American Journal of Artificial Intelligence
2023; 7(1): 6-16
http://www.sciencepublishinggroup.com/j/ajai
doi: 10.11648/j.ajai.20230701.12
ISSN: 2639-9717 (Print); ISSN: 2639-9733 (Online)

Evaluation of Machine Learning Techniques Towards Early Detection of Cardiovascular Diseases

Anietie Ekong

Department of Computer Science, Akwa Ibom State University, Ikot Akpaden, Nigeria

Email address:
anietieekong@aksu.edu.ng

To cite this article:
Anietie Ekong. Evaluation of Machine Learning Techniques Towards Early Detection of Cardiovascular Diseases. *American Journal of Artificial Intelligence*. Vol. 7, No. 1, 2023, pp. 6-16. doi: 10.11648/j.ajai.20230701.12

Received: January 20, 2023; **Accepted**: February 13, 2023; **Published**: April 15, 2023

Abstract: The effectiveness of three Machine Learning (ML) algorithms: Support Vector Machine (SVM), Random Forest (RF) and K-Nearest Neighbour (KNN) techniques for the early diagnosis of heart diseases were evaluated. Heart disease' dataset collected from kaggle.com data repository, which comprised of 303 data points with 13 features and a target variable were used and data preprocessing by data shuffling and dimension reduction were performed. The new dimension of the dataset was chosen such that 85.03% of the original information is retained. The preprocessed dataset was partitioned into 70% of the training set and 30% of the testing set. The ML algorithms were trained and tested for the diagnosis of cardiovascular diseases (CVD). The training performances of these models were evaluated with a k-fold cross-validation algorithm using 10 folds. The k-fold accuracy shows KNN with an accuracy of 0.837662, RF with an accuracy of 0.834091, and SVM with an accuracy of 0.814935. The test results also show KNN with an accuracy of 0.8, SVM with an accuracy of 0.7889, and RF with an accuracy of 0.7667. KNN emerged the best model both in training and test's performances and is recommended for the early diagnosis of CVD.

Keywords: Cardiovascular Disease, Prediction, K-nearest Neighbor, Machine Learning, Random Forest, Support Vector Machine, Machine Learning, Classification, Diagnoses

1. Introduction

Cardiovascular disease (CVD) is the cause of most sudden deaths globally, and over 17 million lives are lost annually. The CVD is a group of medical condition that affects the heart and blood vessels. Most of these deaths are caused by heart attacks and strokes [1]. A great percentage of the CVD can be prevented by imbibing the culture of a healthy lifestyle [2]. However, early and accurate diagnosis is the key to ensuring improved outcome of CVD.

Application of Machine learning (ML) in the diagnosis of CVD and other ailments in the field of medicine is gaining popularity globally. The models make it possible for tasks to be classified with or without human intervention. ML is made up of a conceptual framework that is predicated on models that receive input data and predict possible outcomes [3]. ML techniques have successfully been deployed for the prediction of CVD [4], artery disease [5], breast cancer [6], and H1N1 flu [7], and dengue fever [8], influenza [9] and swine fever [10]. Many authors have proposed several of these techniques for the prediction, detection, and diagnosis of CVD.

Krittanawong et al. [3] proposed the ability of Support Vector Machine (SVM) and Boosting algorithms, in particular, to accurately predict CVD. The result from these algorithms may assist doctors in having a proper understanding of the data and hybrid technique to predict the CVD, applying different ML models such as "Logistic Regression (LR), Adaptive Boosting (AdaBoostM1), Multi-Objective Evolutionary Fuzzy Classifier (MOEFC), Fuzzy Unordered Rule Induction (FURIA), Genetic Fuzzy System-Logit Boost (GFS-LB) and Fuzzy Hybrid Genetic Based Machine Learning (FH-GBML)" was proposed.

Amin Ulhaq et al. [12] further explored ML methods to ensure more automated, precise, and early diagnosis of most CVD. F. S. Alotaibi [13] used some well-known ML algorithms such as feature selection algorithm and cross-validation method were used. Evaluation metrics such as sensitivity, Matthews' correlation coefficient, specificity, classification accuracy, and execution time were used to

identify and classify people with heart disease. C. Friedman, A. Wong, D. Blumenthal [11] posited that Secured and trusted use of these data can speed the progression of knowledge from the laboratory to the patient's bedside and provide a cornerstone for health care reform. Hassani, Tao, Kamyab and Mohammadi [14] proposed improved accuracy of Heart Failure (HF) prediction by using a heart disease dataset and deployed hybrid ML approaches to train the data and predict the HF chances. Vanisree and Singaraju [15] used a novel methodology of the NNDT (Neural Network and Decision Tree) that uses Neural Network and Decision Tree for training and testing model for better classification of heart diseases was proposed. The performance of the proposed approach was compared with Naïve Bayes, Support Vector Machine, Neural Network, Voted Perceptron and Decision Tree algorithms. The diagnosis and treatment of CVD is complex, especially in developing countries and this is occasioned by the dearth of equipment and acute shortage of clinicians which make the prediction, diagnosis and treatment of patients with this health condition a herculean task [16]. Approach for designing and operating of health care logistics evaluation framework may be steady, deterministic or dynamic and computational intelligence is required to handle the dynamics [29]. In their work, Dawes, Marva, W. Shi et al [17] used ML techniques to predict total cardio-metabolic risk assessment of selected healthy adults. The predictive accuracy was compared with a known risk prediction tool called the Hellenic SCORE, a calibration of the ESC SCORE against the combined CVD incidences to evaluate which approach for risk classification will guarantee the correct CVD prediction rate.

According to Umoren, Etuk, Ekong and Udonyah [28], the mortality rate from diseases due to delayed or wrong diagnoses is high especially in low income countries with low quality of healthcare and lack of well-trained medical personnel.

To resolve these complexities in the diagnosis and treatment of CVD, a non-invasive approach using ML is proposed. This paper attempts to evaluate the performance of different ML techniques: Support Vector Machine, Random Forest, and K-Nearest Neighbour on the heart disease dataset gotten from kaggle.com, ranks the models based on their accuracy and recommends the best model. This system is implemented using the R and Python programming languages.

2. Proposed Model

This work proposes a hybrid model for the early detection of cardiovascular disease as illustrated in Figure 1. The framework is comprised of a data repository, data pre-processing, data partitioning, ML models, model evaluation, and model ranking. The heart disease dataset was collected from kaggle.com. This dataset is comprised of features (parameters) and labels (cardiovascular disease) for use in modeling a system for the early prediction of heart disease. The collected dataset is pre-processed via data shuffling and dimension reduction. The dimension reduction algorithm employed is called the Principal Component Analysis (PCA).

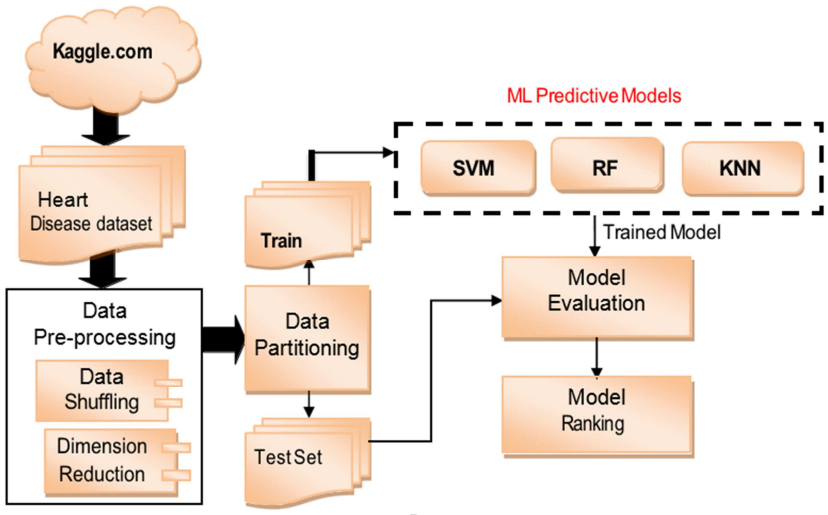

Figure 1. *The proposed model for diagnosis of Cardiovascular Diseases.*

Dimension reduction aims to eliminate redundant information in the dataset. The shuffled and reduced dataset is partitioned into 70% training of the dataset and 30% of the testing dataset. Three (3) different ML models namely; the Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbour (KNN) are trained using the training dataset and tested using the test dataset. The performance of these models are computed using Accuracy and Kappa.

The different modules in the proposed model for early diagnosis of cardiovascular disease are presented as follows;

Module 1: Heart Disease Dataset

The heart disease dataset used in this paper was downloaded from kaggle.com, a machine learning dataset repository.

Module 2: Data Pre-Processing

This module takes care of data shuffling and dimension

reduction. Data shuffling serves the purpose of reducing variance and making sure the models remain general and not affected by overfitting, since the data was sorted by their class/target. With data shuffling, the training and test sets are representative of the overall distribution of the dataset. The PCA algorithm reduces redundancy in the dataset by eliminating feature-to-feature correlation in the dataset.

Module 3: Data Partitioning

The shuffled and reduced dataset is partitioned into 70% training set and 30% test set for models training and testing respectively. Three (3) ML models namely; Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbour (KNN) are employed.

Module 4: Model Evaluation

The performance of Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbour (KNN) are evaluated using Accuracy and Kappa.

Module 5: Model Ranking

The ML models used are ranked based on their accuracy. The model with the highest accuracy is recommended for future use in the early detection of cardiovascular disease.

3. Data Collection

For this work, as earlier mentioned, the heart disease dataset was collected from kaggle.com, a machine learning data repository. The dataset comprised of 13 features, 1 label, and 303 data points. An excerpt of the original dataset is presented in Table 1.

Table 1. *Dataset Features [18].*

Features	Dataset Description
Age	age of the patient in years
Sex	sex (male/female)
Cp	chest pain type
Trestbps	resting blood pressure
Chol	Serum cholesterol
Fbs	fasting blood sugar
Restecg	Resting electrocardiographic results
Thalach	maximum heart rate achieved
Exang	exercise induced angina
Oldpeak	ST depression induced by exercise relative to rest
Slope	slope of the peak exercise ST segment
Ca	number of major vessels
Thal	Thalassemia
Cardiovascular Disease	Target

4. Data Preprocessing

The original dataset was sorted based on the target variable, for which reason data shuffling is required. Data shuffling is aimed at mixing data and ensuring that the logical relationship between columns is retained. It randomly shuffles data from a dataset within an attribute or a set of attributes. Shuffling the heart disease dataset ensures that training and test sets are representatives of the collective distribution of the data. The shuffling algorithm as implemented by Python programming language is presented below:

STEP 1: read excel heart disease dataset
heart_dataset = pandas.read_excel("dataset_url")
STEP 2: Convert data to a DataFrame object
Df = DataFrame(data=heart_dataset)
STEP 3: Shuffle dataset Shuffled_data = Df.reindex (numpy.random.permutation(df.index))
STEP 4: Save shuffled dataset
shuffled_dataset = shuffled_data.to_excel("path_to_excel_file")

The shuffled dataset is presented in Table 1.

Principal Component Analysis (PCA) algorithm is used in this work to eliminate redundancy in the dataset. The PCA breaks the correlation that exists among features of the dataset while encouraging features to target (labels) correlation. The number of dimensions (components) of the reduced dataset is chosen such that over 70% of the original information is retained. The screen plot that shows the cumulative explained variance for any given number of components is presented in Figure 2. The chosen number of dimensions is nine (9) with a cumulative explained variance of 85.03. The number of dimensions with their respective cumulative explained variance is presented in Table 3. The reduced dataset is presented in Table 4. Each column in the reduced dataset is called a Principal Component. Each Principal Component represents a feature in the dataset. A mapping between the Principal Component of the PCA algorithm and features of the original dataset is presented in Table 5 while table 6 contains the sample dataset. The correlation plots before and after dimension reduction are presented in Figure 3 and Figure 4.

$$(x_1, y_1) \ldots\ldots (x_n, y_n) \quad (1)$$

Where: x represents the features and y represents the target class called "cardiovascular"

$$x_i = x_i{}^1, x\,i_2, \ldots\ldots, \quad (2)$$

Where: $x_i{}^d$ is a real value, and $y_i = \{-1, 1\}$ with -1 denoting "No Heart Disease" and 1, "Heart Disease". RBF (Radial Basis Function) Kernel maps the non-separable training data from input space to feature space to obtain an optimised hyperplane that correctly segregates the data. The Radial Basis Kernel Function is presented in Equation 3.

$$k(\overrightarrow{x_i}, \overrightarrow{x_j}) = \emptyset(\overrightarrow{x_i})^T \emptyset(\overrightarrow{x_i}) \quad (3)$$

$$k(\overrightarrow{x_i}, \overrightarrow{x_j}) = \exp(-\gamma ||x_i - x_j||^2) \quad (4)$$

where: $\gamma: \frac{1}{2\sigma^2} > 0, x_i$ is the support vector points, x_j are the feature vector points in the transformed space, $k(\overrightarrow{x_i}, \overrightarrow{x_j})$ is the kernel function.

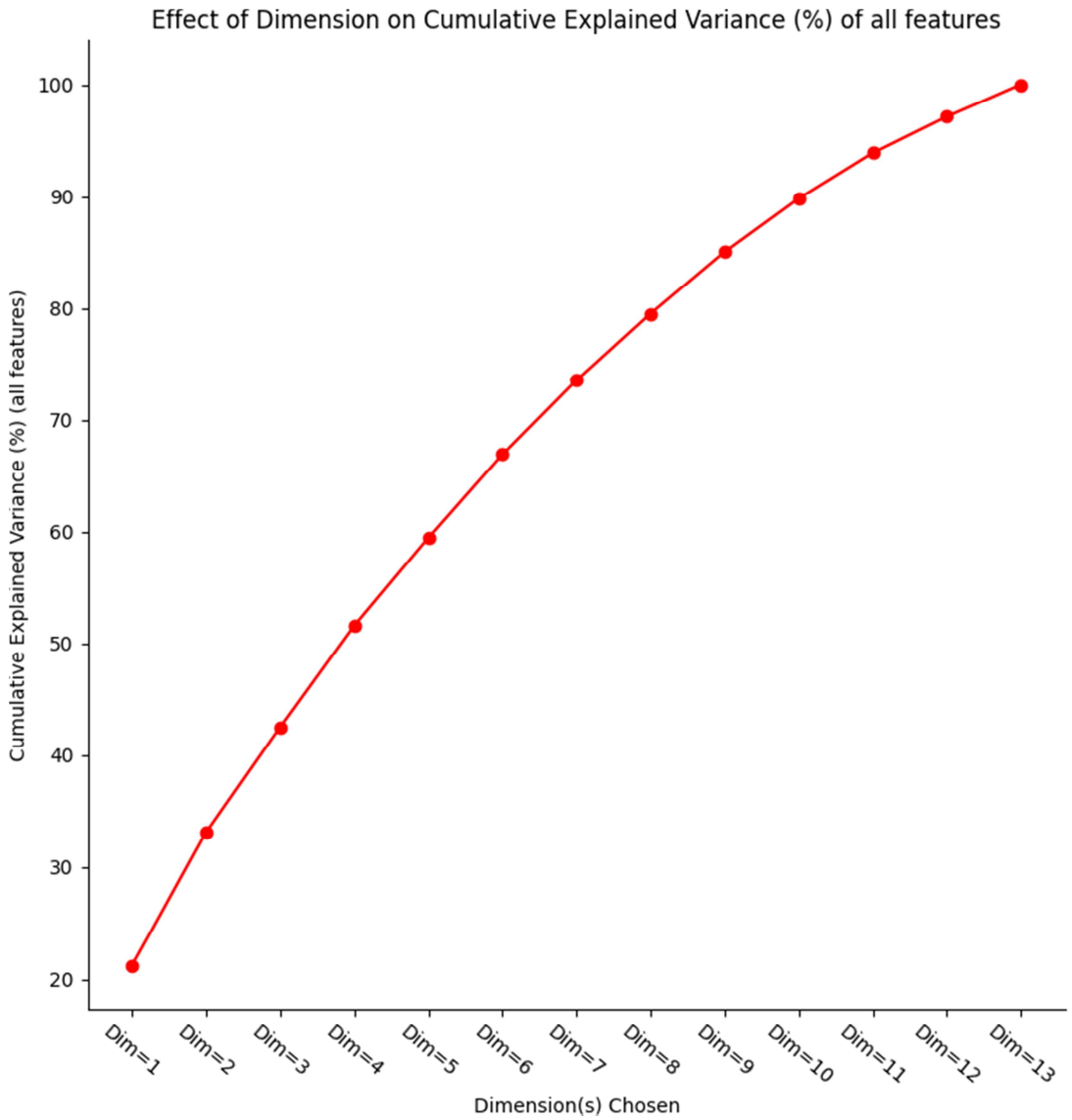

Figure 2. *Effect of Dimension on cumulative explained variance.*

Table 2. *Shuffled Dataset.*

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Cardio vascular
58	0	0	170	225	1	0	146	1	2.8	1	2	1	0
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
65	0	2	155	269	0	1	148	0	0.8	2	0	2	1
44	1	2	120	226	0	1	169	0	0	2	0	2	1
46	1	2	150	231	0	1	147	0	3.6	1	0	2	0
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
48	1	0	122	222	0	0	186	0	0	2	0	2	1
65	1	3	138	282	1	0	174	0	1.4	1	1	2	0
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
46	0	1	105	204	0	1	172	0	0	2	0	2	1
42	1	3	148	244	0	0	178	0	0.8	2	2	2	1
45	1	0	115	260	0	0	185	0	0	2	0	2	1
59	1	3	178	270	0	0	145	0	4.2	0	0	3	1
55	1	0	140	217	0	1	111	1	5.6	0	0	3	0
58	0	3	150	283	1	0	162	0	1	2	0	2	1
70	1	0	130	322	0	0	109	0	2.4	1	3	2	0
59	1	0	138	271	0	0	182	0	0	2	0	2	1
52	1	1	134	201	0	1	158	0	0.8	2	1	2	1
50	1	2	129	196	0	1	163	0	0	2	0	2	1
54	1	2	150	232	0	0	165	0	1.6	2	0	3	1

Table 3. *Effect of Dimension on Cumulative Explained Variance.*

Dimension chosen	Dimension = 1	Dimension = 2	Dimension = 3	Dimension = 4	Dimension = 5	Dimension = 6	Dimension = 7	Dimension = 8	Dimension = 9	Dimension = 10	Dimension = 11	Dimension = 12	Dimension = 13
Cumulative Explained Variance	21.25	33.07	42.48	51.57	59.43	66.89	73.53	79.5	85.03	89.81	93.89	97.14	100

Table 4. *Reduced Heart Disease dataset.*

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	Cardio-vascular
2.2418	2.279	0.6388	0.9642	-0.9492	1.968	1.7679	0.1133	2.3613	0
-0.5576	1.7478	-0.9311	-0.0658	0.6092	0.091	-0.745	0.1847	0.4327	1
-0.877	2.0146	-0.9691	0.7632	-0.335	-1.2027	0.3389	-1.1539	-0.3285	1
-2.5467	-0.8094	0.442	0.1082	-0.0327	-0.3353	-0.0273	-0.0447	-0.2703	1
0.1703	-0.2496	1.0415	2.1041	1.1212	-0.9889	-0.0394	-0.2018	0.6454	0
-0.6489	-0.3829	-0.0768	1.494	-1.1375	0.709	0.1059	-1.1348	0.6201	1
-1.9295	-0.7067	0.0499	-1.0743	0.5019	1.1513	-0.3542	-0.4339	0.6461	1
-0.0183	2.5456	2.331	0.1285	0.3955	0.5479	-0.2129	1.1555	-0.8347	0
0.4786	-0.3661	0.6015	1.1889	0.5083	1.0611	-1.1066	-0.208	-1.6523	1
-2.6683	-0.7164	-1.0983	0.4755	-0.6227	-0.1117	0.189	0.5997	0.2288	1
-1.6433	0.8168	1.6613	-0.6878	0.8027	-0.0698	-1.093	-0.255	1.3113	1
-2.0026	-0.752	-0.3069	-1.3207	0.7493	1.1939	-0.3815	0.1582	0.5021	1
2.2143	1.5765	1.8094	1.977	3.1768	-1.2621	-0.8224	-0.8244	-0.052	1
4.009	-1.9531	0.2455	2.38	1.2468	-0.9495	0.5588	0.0315	0.067	0
-1.139	3.3151	0.9219	0.203	0.4279	0.6002	0.944	0.5141	-0.6754	1
2.8353	1.011	-0.5184	-0.6417	-0.9867	0.1799	-2.3067	0.4806	0.2938	0
-1.1962	0.518	-0.2918	-1.3717	0.5478	0.7536	-0.2926	-0.9688	0.2956	1

Table 5. *PCA Component to Feature Map.*

Principal Component	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Related Features	Age	Sex	chol	Fbs	thalach	exang	Oldpeak	Slope	ca	Thal	restecg	Trestbps	Cp

Table 6. *Sample Dataset [18].*

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thakach	Exang	Oldpeak	Slope	Ca	Thal	Target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
42	1	0	140	226	0	1	178	0	0	2	0	2	1
61	1	2	150	243	1	1	137	1	1	1	0	2	1

From Figure 3 and Figure 4, it is observed that PCA algorithm has successfully eliminated the dependency that existed between features of the heart disease dataset.

4.1. Principal Component Analysis (PCA)

This paper employs Principal Component Analysis (PCA) for feature extraction (extraction of relevant features from the dataset thereby eliminating redundant ones). In this way, the dimension of the original dataset is reduced. From the original dataset, the PCA algorithm processes and outputs a reduced and non-redundant dataset. PCA algorithm is presented below;

Step 1: Calculate the covariance matrix [19]

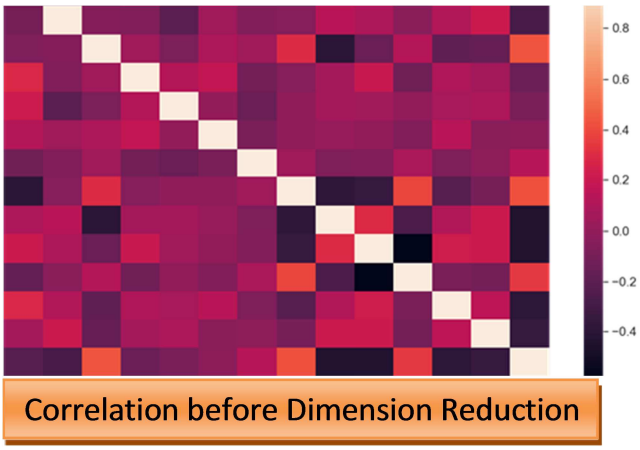

Figure 3. *Correlation before dimension reduction.*

Figure 4. *Correlation after dimension reduction.*

Step 2: Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components.

Step 3: Choose K eigenvectors which correspond to the largest K eigenvalues to be the principal components of the dataset.

Step 4: Project the data as $Y = X\,v$

$v = [v\,1 \ldots v\,K]$ is a $d \times K$ matrix where columns $v\,i$ are the eigenvectors which correspond with the largest K eigenvalues.

4.2. Support Vector Machine (SVM)

A non-linear binary Support Vector Machine (SVM) [20] is used to predict cases of cardiovascular disease. SVM classifies a patient into a negative class (-1) for no heart disease and positive class (+1) for heart disease. The SVM training dataset is formulated as in Equation 1. The kernel function computes the dot product of the mapped data points in the transformed feature space. The optimal hyperplane that segregates [21] between "Heart Disease" and "No Heart Disease" is found using Equation 5.

$$w^T.x + b = \sum_{i=1}^{l} a_i\, y_i\, \emptyset(\overrightarrow{x_i})^T \emptyset(\overrightarrow{x_i}) + b = 0 \quad (5)$$

The classification frontier are found by the following equation;

$w\,\emptyset(x) + b = 1$: for "Heart Disease" data points

$w\,\emptyset(x) + b = -1$: for "No Heart Disease" data points

The optimal weight vector (w) is presented in Equation 6 as;

$$\vec{w} = \sum_{i=1}^{l} a_i\, y_i\, \emptyset(\overrightarrow{x_i}) \quad (6)$$

The dual formulation of SVM algorithm is used in this work. This formulation is presented in Equation 7

$$\max \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i\, a_j\, y_i\, y_j \emptyset(\overrightarrow{x_i})^T \emptyset(\overrightarrow{x_i}) \quad (7)$$

Subject to

$$0 < a < C$$

and $\sum_{i=1}^{l} a_i\, y_i = 0$

Where:

$\alpha\, i$: weight vector

y : label vector

$\emptyset(\overrightarrow{x_i})^T \emptyset(\overrightarrow{x_i})$: kernel function

C: the intercept.

To make prediction, the decision function g is presented in equation (8).

$$g(\vec{x}) = sgn(\vec{w}^T\vec{x} + b) \Rightarrow sgn\left(\sum_{i=1}^{l} a_i\, y_i\, \emptyset(\overrightarrow{x_i})^T \emptyset(\overrightarrow{x_i}) + b\right) \quad (8)$$

Where: $g(\vec{x})$ is the predicted outcome, sgn is in the sign of $(\vec{w}^T\vec{x} + b)$ (i.e. 1 0r -1), and a_i is the weight of the vector.

4.3. Random Forest (RF)

The Random Forest applies the technique of bagging (or bootstrap aggregation) which is a method of generating a new dataset with replacement from an existing dataset. The following are features of Random Forest [22];

(1) Ensemble learning used in Random Forest prevents it from overfitting

(2) Bagging enables random forest to work well with a small dataset

(3) Random Forest predictors can be trained in parallel

(4) Automatic feature selection is enabled by decision tree learning in random forest.

In this work, Random forest is used to classify the heart disease dataset into a negative class "No Heart Disease" and the positive class "Heart Disease". The Random Forest algorithm used in this work is presented below [23]:

Step 1: Select k features from total m features randomly. Where: k < m

Step 2: Among the "k" features, calculate node "d" using the best split point.

Step 3: Split the node into daughter nodes using the best split.

Step 4: Repeat 1 to 3 steps until the "p" number of nodes has been reached.

Step 5: Build forest by repeating steps 1 to 4 for "n" number of times to create "n" number of trees.

4.4. K-Nearest Neighbour (KNN)

The K-Nearest Neighbour (KNN) [24, 25] is an instance-based learning algorithm based on the observation that similar features belong to the same class. The parameter K, in KNN, is the number of neighbours used to classify one particular sample point. A majority vote approach is used to assign a point to a class. The KNN algorithm used in this work is presented below:

Let K = the number of nearest neighbours and

$D = P_1, \ldots, P_n$ be the training dataset in the form

$P_i = x_i, c_i$), where x_i is the d-dimensional feature vector of the point P_i and c_i is the class of P_i. $p' = (x', c')$

Step 1: Compute the distance (x', xi) between $p'$and all pi belonging to D

Step 2: Sort all points, pi according to the key $(x',)$

Step 3: Select the first k points from the sorted list, those are the k closest training samples to p'

Step 4: Assign a class to p' based on majority vote: $c' = argmax\, \Sigma(xi, ci)$ belonging to D, $I(y = ci)$

End

4.5. Cross-Validation

The K-Fold cross-validation algorithm evaluates the performance of the ML models used in this work. In K-Fold cross-validation, all the entries in the cardiovascular disease dataset are used for both training and validation just ones [26]. The value of K was chosen as ten (10) for this work. The steps are presented below:

Step 1: Original dataset is partitioned into k equal subset called a fold. Let the folds be represented as $f\,1, f\,1, \ldots, f\,k$

Step 2: For i = 1 to k

- Keep the fold, $f\,1$ as the validation set with the remaining $k - 1$ folds as the training set and then train the ML models with the training set and calculate the accuracy of the models by validating the predicted results on the validation set.

Step 3: The accuracy of the model is evaluated by averaging the accuracies of all the k cases of cross validation.

4.6. Performance Evaluation

The Performance metrics such as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are used in computing Accuracy and Kappa for the evaluation of the SVM, RN, and KNN performance. The general form of a confusion matrix is presented in Table 7 and the elements of the confusion matrix are described below:

Table 7. *The Confusion Matrix.*

		Predicted Class	
Actual Class		+1	-1
	+1	True Positive Prediction	False Negative Prediction
	-1	False Positive Prediction	True Negative Prediction

True Positive Prediction (TP) – These are the correct positive predictions. This implies that the actual and predicted values are both positive for the given data point.

True Negative Prediction (TN) – These are the correct negative predictions. This implies that the actual and predicted values are both negative for the given data point.

False Positive Prediction (FP) – When the actual class is negative (i.e "no cardiovascular disease") while the predicted class is positive (i.e "cardiovascular disease"). This is a case of an incorrect prediction.

False Negative Prediction (FN) – When the actual class is positive (i.e "cardiovascular disease") while the predicted class is negative (i.e "no cardiovascular disease"). This is also a case of wrongly predicted cardiovascular situation. Hence TP, and TN are desirable, while FP, and FN are not. From the TP, TN, FP, and FN, the model's accuracy and Kappa are computed as follows:

Accuracy – this is the ratio of correct predictions to the total predictions. Accuracy is computed as:

$$\text{Accuracy} = \frac{\text{True Positive(TP)} + \text{True Negative(TN)}}{\text{Total Sample(N)}}$$

Kappa – measures the inter-rater (and intra-rater) reliability of the model [27].

$$\text{K} = \frac{\text{Po} - \text{Pe}}{1 - Pe}$$

Where:

Po: is the overall accuracy of the model

PE: is the hypothetical probability of chance agreement, using the observed cardiovascular dataset to calculate the probabilities of each observer randomly seeing each category.

5. Results and Discussion

The heart disease dataset collected from kaggle.com was shuffled and reduced using PCA algorithm. The reduced cardiovascular dataset was partitioned into 70% training data and 30% testing data for model training and testing respectively. The training set comprises the same set of reduced inputs or features as the test set. The reduced features are outputs of the principal component analysis, namely; age, sex, chol, fbs, thalach, exang, oldpeak, slope, ca, thal, restecg, trestbps and cp. The output of this system is presented as a target variable in the last column of the heart disease dataset. In this work, the output variable is called "Cardiovascular". Cardiovascular disease can either be negative ("Non Cardiovascular Disease") or positive ("Cardiovascular Disease"). The training and test sets are presented in Tables 8 and 9 respectively. Figure 5 is the class distribution showing the frequency of positive (cardiovascular) class and negative (none-cardiovascular) class in the training and test datasets.

From the training dataset class distribution in Figure 5, it is

observed that, 46% of the dataset represent none-cardiovascular cases while 54% represent cardiovascular cases. Also, from the test dataset, 46% of the dataset represents none-cardiovascular cases, while 54% represent cardiovascular cases. This shows that the target class is evenly distributed throughout the dataset. Hence, these datasets are suitable for use in training the ML algorithms for the early detection of cardiovascular diseases.

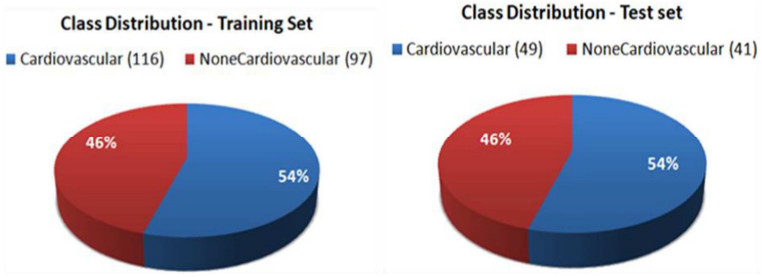

Figure 5. *Class Distribution (Training set and Test set).*

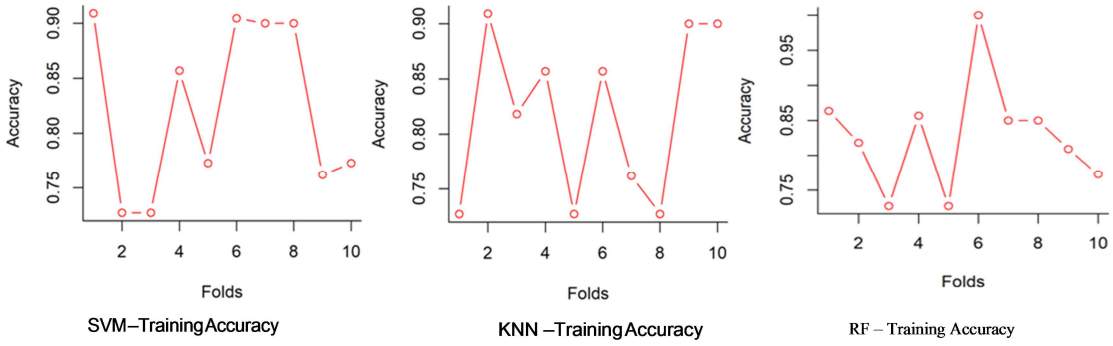

Figure 6. *K-Fold Cross Validation Result.*

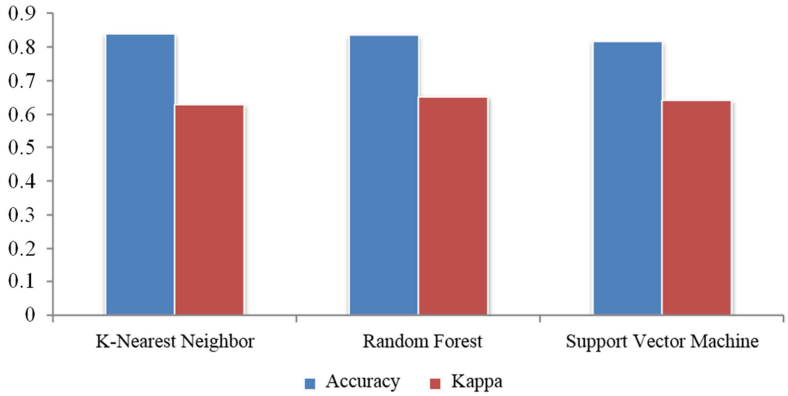

Figure 7. *Models' Training Performance.*

Table 8. *Training Dataset.*

Age	Sex	Chol	Fbs	Thalach	exang	oldpeak	Slope	Ca	CLASS
-0.55757	1.747755	-0.93106	-0.06584	0.609208	0.091027	-0.745	0.184688	0.432694	Cardiovascular
-0.87698	2.014563	-0.96911	0.763245	-0.33501	-1.20268	0.338935	-1.15395	-0.32854	Cardiovascular
-2.54674	-0.80938	0.441973	0.108242	-0.03275	-0.33535	-0.02726	-0.04471	-0.27035	Cardiovascular
0.170318	-0.24963	1.041516	2.104106	1.121186	-0.98893	-0.03937	-0.20182	0.645436	None Cardiovascular
-0.01834	2.545637	2.330987	0.128511	0.395522	0.547934	-0.21286	1.155529	-0.83475	None Cardiovascular
-2.66827	-0.71642	-1.09833	0.475516	-0.6227	-0.11168	0.188994	0.59966	0.228806	Cardiovascular
-1.6433	0.81683	1.661252	-0.68782	0.802718	-0.06979	-1.09301	-0.25503	1.311265	Cardiovascular
2.214332	1.576538	1.809372	1.977024	3.176813	-1.26214	-0.8224	-0.82441	-0.05202	Cardiovascular
4.008957	-1.95315	0.245471	2.379965	1.246804	-0.94945	0.558783	0.031476	0.066956	None Cardiovascular
-1.13905	3.31505	0.921917	0.203043	0.427918	0.600226	0.944035	0.514063	-0.67545	Cardiovascular
2.835341	1.01099	-0.51837	-0.64173	-0.98671	0.179852	-2.30666	0.480645	0.293833	None Cardiovascular
-1.19624	0.517978	-0.29182	-1.37172	0.547771	0.753633	-0.29262	-0.96881	0.295602	Cardiovascular
-0.61287	0.438221	1.139229	-0.5766	1.570201	-0.57239	-0.53747	-1.20466	-0.24045	Cardiovascular
-0.70035	-0.53871	-0.09187	-0.35141	-0.51233	-0.03642	1.124085	-1.51293	-0.34722	Cardiovascular
0.050364	-0.29446	-1.40132	0.298128	0.790502	1.62981	1.085971	-0.40668	0.886336	Cardiovascular

Table 9. *Test Set.*

Age	Sex	Chol	Fbs	Thalach	exang	oldpeak	slope	Ca	CLASS
2.24176	2.279011	0.638813	0.964192	-0.94925	1.968042	1.767873	0.113295	2.36131	None Cardiovascular
-0.64887	-0.38288	-0.07679	1.493981	-1.13755	0.708983	0.105895	-1.13483	0.620053	Cardiovascular
-1.92949	-0.70672	0.049902	-1.07433	0.501943	1.151281	-0.35424	-0.43389	0.646092	Cardiovascular
0.47858	-0.3661	0.601524	1.188904	0.50825	1.061121	-1.10658	-0.20797	-1.65233	Cardiovascular
-2.00257	-0.75201	-0.30688	-1.32066	0.749263	1.193857	-0.38146	0.158158	0.502063	Cardiovascular
-1.15501	-0.54659	0.552373	-0.01295	-0.87535	-0.59128	-0.23079	-0.66068	0.549008	Cardiovascular
-2.20305	-0.54715	0.67942	0.353594	-0.40968	-0.42003	-0.05949	-0.77794	-0.34938	Cardiovascular
1.66871	-0.08609	-0.17401	-0.31218	0.176561	-0.00643	-1.72506	0.147647	-1.50676	None Cardiovascular
-1.69288	-1.05478	0.341105	-0.13854	0.419479	-0.07173	1.13416	-0.44533	0.027523	Cardiovascular
0.736185	3.199744	-3.12699	-1.46135	3.028836	-1.05711	-0.96927	2.770744	-2.07961	Cardiovascular
-1.81927	-1.30075	0.266253	-1.16362	-0.19981	1.250712	-0.87806	0.107906	0.976966	None Cardiovascular
0.895162	-0.52921	0.964552	-0.96043	-1.67725	-1.75308	-1.54622	0.699004	0.762138	Cardiovascular
-0.15066	2.38153	-0.73399	0.408424	0.198297	0.14933	-0.9287	-0.43002	-0.07822	Cardiovascular
-0.64747	1.826586	1.114917	-2.15351	1.830379	-1.02071	0.476667	-2.13513	1.430875	None Cardiovascular
-0.2807	-2.26959	0.070851	0.343628	0.548776	-0.60547	0.224755	0.523504	-0.04697	None Cardiovascular

i. Model Training

The performance of SVM, RF, and KNN trained using the training dataset is evaluated using 10- fold cross-validation algorithm. The accuracy of the ML models across 10-folds of the 10-fold cross- validation algorithm are presented in Figure 6. The accuracy and kappa from performance values for SVM, KNN and RF are presented in Table 10.

To detect the best performing models, Bar Chart is used. The Accuracy and Kappa of the ML models on the training dataset using K-Fold cross-validation are presented in Figure 7. It shows that K-Nearest neighbour is the best, followed

ii. Model Testing

Our ML algorithms; K-Nearest Neighbour, Random Forest and Support Vector Machine are tested using 30% of the heart disease's dataset.

Table 10. *Model's Training Performance.*

	Accuracy	Kappa
Support Vector Machine	0.814935	0.640725
Random Forest	0.834091	0.650544
K-Nearest Neighbour	0.837662	0.627812

Support Vector Machine

Actual \ Predicted	1	-1
1	40	10
-1	9	31

Random Forest

Actual \ Predicted	1	-1
1	39	11
-1	10	30

K-Nearest Neighbour

Actual \ Predicted	1	-1
1	44	13
-1	5	28

Figure 8. *Confusion Matrix.*

Table 11. *Model Testing Result.*

Chol	Fbs	Thalach	Exang	Oldpeak	Slope	Ca	Actual	SVM	KNN	RN
0.639	0.964	-0.949	1.968	1.768	0.1133	2.3613	-1	-1	-1	-1
-0.077	1.494	-1.138	0.709	0.106	-1.135	0.6201	1	1	1	1
0.05	-1.074	0.502	1.151	-0.354	-0.434	0.6461	1	1	1	1
0.602	1.189	0.508	1.061	-1.107	-0.208	-1.652	1	1	1	1
-0.307	-1.321	0.749	1.194	-0.381	0.1582	0.5021	1	1	1	1
0.552	-0.013	-0.875	-0.59	-0.231	-0.661	0.549	1	1	1	1
0.679	0.354	-0.41	-0.42	-0.059	-0.778	-0.349	1	1	1	1
-0.174	-0.312	0.177	-0.01	-1.725	0.1476	-1.507	-1	-1	-1	-1
0.341	-0.139	0.419	-0.07	1.134	-0.445	0.0275	1	1	1	1
-3.127	-1.461	3.029	-1.06	-0.969	2.7707	-2.08	1	-1	1	-1
0.266	-1.164	-0.2	1.251	-0.878	0.1079	0.977	-1	1	1	1
0.965	-0.96	-1.677	-1.75	-1.546	0.699	0.7621	1	-1	-1	-1
-0.734	0.408	0.198	0.149	-0.929	-0.43	-0.078	1	1	1	1
1.115	-2.154	1.83	-1.02	0.477	-2.135	1.4309	-1	-1	1	-1
0.071	0.344	0.549	-0.61	0.225	0.5235	-0.047	-1	-1	1	1
0.022	0.324	-2.423	0.243	-0.012	-1.359	0.909	-1	-1	-1	-1
0.311	1.334	-2.669	0.505	0.994	0.2842	-2.054	1	1	1	1
-0.69	0.23	0.224	0.597	-0.744	0.1638	0.8018	-1	1	1	1
-0.229	-0.755	-0.16	-1.17	-0.093	-0.421	-1.357	-1	1	1	1
-0.476	-0.568	0.204	-0.22	0.266	0.4193	-0.165	1	1	1	1

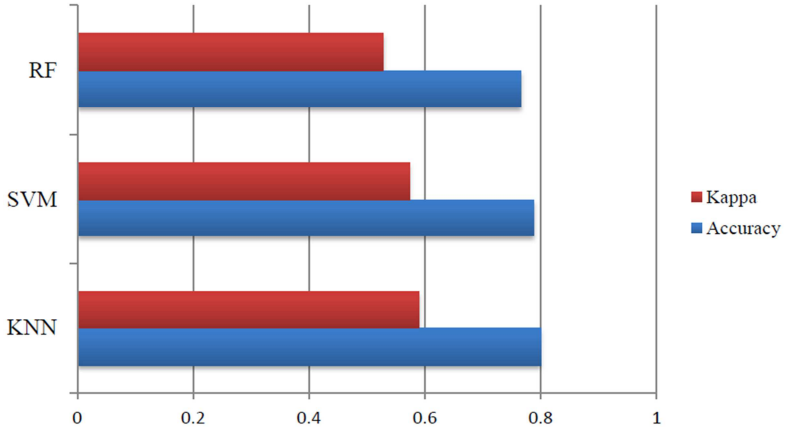

Figure 9. *The Models' Performance Plot.*

The input values from the test sets, actual, and predicted results are presented in Table 11. The model testing result in Table 10 shows the predicted values from the three ML algorithms. Cells highlighted with yellow color represent incorrect prediction, while all other cells represent correct prediction. Here, 1 represents the positive class (i.e. cardiovascular disease) and -1 represents the negative class (i.e. none cardiovascular disease).

From Table 10, it is observed that SVM has the least number of misclassifications. The confusion matrix that shows the number of false predictions (i.e. False Negative and False Positive) and true prediction (i.e. True Negative and True Positive) on the test dataset is presented for the SVM, RF, and KNN in Figure 8 which 8 shows the true positive value for SVM as 40, true negative prediction 31, false-positive prediction 9 and false-negative prediction of 10. Figure 9 shows the Models' Performance Plot.

For Random Forest, true-positive prediction is 39, true negative prediction 30, false positive-prediction is 10 and false-negative prediction of 11.

For K-Nearest Neighbour, true positive-prediction is 44, true negative-prediction is 28, and false-positive prediction is 5 and false-negative prediction of 28. The performances of these models are presented in Table 12 and visualized in Figure 9 which shows that the KNN outperforms the SVM and the RF in accurately predicting CVD. Hence, KNN is recommended for the prediction of CVD.

Table 12. *Models' performance.*

	KNN	SVM	RF
Accuracy	0.8	0.7889	0.7667
Kappa	0.5903	0.5736	0.5287

6. Conclusion

In this work, a combined technique for early diagnosis of cardiovascular diseases using machine learning approach was undertaken. The framework of this system is comprised of kaggle.com data repository, heart disease dataset, data preprocessing modules, data partitioning, and ML models, model evaluation, and model ranking. We collected 300 data points from kaggle.com, preprocessed it through data shuffling and dimension reduction using PCA. The result of data preprocessing is a reduced dataset with 9 features. The reduced dataset was partitioned into training and test sets. The training dataset was used in the training of SVM, KNN, and RF. The training performance of the ML models was evaluated using 10-fold cross-validation algorithm. The training and test performance show K-Nearest Neighbour as the best in the diagnosis of cardiovascular diseases.

Conflict of Interests

The authors declare that they have no competing interests in this research.

References

[1] Cardiovascular Disease. Accessed: February 25, 2021 [online]. Available: https://www.who.int/health topics/cardiovascular-diseases

[2] *Global Atlas on Cardiovascular Disease Prevention and Control*, WHO, World Heart Federation, and World Stroke Organization, 2011.

[3] C. Krittanawong, H. U. H. Virk, S. Bangalore, et al. "Machine learning prediction in cardiovascular diseases: a meta-analysis." Sci Rep 10, 16057 (2020). https://doi.org/10.1038/s41598-020-72685- 1.

[4] K. Anderson, P. Odell, P. Wilson, W. Kannel, "Cardiovascular Disease risk Profiles". American Heart Journal, vol. 121, no.1, pp293 – 298, 1991.

[5] P. Lapuerta, S. Azen, L. Labree, "Use of Neural Networks in Predicting the risk of Coronary Artery Disease". Computers and Biomedical Research, vol. 28, no.1, pp38 – 52, 1995.

[6] H. Asri, H. Mousannif, H. Moatassime, T. Noel, "Using Machine Learning Algorithms for Breast Cancer risk Prediction and Diagnosis". Procedia Computer Science, Vol. 83, pp1064 – 1069, 2016.

[7] F. Koike and N. Morimoto, "Supervised Forecasting of the Range Expansion of Novel Nonindigenous Organisms: Alien Pest Organisms and the 2009 H1N1 flu pandemic". Global Ecology and Biogeography, vol. 27, 991-1000, 2018. doi: 10.1111/geb.12754.

[8] S. Anno, T. Hara, H. Kai, M. A. Lee, M. A, Y. Chang, K. Oyoshi, Y. Mizukami, T. Tadono, "Spatiotemporal Dengue Fever Hotspots Associated with Climatic factors in Taiwan including Outbreak Predictions Based on Machine-Learning". Geospatial Health, vol. 14, pp183-194, 2019. doi: 10.4081/gh.2019.771.

[9] L. Tapak, O. Hamidi, M. Fathian, M. Karami, "Comparative evaluation of time series models for predicting influenza outbreaks: Application of influenza-like illness data from sentinel sites of # healthcare centers in Iran". BMC Research Notes, vol. 12, 2019. doi: 10.1186/s13104-019-4393-y.

[10] M. Alsharqi, W. J. Woodward, J. A. Mumith, D. C. Markham, R. Upton, and P. Leeson, "Artificial intelligence and echocardiography," Echo Research and Practice, vol. 5, pp. R115–R125, 2018.

[11] C. P. Friedman, A. K Wong, D. Blumenthal "Achieving a Nationwide Learning Health System." SciTransl Med. Vol. 2, no.57, pp57–29, 2010.

[12] Amin UlHaq, Jian Ping Li, Muhammad Hammad emon, Shah Nazir, Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", Mobile Information Systems, vol. 2018, Article ID 3860146, vol. 21 pages, 2018. https://doi.org/10.1155/2018/3860146.

[13] F. S. Alotaibi, "Implementation of Machine Learning Model to Predict Heart Failure Disease" International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10 no. 6, 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100637.

[14] M. A. Hassani, R. Tao, M. Kamyab, M. H. Mohammadi. "An Approach of Predicting Heart Disease Using a Hybrid Neural Network and Decision Tree." In Proceedings of the 2020 5th International Conference on Big Data and Computing (ICBDC 2020). Association for Computing Machinery, New York, NY, USA, pp84–89, 2020. DOI: https://doi.org/10.1145/3404687.3404704.

[15] K. Vanisree and J. Singaraju, "Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks," International Journal of Computer Applications, vol. 19, no. 6, pp6–12, 2011.

[16] A. C. Dimopoulos, M. Nikolaidou, F. F. Caballero, et. al, "Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk." BMC Med Res Methodol 18, 179 (2018). https://doi.org/10.1186/s12874-018-0644-1.

[17] T. J. W. Dawes, A. de Marvao, W. Shi et al., "Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: a cardiac MR imaging study," *Radiology*, vol. 283, no. 2, pp381–390, 2017.

[18] David W. Aha "Heart Disease UCI, Version 1", 2019. Retrieved February 12, 2021 from https://www.kaggle.com/ronitf/heart-disease-uci

[19] I. T. Jolliffe, "Principal Component Analysis." *Springer Series in Statistics. New York: Springer-Verlag.* 2002. doi: 10.1007/b98835. ISBN 978-0-387-95442-4.

[20] Cortes, Corinna; Vapnik, N. Vladimir, "Support-vector networks". Machine Learning. 20 (3): 273–297, 1995. CiteSeerX 10.1.1.15.9362. doi: 10.1007/BF00994018. S2CID 206787478.

[21] M. Adankon, M. Cheriet, "Support Vector machine." In: Li S. Z., Jain A. (eds) Encyclopedia of Biometrics. Springer, Boston, MA. (2009). https://doi.org/10.1007/978-0-387-73003-5_299

[22] S. Chopra, "An Introduction to Building a classification model using Random Forest in python. http://www.datascience.com/blog/classification-random-forest-in-python. Retrieved: 26 July, 2019.

[23] Mohammed Zakariah "Classification of large datasets using Random Forest Algorithm in various applications: Survey" International journal of Engineering and Innovative Technology (IJEIT) Volume 4, Issue 3, September 2014.

[24] Fix, Evelyn; Hodges, Joseph L. "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties" (Report). 1951. *USAF School of Aviation Medicine, Randolph Field, Texas.*

[25] Altman, Naomi S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". *The American Statistician*. Vol 46, no. 3, pp 175–185. 199 doi: 10.1080/00031305.1992.10475879.

[26] J. Brownlee, "K-Fold cross validation technique," http://machinelearningmastery.com/k-fold-cross validation 2019. Retrieved: 27 July, 2019.

[27] McHugh, Mary L. "Interrater reliability: The Kappa statistic". Biochemia Medica. Vol. 22, no. 3, pp 276-282, 2021. *Supplementary material.*

[28] I. J. Umoren, U. E. Etuk, A. P. Ekong, K. C. Udonyah (2021), Healthcare Logistics Optimization Framework for Efficient Supply Chain Management in Niger Delta Region of Nigeria. *International Journal of Advanced Computer Science and Applications*, vol 12, no. 4, pp 594-604.

[29] A. Ekong, H. Odikwa, O. Ekong (2021), Minimizing Symptom-based Diagnostic Errors Using Weighted Input Variables and Fuzzy Logic Rules in Clinical Decision Support Systems. *International Journal of Advanced Trends in Computer Science and Engineering, vol. 10, no. 3, pp 1567-1575* https://doi.org/10.30534/ijatcse/2021/121032021