where expertise and technologies are not readily available. For example, in remote locations where medical facilities or experts are limited, a Machine Learning Classifier model can be deployed as an application that can work with less complexities-even on a mobile phone, to help make life saving decisions.

Although the classifier cannot predict if an individual will develop a cardiovascular disease in the future; the results clearly shows that the system is effective in classifying if an individual has a cardiovascular disease or if the individual is free from cardiovascular disease.

Still, some studies show the impact of genetic factors or family history in CVD diagnosis (Atkov *et al.*, 2012). However, for this Paper study the databank considered had no record of family history of the individuals used for the study. Hence, the diagnosis was made from the routine clinical data in the databank at the time. This factor can prove very useful as this can go a long way to reduce the false negatives identified by the model in this study.

Conflict of Interest

The authors declare no conflict of interest.

Authors Contributions

Oyetunde P. Oyelude do the works on conceptualization, formal analysis, methodology, investigation, writing-original draft.

Dr. Rene V. Mayorga do the works on formal analysis, methodology, supervision, paper review & editing, funding acquisition.

Acknowledgement

This paper research was supported by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

[1] Z. Cheng and Z. Lu, "A novel efficient feature dimensionality reduction method and its application in engineering," *Complexity*, 2018.

[2] S. Shylaja and R. Muralidharan, "Classification of data for cardiovascular disease prediction system using multi layer perceptron," *Journal of Adv Research in Dynamical & Control Systems*, vol. 11, pp. 850-856, 2019.

[3] K. Subhadra and B. Vikas, "Neural network based intelligent system for predicting heart disease," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* , vol. 8, no. 5, pp. 484-487, 2019.

[4] J. J. Nirschl, A. Janowczyk, E. G. Peyster, R. Frank, K. B. Margulies, M. D. Feldman, and A. Madabhushi, , "A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue," *PLos One*, vol. 13, no. 4, pp. 1-16, 2018.

[5] S. Ulianova. (2019). Cardiovascular Disease dataset (Version 1). [Online]. Available: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

[6] V. Mokin. (2020). 20 models for Cardiovascular Disease prediction. [Online]. Available: https://www.kaggle.com/vbmokin/20-models-for-cardiovascular-disease-prediction

[7] [AHA]American Heart Association (2019). *How To Manage Blood Pressure.* [Online]. Available: https://www.heart.org/en/healthy-living/healthy-lifestyle/my-life-check--lifes-simple-7/ls7-blood-pressure-infographic

[8] [NHLBI]National Heart, Lung, and Blood Institute. (1998). Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults: The Evidence Report. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK1997/

[9] Pedregosa *et al*., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[10] Scikit-learn. (2020). *Sklearn: User Guide 0.23.2.* [Online]. Available: https://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf.

[11] O. P. Oyelude, "A machine learning classifiers approach for cardiovascular. disease diagnose," M.A.Sc. Thesis, Faculty of Engineering and Applied Science, University of Regina, November 2020.

[12] W. Wang, "Some fundamental issues in ensemble methods," in *Proc. 2008 International Joint Conference on Neural Networks*, 2008, pp. 2243-2250

[13] A. Rufai, U. S. Idriss, and M. Umar, "Using artificial neural networks to diagnose heart disease," *International Journal of Computer Applications*, vol. 182, no. 19, pp. 1-6, 2018.

[14] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. Rudd, and M. V. Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants," *PLOS ONE*, vol. 14, no. 5, 2019.

[15] O. Y. Atkov, S. G. Gorokhova, A. G. Sboev, E. V. Generozov *et al*., "Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters," *Journal of Cardiology*, vol. 59, no. 2, pp. 190-194, 2012.

Oyetunde P. Oyelude received a BTech in mechanical engineering from Ladoke Akintola University of Technology, Nigeria, in 2012; and a master of applied science (MASc) in industrial systems engineering, from the University of Regina, Canada in 2020.

He currently works as a program advisor at the University of Regina. Previously, he held roles as a project engineer and project coordinator, in Nigeria. Mr Oyelude is a member of the Project Management Institute (PMI), and association of Professional Engineers and Geoscientists of Saskatchewan, (APEGS).

Rene V. Mayorga is a professor in the Department of Industrial Systems Engineering, at University of Regina, Canada. His research activities are dedicated to the development of Artificial-Computational Sapience (Wisdom) as new disciplines/fields, and to Intelligent,-Sapient (Wise) Systems applied on diverse areas. Over the years he has been in the editorial board of several international journals. He was the Editor in Chief for *Applied Bionics and Biomechanics* from 2003 to 2016. He is the co-editor of *"Toward Artificial Sapience: Principles and Methods for Wise Systems"*, Springer 2008. He has published papers widely in scientific journals, international conferences proceedings, books, and monographs. Also, he has edited several international conference proceedings. Over the years he has also served in several occasions as General Chair and Program Chair for several International Conferences.