Available online at http://icsejournal.com/index.php/JCSE
Journal of Computer Science an Engineering (JCSE)
e-ISSN 2721-0251
Vol. 4, No. 1, February 2023, pp. 1-14

Heart Disease Prediction Using Principal Component Analysis and Decision Tree Algorithm

Moshood Abiola Hambali[1,], Morufat Damola Gbolagade[2], Yinusa Ademola Olasupo[3*]

[1, 3,] Computer Science Department, Federal University Wukari, Taraba State, Nigeria
[2.] Department of Computer Science, Al-Hikmah University, Kwara State, Nigeria.
Email: [1]hambali@fuwukari.edu.ng [2]dammyconsult@gmail.com, [3]yinusa@fuwukari.edu.ng
* corresponding author

ARTICLE INFO

Article History:
Received December 3, 2022
Revised December 20, 2022
Accepted February 28, 2023

Keywords:
Heart disease,
cardiovascular disease,
decision tree,
PCA

Correspondence:
E-mail: yinusa@fuwukari.edu.ng

ABSTRACT

Globally, cardiovascular disease is among the major diseases that lead to death. Early forecasts are crucial. Using the patient's medical record, the supervised learning algorithm for predicting heart disease at an early stage was proposed. The principal component analysis (PCA) classifier and decision tree algorithm were created to classify medical record data. To predict cardiovascular diseases, data mining was utilized. The proposed strategy improves the diagnostic efficiency of physicians. Using data received from the UCI repository, the classifier's efficacy was confirmed. PCA offers 98% precision, 100% sensitivity, and 98% accuracy. In terms of accuracy, sensitivity, and precision, the results showed that the PCA outperformed the decision tree.

1. Introduction

Medical diagnosis is essential yet difficult task that must be performed effectively; therefore, its automation would be quite beneficial [1]. Unfortunately, physicians are in low supply, and not all are adept in every discipline. However, efficient computer-based information system or decision support systems aid in lower the cost of clinical trials and medical care. In addition, the precise and effective deployment of computerized system necessitates comparison of the various methodologies. In reality, the majority of hospitals employ clinical information systems to manage patient and their information. Sadly, these techniques frequently create vast quantities of data that are seldom used to inform clinical decision-making. These statistics provide an abundance of esoteric information that has been mostly disregarded.

Cardiovascular disorders are among the most prevalent illnesses in the world today [2]. Based on the World Health Organization (WHO) fact sheet of 2019, heart disease killed over 17.9 million people, representing about 32% of global death [3]. Numerous organizations in the medical domain have utilized data mining aggressively and extensively. Consequently, decision-making is enhanced by identifying patterns and styles in vast quantities of complex data.

Data mining is an act of discovering patterns within enormous quantities of data [4][5]. Data mining utilizes classification and clustering for supervised and unsupervised learning, respectively. Using a variety of clinical reports and other patient concerns, the medical sciences collect enormous amount of data. Medical datasets can be extensively mined for the same reason, and the uncovered hidden patterns can be developed for clinical diagnosis. Medical database is vast, complex, and globally dispersed. These datasets must be arranged and incorporated into the hospital management systems.

Recently, a number of researchers have employed data mining approaches to develop diagnostic procedures for various forms of cardiac disease. The major objective of this study is to collect information that assist healthcare experts in creating intelligent clinical decisions. This study

http://dx.doi.org/10.36596/jcse.v4i1.617

employs classification approach to efficiently predict and diagnose heart disease by reducing the number of attributes.

Cardiovascular disease is becoming a prominent cause of death on a global scale. As a result, heart disease prediction has sparked widespread interest in the healthcare community [6]. As a result, a number of studies have developed machine-learning (ML) algorithms for the on-time diagnosis of heart problems to aid in the development of medical therapies. A substantial amount of research has been conducted in order to develop accurate methodologies and instruments for the diagnosis and prediction of cardiac disease.

Ref[7] created an intelligent heart disease prediction system (IHDPS) that is compatible with the .NET framework. It is an easy-to-use web-based cardiac disease prediction program. It was created using data mining approaches such as Nave Bayes, neural networks, and decision tree algorithms. The dataset includes medical characteristics associated with heart disease cases. The outcomes of their work showed that Nave Bayes outperforms the other two algorithms in terms of accurate prediction for the entire population tested. The accuracy of Naive Bayes predictions is 86.12%, followed by neural networks (85.68%) and decision trees (80.40%).

Ref[8] proposed a system based on data mining approach for detecting heart disease. For accurate evaluation, the Tanagra tool was employed to classify the dataset using 10-fold cross validation. The classification algorithms and accuracy on the datasets used, respectively, are k-NN (45.67%), Naive Bayes (52.33%), and Decision list (50.00%). The results showed that the Bayes Algorithm performed well, with a best compact time of 609ms for processing the dataset and making accurate predictions.

Ref[9] employed equal frequency voting method with a gain ratio decision tree. This was evaluated using the J48 decision tree and bagging algorithm. It has been demonstrated that using voting discretization with a Gain ratio decision type improves performance, sensitivity, and precision. The proposed model's sensitivity and accuracy were increased from 72.01% to 77.90% and 78.90% to 84.10%, respectively, whereas those of the bagging algorithm are 74.93% and 81.40%. They concluded that the proposed methodology resulted in more accurate for health condition diagnoses.

Ref[10] developed a model that examined various cardiac disease detection techniques. Numbers of supervised machine learning algorithms were used such as Nave Bayes, KNN, and Decision tree with the aid of Tanagra tools for data classification. They used 10-fold cross validation to categorize the dataset and discovered that the neural network algorithm had the best accuracy of 100%. When combined Decision tree algorithm with a genetic algorithm using six attributes, the result was ranked second with an efficiency of 99.20% accuracy.

Ref[11] developed a neural network-based heart disease diagnosis system. The implementation made use of a dataset with thirteen features, including blood pressure, gender, cholesterol, obesity, and smoking. The result showed that an artificial neural network can predict cardiac disease with near-perfect accuracy.

Ref[12] proposed a hybrid method that uses natural network and generic algorithm to diagnosed cardiac disease based on a set of risk factor data. In comparison to the back propagation algorithm, the hybridization of this method was implemented in Matlab, and the resulting accuracy was 89%.

Ref[13] offered standard data mining classifiers such as decision table (DT), Classification and Regression Tree (CART) and Iterative Dichotomized 3 (ID3). They derived an unbiased estimate from the massive dataset by applying 10-fold cross-validation techniques. The testing results show, the CART method is the most accurate classifier, with an accuracy of 83.40 percent and a model-building time of 0.23 seconds. CART also has the lowest average error rate of 0.30 when compared to other models.

http://dx.doi.org/10.36596/jcse.v4i1.617

Ref[14] excerpted hidden patterns from the data provided obtained from International Cardiovascular Hospital using data mining techniques. For the diagnose of heart diesease, the machine learning tools utilized were Waikato Environment Knowledge Analysis (WEKA), and the method is a combination of Information Gain and Adaptive Neuro-Fuzzy Inference System. In compared to other methodologies, the result indicated 98.24% accuracy rate.

Ref[15] developed a technique for data mining that included REPTREE, Naive Bayes, Bayes Net, J48 and CART. As a criterion for the diagnosis of cardaic attack disease, they obtained 11-attribute patient information from medical practitioners in South Africa. Age, gender, dummy values, heart rate, chest pain, cholesterol, blood sugar, blood pressure, cardiogram, alcohol intake and smoking are the attributes. The outcome revealed the following regarding the predicted accuracy of the classifiers: J48 (99%), REPTREE (99.07%), CART (99.07%), Bayes Net (98.15%) and Naive Bayes (97.22%). Also, [16] conducted experiments with a dataset contains 294 records and 13 attributes using the Weka 3.6 software. The experiment demonstrated the effectiveness of numerous data mining approaches. Nave Bayes produced an accuracy of 85.03 percent, followed by Decision Tree with an accuracy of 84.01%. Thus, Nave Bayes outperformed decision tree and the other method.

For additional prediction and diagnosis of cardiac illness, [17] developed various classifiers to evaluate their performance based on F-measure, precision, specificity, accuracy, sensitivity and ROC curve. K- Nearest Neighbors (KNN), Decision Tree, Sequential Minimal Optimization (SMO), J48 and Naive Bayes (NB) are the classifiers employed in their work. Using a 10-folds cross validation dataset, these classifiers are applied to categorize existing data, and the obtained results were compared. The classification accuracy is as follows: J48 (83.732%), KNN (82.775%), SMO (82.775%) and NB (81.818%). The results indicate that J48 achieved better than the other classifiers utilized.

Ref[18] provided a model employing two methodologies on collected data. In the second technique, weka was utilized to evaluate selected attributes in the form of attribute relation file format (ARFF). Three algorithms; Neural network, decision tree and Nave Bayes were employed for implementation, with Nave Bayes having the highest prediction accuracy at 82.914%. The performance of Naives surpasses that of the other. In addition, [19] presented a model for assessing the effectiveness of several classifiers in predicting cardiovascular disease.Their models include the application of C5.0 decision tree, KNN, Support Vector Machine (SVM) and Neural Network. The dataset used was acquired from the University of California, Irvine (UCI) and has two labeled classess: those with and without cardiac disease. With a prediction accuracy of 93.02%, the outcome indicates that the decision tree has better performance. The accuracy of the remaining classifiers, neural network, SVM and KNN were 80.20%, 86.05% and 88.37% respectively.

Ref[20] developed a hybrid technique for diagnosing coronary heart disease by employing C4.5, multi-layer perceptron (MLP), multinomial logistic regression (MLR) and fuzzy unordered rule induction algorithm (FURIA). In addition to the innovative hybrid strategy, they created a model of K-mean clustering algorithms, particle swam optimization (PSO) and correlation-based feature subset (CFS) for distinguishing the risk factor from the Non-Invasive Clinical Dataset. Experiments were conducted on dataset obatained from the department of cardiology at the Indira Gandhi Medical College in Shimla, India that consist of 335 samples with 26 features. MLR which has the best accuracy of 88.40%, outperforms other algorithms on cleavaland heart disease benchmark data.

Ref[21] suggested a model to predict the heart risk rate utilizing a variety of disease classifiers, including Nave Bayes, decision tree algorithm, KNN, Memoral network, and ID3 algorithm. The unique identifier of the patient is the scenario's key attribute, whereas smoking and a history of heart disease are the prediction attributes. The suggested system examines the input qualities before classifying them using the KNN algorithm. The ID3 was subsequently used to assess heart disease risk. The dataset's properties include input, key and prediction attributes. The input pulse rate characteristics include cholesterol levels, blood pressure, gender and age. They discovered that with the increase in the number of attributes, the accuracy of KNN and ID3 also increased. For instance, the prediction rate accuracy was 40.30% prior to the addition of previous heart disease and smoking, but improved to 80.60% after the addition of these factors.

Ref[22] presented a method that combines multiple kernel learning (MKL) with an Adaptive Neuro-Fuzzy inference system (ANFIS). The method of deep learning (MKL and ANFIS) employs a two-fold strategy.Using MKL, the dataset is initially separated into two groups: normal and heart disease patients. The second strategy involves submitting the MKL result to the ANFIS in order to classify heart diseased and healthy patients. They used KEGG metabolic reaction network dataset. The proposed approaches generated high specificity of 98%, 99% sensitivity, and reduced mean square error (MSE) of 0.01.

Ref[23] created a model that divides data into ten subsets. They conducted the experiment ten times using a 10-cross validation dataset that was applied to the neural network model, Gini index prediction model, Equal frequency model, J48 algorithm, and Bagging algorithm. The framework consists of two components: the neural network component and the classification component, the models aid in prediction of cardiac disease in a MATLAB environment with 93% precision. Their developed neural network prediction model (yielded an accuracy of 87.89%) and a Gini index decision tree model (77.90% accuracy) for the prediction of cardiac disease that surpasses their prior model.

Ref[24] introduced a model that partitioned 303 dataset records into two distinct sets: the training and test sets. 60% of the dataset was considered for testing set, whereas the remaining data were used for training set. Multiplayer perceptron neural network (MLPNN) with backpropagation (BP) algorithm was employed for the training and prediction. The investigation was conducted using Weka 3.6.11 tools for estimating heart disease risk. The outcome demonstrated that the suggested method predicted the risk of heart disease with an accuracy of almost 100 percent.

Ref[25] have suggested a unique SVM-based GA optimization. The SVM and GA discovered seven features. Seven features were selected and supplied to the SVM classifier to determine its accuracy. With the entire feature, SVM achieved an accuracy of 83.70%, while the extracted feature provided an accuracy of 88.34%. They revealled that the GA-SVM model outperformed existing feature selection algorithms such as Gain ratio, chi-squared, GA, etc.

Ref[26] proposed decision tree and artificial neural network model for the early identification of heart disease. Weka tool was used to hybridize the classifiers for improved performance in predicting heart disease. The UCI repository patient record for the impacted patient was subjected to a ten-fold validation test. A method called prunning is used to minimize noise in a data collection of patients with heart disease. Compared to the result of ANN and C4.5, the accuracy of the hybrid technique is highest at 78.14 percent, while individual classifiers ANN (77.40 percent) and C4.5 (76.66 percent) have a lower accuracy.

Ref[2] studied a 2015 dataset of 299 people with heart failure. To rank the features linked with the most significant risk variables and forecast the patients' survival, they applied a number of machine learning (ML) classifiers. Furthermore, they used conventional biostatistical tests to do a feature ranking analysis and compared the outcomes with those produced by ML techniques. The feature ranking algorithm identified that both ejection fraction and serum creatinine as the two most important characteristics, hence these two variables alone are used to construct the ML survival prediction models.

Using a hybrid genetic algorithm (GA) and particle swarm optimization (PSO) optimized approach based on random forest (RF), Ref[6] proposed GAPSO-RF to identify the best attributes that can help improve the heart disease prediction accuracy. The suggested GAPSO-first RF's phase involves choosing the most important attributes to include in the starting population using multivariate statistical analysis. A discriminating mutation method was then used in GA. GA was

http://dx.doi.org/10.36596/jcse.v4i1.617

modified to use global search technique while PSO also modified to use local search technique, are combined into one algorithm by GAPSO-RF. PSO also grasped the idea of rehabilitating people who had been passed over throughout the selection process. Area under the receiver, sensitivity, specificity and accuracy were some of the evaluation criteria that were used to verify the effectiveness of the suggested GAPSO-RF technique.

Ref[1] compared wrapper and filter selection approaches for classification. The CFS, IG, and CS filter methods are considered. Wrapper approaches include BFS, LFS, and GSS. This investigation uses a WEKA-implemented Decision Tree classifier. CFS outperformed other filter algorithms in accuracy and execution time. The CFS technique was 93.8% accurate for heart disease, 89.5% for diabetes, and 96.8% for breast cancer. Using the same procedure, 1.08s, 1.02s, and 1.01s latency delays were recorded. BFS outperformed other wrapper approaches. Heart disease, diabetes, and breast cancer datasets have 94.7%, 95.8%, and 96.8% accuracy. Using the same procedure, 1.42s, 1.44s, and 132s latency delays were recorded. It has been proposed to use a novel hybrid attribute evaluator technique that combines improved K-Means clustering with the CFS filter approach and the BFS wrapper approach. It was decided to test the hybrid approach using a decision tree classifier. Clustering and classification were merged into a better decision tree classifier. The average results obtained showed that precision, f-score, sensitivity, and specificity were 96.7%, 96.2%, 95.6%, and 96.5%, respectively.

To identify cardiac illness, Ref[27] employed the following four ML models: KNN, adaBoost (AB), DT and RF. A generalized approach was created to evaluate the strength of the important features that influence the prediction of cardiac disease. Kaggle datasets from Cleveland, Hungary, Switzerland, and Long Beach (CHSLB) were used to test the models. On the CHSLB dataset, the KNN, AB, DT, and RF models' respective accuracy rates were 100%, 100%, 96.10% and 99.03%. In the case of a single (Cleveland) dataset, just two models— KNN and RF —show good accuracy, with 97.83% and 93.437%, respectively. Finally, utilizing Streamlit, an online cloud hosting platform, the researchers created a computer-aided smart system for disease prediction.

2. Method

This study obtained cardiovascular heart disease-related medical datasets from the UCI repository. The dataset was initially subjected to feature selection to exclude unnecessary features, and then a classification technique was applied using a decision tree and principal component analysis (PCA). The decision tree algorithm has proven to be an excellent classifier. The outcomes of this study revealed a new path for diagnosing and preventing heart disease in the future.

2.1 Methodology and Approach Solution

This paper presents an intelligent classification method based on a dynamically reduced subset of features. Coronary Heart Disease (CHD) cases were identified using decision tree classification and Principal Component Analysis for prediction and compare their performances. PCA was initially used as extraction of relevant features before classifiers were employed to predict the heart disease.

2.2 Principal Components Analysis

PCA is an instrument utilized in prediction models and exploratory data sets. It minimizes the complexity of a dataset while retaining as much variation as feasible and can be used to vividly illustrate the genetic distance and relatedness between populations. After the original data has been normalized, PCA is performed by first calculating the covariance or correlation matrix and then doing the eigenvalue decomposition on the correlation matrix [28].

Principal components are the basis vectors created on a dimensional space in distinct data points that are uncorrelated as a result of the repeated development of best–fitting lines to yield an orthogonal basis. The ideal line is the one that minimizes the average squared distribution's separation from the line and a particular point. PCA is used to categorize and investigate data similarity and difference. The relevance of PCA lies in its ability to generate Principal Components (PCs), which are tiny variables that operate according to the variance estimation theory. A new

http://dx.doi.org/10.36596/jcse.v4i1.617

orthogonal coordinate system is created using PCA that more accurately captures the variance in a given dataset. It has a relationship to canonical correlation analysis (CCA), where coordinate systems are used to characterize how well two datasets vary from one another. When applied to a dataset, PCA recalls the most basic model of a real eigenvector-based multivariate analysis. PCA and factor analysis are very closely related. A factor analysis solves the eigenvectors of a somewhat different matrix and makes more domain-specific assumptions about the underlying structure. A PCA is an orthogonal linear transformation that shifts the data to a new coordinate system so that the first coordinate contains the most variance when the data are projected as scalars. Algorithm 1 has displayed the PCA's pseudo-code, which was employed in this work.

Algorithm 1: Pseudo-code of PCA

```
Initialize First solution Prev_Config
Best _Fitness = Fitness (Prev_Config)
For n = 0 to # of iterations
    Perturbation ( )
    If Fitness (New_Config) > Fitness (Prev_Config)
        If Fitness (New_Config) > Best Fitness
            Best_Fitness = Fitness (New_Config)
        End-If
        Prev_Config = New_Config
        Exploration ( )
    Else
        Scaterring ( )
    End-If
End-For

Exploration ( )
    For n = 0 to # of iterations
    small Perturbation( )
    If Fitness (New _Config) > Fitness(Prev_Config)
        If Fitness (New_Config) > Best_Fitness
            Best_Fitness = Fitness (New_Config)
        End-If
        Prev_Config = New_Config
    End-IF
Return
Scaterring ( )
    $P_{scattering}$= 1 - (Fitness (New_Config) / Best _Fitness)
    If $P_{scattering}$ > random (0, I)
        Prev_Config = Random Solution
    Else
        Exploration ( )
    End-If
Return
```

2.3 Decision Tree Algorithm

A decision tree (DT) is a type of classifier that employs a tree-like structure to show how a decision is made. It is a replica that looks like a flowchart. There are branches, non-leaf nodes, and leaf nodes in a decision tree. The root node is the node at the very top, and each leaf node has a class label. A specific problem attribute is expressed by the non-leaf nodes. A decision tree's ability to solve any problem without the need for specific subject expertise is one of its many advantages [22], [29], [30]. The DT pseudo-code utilized in this work is displayed in algorithm 2.

Algorithm 2: Pseudocode of Decision Tree Algorithm

```
#Generate Decision Tree, S=sample, F=Feature
GenDecTree(S, F)
```

http://dx.doi.org/10.36596/jcse.v4i1.617

```
If stopping_condition(S,F) = true then
        Leaf = createNode()
        leaf-Label = classify(s)
        return Leaf
root = createNode()
root.test_condition =findBestSpilt(S,F)
V = (|v| v a possible outcomecfroot.test_condition)
For each value v ∈ V.
          S_v = {s\ root.test_condition(s) = v and s ∈ S};
          Child = TreeGrowth (S_v,F);
          Add child as descent of root and label the edge {root→child} as v
return root
```

2.4 Data Source, Description and Processing

Making a tool that can draw useful patterns out of the CHD data warehouse is the objective. The data was gathered from the UCI repository, which has to be cleaned and filtered because it has duplicate, inconsistent, and missing data. A data processing phase and the application of a data mining method are necessary to prevent the development of misleading or unsuitable patterns or rules. In this work, the CHD dataset was enhanced by removing redundant entries, converting and normalizing the values used to represent data in the database, filling in for missing data and compensating for missing data fields. A few pertinent attributes were added to the datasets created from the raw data. Additionally, data consolidation could be necessary to lower the number of datasets and the amount of memory and processing power that the data mining method needs.

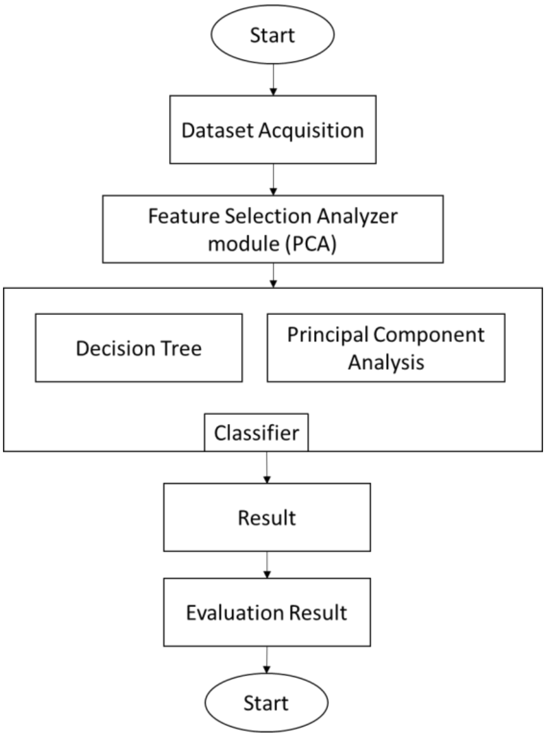


Fig 1. Proposed System framework

http://dx.doi.org/10.36596/jcse.v4i1.617

3. Results and Discussion

For the purpose of predicting heart disease, the proposed models used PCA and decision tree classifiers. The created platform used classifiers to determine if a heart disease situation is favorable or unfavorable in relation to the desired or predetermined outcome.

R programming was used to carry out and build the experimental setup; numerous functions were made and connected to a graphical user interface for user responsiveness. The produced systems utilized a variety of component settings in R studio to develop and output the data mining task's findings.

3.1 Dataset Configuration

The dataset was obtained online from a UCI research institute repository. A sample with eight attributes, seven predicting variables, and two class label has shown in Table 1. The class label represents the state of an individual's heart, which can be either positive or negative.

Table 1. Sample of Original Dataset Used

age	chest_pain	rest_bpress	blood_sugar	rest_electro	max_heart_ rate	exercice_ angina	disease
43	Asympt	140	F	Normal	135	yes	positive
39	atyp_angina	120	F	Normal	160	yes	negative
39	non_anginal	160	T	Normal	160	No	negative
42	non_anginal	160	F	Normal	146	No	negative
49	Asympt	140	F	Normal	130	No	negative
50	Asympt	140	F	Normal	135	No	negative
59	Asympt	140	T	left_vent_hyper	119	Yes	positive
54	Asympt	200	F	Normal	142	Yes	positive
59	Asympt	130	F	Normal	125	No	positive

3.2 Result Presentation

In a progressive mode, the implemented model and developed application for heart disease using PCA and Decision tree are shown below.

3.3 Normalization of Datasets

The data in this study was normalized by converting string variables to numeric variables in order to present the network with well-formatted system data. The normalized data sample is shown in table 2.

Table 2. Sample of Dataset Normalized

Age	ches2_pain	res2_bpress	blood_sugar	res2_elec2ro	max_hear2_ ra2e	exercice_ angina	disease
43	1	140	1	1	135	1	1
39	2	120	1	1	160	1	2
39	3	160	2	1	160	2	2
42	3	160	1	1	146	2	2
49	1	140	1	1	130	2	2
50	1	140	1	1	135	2	2
59	1	140	2	2	119	1	1
54	1	200	1	1	142	1	1
59	1	130	1	1	125	2	1

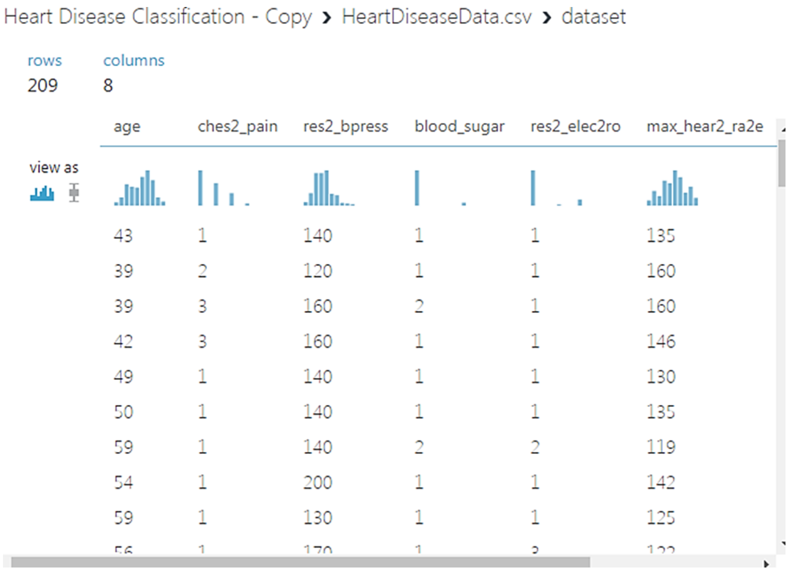
http://dx.doi.org/10.36596/jcse.v4i1.617

3.4 Training

The dataset was partitioned into three sets to create the model's training set (Training, validation and testing).The training took 70% of the whole dataset as it helps to formulate an experimental knowledge for the heart disease dataset used. The validation set contained 15 % of the dataset, this really help to improve generalization of the model and also validate the model's performance. The remaining 15% was allotted for the testing dataset this assume a fresh prediction by the classifiers, by discerning the target to predict new target.

3.5 Feature Selection

The process of obtaining an important feature subsets to be used for model development is called feature selection. At this stage, the original dataset's significant and instructive features were extracted using PCA to automatically choose the attributes. Sometimes, a single feature is viewed as irrelevant if it is studied alone, but when it is combined with other features, it becomes a important feature. This is as a result of the presence of interactions between features. Fig 2 and 3 depict sample of dataset before and after selection of relevant features.

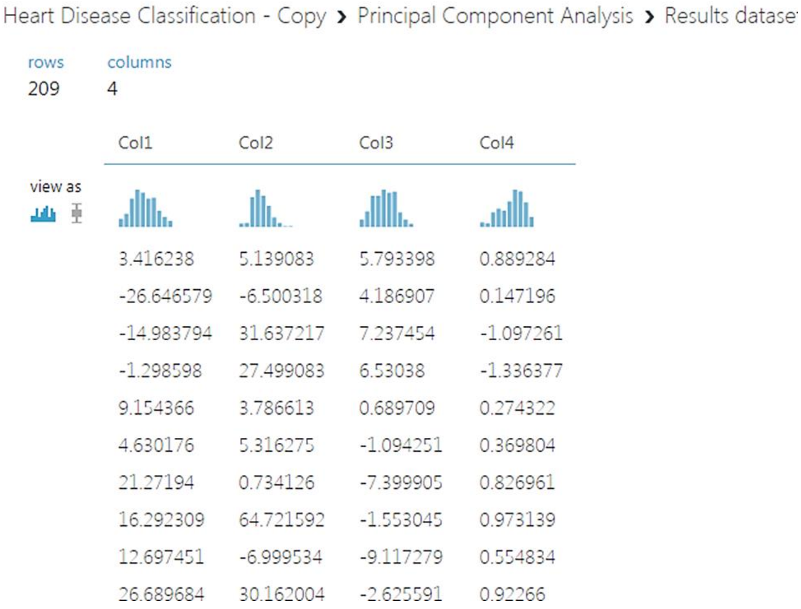


Heart Disease Classification - Copy › HeartDiseaseData.csv › dataset

rows 209
columns 8

age	ches2_pain	res2_bpress	blood_sugar	res2_elec2ro	max_hear2_ra2e
43	1	140	1	1	135
39	2	120	1	1	160
39	3	160	2	1	160
42	3	160	1	1	146
49	1	140	1	1	130
50	1	140	1	1	135
59	1	140	2	2	119
54	1	200	1	1	142
59	1	130	1	1	125
56	1	170	1	2	122

Fig 2. Heart disease dataset before feature selection.



Heart Disease Classification - Copy › Principal Component Analysis › Results datase

rows 209
columns 4

Col1	Col2	Col3	Col4
3.416238	5.139083	5.793398	0.889284
-26.646579	-6.500318	4.186907	0.147196
-14.983794	31.637217	7.237454	-1.097261
-1.298598	27.499083	6.53038	-1.336377
9.154366	3.786613	0.689709	0.274322
4.630176	5.316275	-1.094251	0.369804
21.27194	0.734126	-7.399905	0.826961
16.292309	64.721592	-1.553045	0.973139
12.697451	-6.999534	-9.117279	0.554834
26.689684	30.162004	-2.625591	0.92266

Fig 3. Feature selection module with PCA algorithm

http://dx.doi.org/10.36596/jcse.v4i1.617

3.6. Performance Measures of Principal Component Analysis

ROC curve was obtained by plotting the true positive rate (TPR) versus the false positive rate (FPR) at various levels of threshold. In other words, the true-positive rate can also be referred to as likelihood of detection, or recall or sensitivity. True Positives (TP) are successfully predicted positive values, showing that both the actual and projected class values are true. Also, when the real class is “NO” and the predicted class is “YES”, the false-positive rate (fall-out) occurs.

Fig 4 depicts the Receiver Operating Characteristics (ROC) result for the classified components which was extracted using PCA technique; it uses the ROC curve in studying the output of the classifier.

Confusion matrix is a measure employed to assess the quality of our ROC curve by examining AUC (Area under the Curve). A confusion matrix table is usually employed to determine the goodness of a classification model on a given set of test data for which the true values are already knew.

Fig 5 shows the confusion matrix of PCA which is used to evaluate the performance of PCA on heart disease prediction. 62 0ut 0f 63 samples used to test the PCA model are correctly classified and only one missed classified, which showed that the model is good for classification of heart disease.

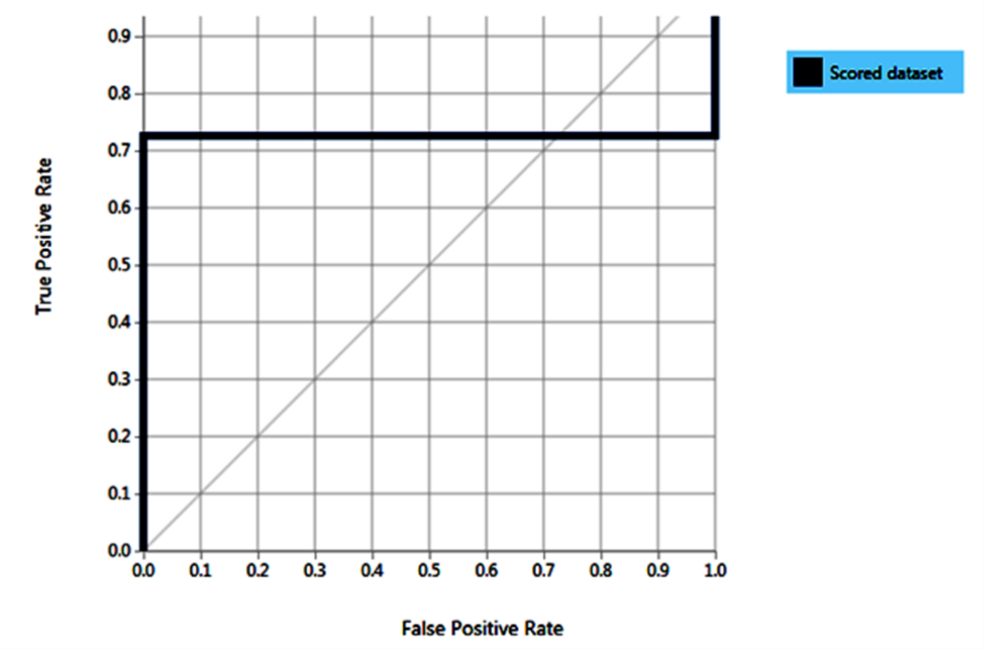
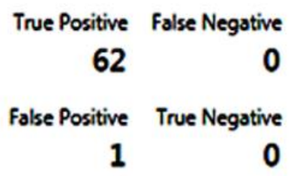


Fig 4. ROC curve of PCA Algorithm



True Positive	False Negative
62	0
False Positive	True Negative
1	0

Fig 5. Confusion matrix of the PCA algorithm

3.7. Performance Measures of Decision Tree Algorithm

Fig 6 displays the Receiver Operating Characteristics (ROC) result for the components that were identified and extracted using the decision tree technique. The ROC curve is used to analyze the classifier's output. Confusion matrix is a measure employed to assess the quality of our ROC curve by examining AUC (Area under the Curve).

The confusion matrix of the Decision Tree method, which was used to assess the effectiveness of the classifier in predicting heart disease, is shown in Fig 7. The decision tree model tested on 63 samples was able to accurately classify 46 of them, demonstrating above-average model performance for the categorization of heart disease.

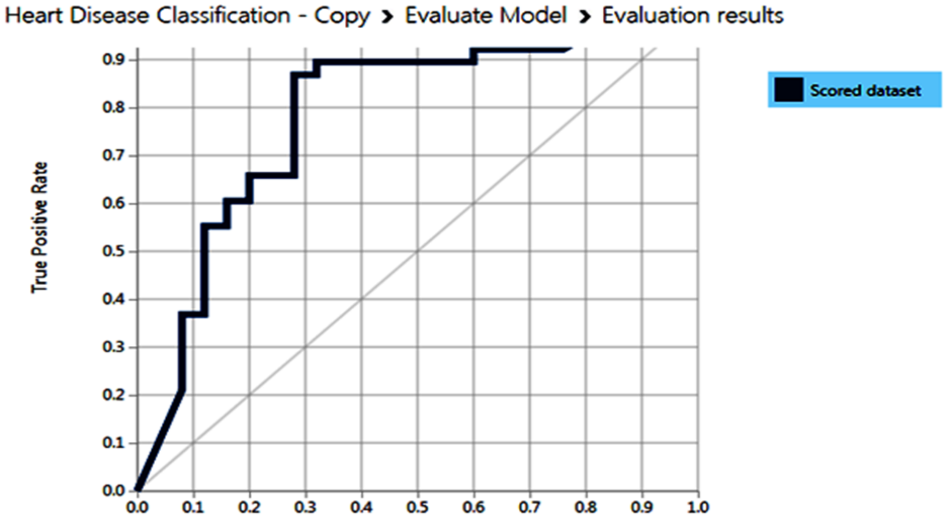
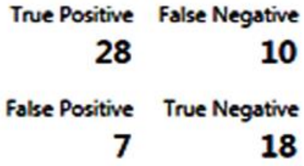


Fig 6. ROC curve of Decision Tree Algorithm



True Positive	False Negative
28	10
False Positive	**True Negative**
7	18

Fig 7. Confusion matrix of Decision tree algorithm

The PCA algorithm produced an AUC of 72.6%, Accuracy of 98.4%, F1 score of 99.2%, Precision of 98.4%, and Recall of 100%, according to Table 3. A graphic comparison is presented in Fig 8 for the Decision Tree Algorithm's 79.7% AUC, Accuracy of 73%, F1 of 76.7%, Precision of 80%, and Recall of 73.7%. According on these findings, PCA outperformed the Decision Tree algorithm.

Table 3. Performance of the PCA and Decision tree classifiers on heart disease predictions

Methods	AUC (%)	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)
PCA	72.6	98.4	99.2	98.4	100
Decision Tree	79.7	73.0	76.7	80.0	73.7

http://dx.doi.org/10.36596/jcse.v4i1.617

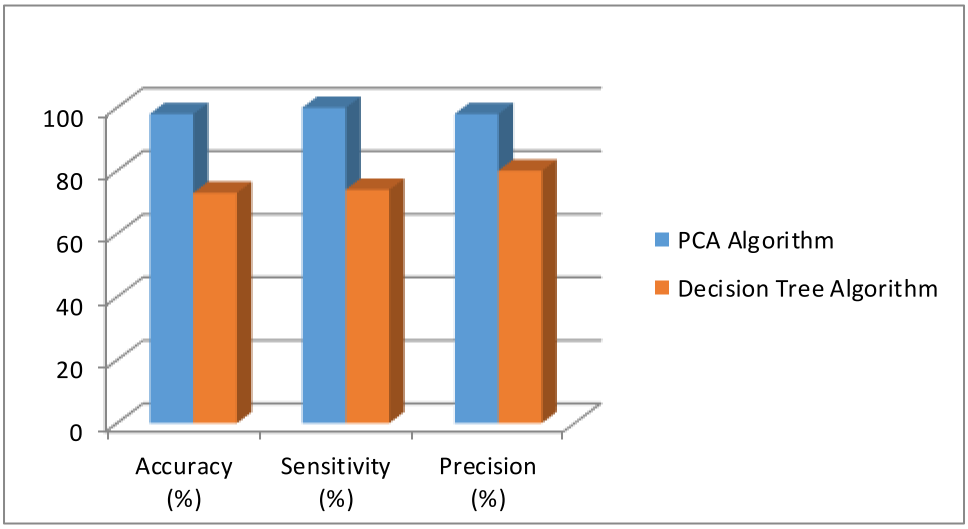


Fig 8. Performance metrics of the two classifiers

From the result obtained from this research work, it was revealed that the PCA performs better than Decision tree algorithm. It achieved a very good classification accuracy, precision, F1-score and Recall than the Decision tree algorithm.

4. CONCLUSION

Heart disease is a lethal condition, and a misdiagnosis of this condition can lead to grave complications, including cardiac arrest and death. The classification accuracy of the best model proposed for predicting heart disease yielded 98.4 %, and there is still much work to be done to close the gap of 1.4 % misclassified cases.

This work proved that cardiovascular disease cases can be effectively modeled and predicted using data mining approaches. The results of this study can help cardiologists diagnose heart problems more precisely. The model that was produced also recorded a high rate of specificity that makes it a helpful tool for junior cardiologists to identify patients who have a high likelihood of having the heart disease condition and refer them to senior cardiologists for additional evaluation.

References

[1] 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, p. 8137, 2020. https://doi.org/10.3390/app10228137.

[2] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," vol. 5, pp. 1–16, 2020. https://doi.org/10.1186/s12911-020-1023-5.

[3] WHO, "Cardiovascular Diseases," *World Health Organization*, 2021.

[4] M. A. Hambali, T. O. Oladele, K. S. Adewole, A. K. Sangaiah, and W. Gao, "Feature selection and computational optimization in high-dimensional microarray cancer datasets via InfoGain-modified bat algorithm," *Multimed. Tools Appl.*, vol. 1213, pp. 1–45, 2022. https://doi.org/10.1007/s11042-022-13532-5.

[5] S. M. R. Haque, A. Das, and S. I. Leon, "Heart Disease Prediction Using Data Mining," vol. 12, no. 03, pp. 3071–3078, 2021. https://doi.org/10.1088/1742-6596/1913/1/012099

[6] M. G. El-Shafiey, A. Hagag, E.-S. A. El-Dahshan, and M. A. Ismail, "A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest," *Multimed. Tools*

http://dx.doi.org/10.36596/jcse.v4i1.617

Appl., vol. 81, no. 13, pp. 18155–18179, 2022. https://doi.org/10.1007/s11042-022-12425-x.

[7] S. Palaniappan and R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques," *Int. J. Comput. Sci. Netw. Secur.*, vol. 8, no. 8, pp. 343–350, 2008. https://doi.org/10.1109/AICCSA.2008.4493524.

[8] R. Asha and R. G. Sophia, "Diagnosis Of Heart Disease Using Datamining Algorithm," *Glob. J. Comput. Sci. Technol.*, vol. 10, no. 10, pp. 1–6, 2010.

[9] M. Shouman, T. Turner, and R. Stocker, "Using decision tree for diagnosing heart disease patients," in *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*, 2011, pp. 23–30.

[10] N. Bhatla and K. Jyoti, "An analysis of heart disease prediction using different data mining techniques," *Int. J. Eng.*, vol. 1, no. 8, pp. 1–4, 2012.

[11] C. Dangare and S. Apte, "A data mining approach for prediction of heart disease using neural networks," *Int. J. Comput. Eng. Technol.*, vol. 3, no. 3, 2012.

[12] S. U. Amin, K. Agarwal, and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," in *2013 IEEE Conference on Information & Communication Technologies*, 2013, pp. 1227–1231. https://doi.org/10.1109/CICT.2013.6558288.

[13] V. Chaurasia and S. Pal, "Early prediction of heart diseases using data mining techniques," *Caribb. J. Sci. Technol.*, vol. 1, pp. 208–217, 2013.

[14] D. Chandna, "Diagnosis of heart disease using data mining algorithm," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 2, pp. 1678–1680, 2014.

[15] H. D. Masethe and M. A. Masethe, "Prediction of heart disease using classification algorithms," in *Proceedings of the world Congress on Engineering and computer Science*, 2014, vol. 2, no. 1, pp. 25–29.

[16] B. Venkatalakshmi and M. V. Shivsankar, "Heart Disease Diagnosis Using Predictive Data mining," *Int. J. Innov. Res. Sci. Eng. Technol.*, vol. 3, no. 3, pp. 1–5, 2014.

[17] B. Bahrami and M. H. Shirvani, "Prediction and diagnosis of heart disease by data mining techniques," *J. Multidiscip. Eng. Sci. Technol.*, vol. 2, no. 2, pp. 164–168, 2015.

[18] U. Shafique, F. Majeed, H. Qaiser, and I. U. Mustafa, "Data mining in healthcare for heart diseases," *Int. J. Innov. Appl. Stud.*, vol. 10, no. 4, p. 1312, 2015.

[19] M. Abdar, S. R. N. Kalhori, T. Sutikno, I. M. I. Subroto, and G. Arji, "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases.," *Int. J. Electr. Comput. Eng.*, vol. 5, no. 6, 2015. http://doi.org/10.11591/ijece.v5i6.pp1569-1576.

[20] L. Verma, S. Srivastava, and P. C. Negi, "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data," *J. Med. Syst.*, vol. 40, no. 7, pp. 1–7, 2016. https://doi.org/10.1007/s10916-016-0536-z.

[21] J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," in *2016 international conference on circuit, power and computing technologies (ICCPCT)*, 2016, pp. 1–5. https://doi.org/10.1109/ICCPCT.2016.7530265.

[22] G. Manogaran, R. Varatharajan, and M. K. Priyan, "Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system," *Multimed. Tools Appl.*, vol. 77, no. 4, pp. 4379–4399, 2018. https://doi.org/10.1007/s11042-017-5515-y.

[23] K. Mathan, P. M. Kumar, P. Panchatcharam, G. Manogaran, and R. Varadharajan, "A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease," *Des. Autom. Embed. Syst.*, vol. 22, no. 3, pp. 225–242, 2018.

[24] S. Poornima, S. Sanjay, and S. P.-J. Gayatric, "Effective heart disease prediction system using data mining techniques," *Int. J. Nanomedicine*, pp. 121–124, 2018.

[25] C. B. Gokulnath and S. P. Shantharajah, "An optimized feature selection based on genetic approach and support vector machine for heart disease," *Cluster Comput.*, vol. 22, no. 6, pp. 14777–14787, 2019. https://doi.org/10.1007/s10586-018-2416-4.

[26] S. Maji and S. Arora, "Decision tree algorithms for prediction of heart disease," in *Information and communication technology for competitive strategies*, Springer, 2019, pp. 447–454. https://doi.org/10.1007/978-981-13-0586-3_45.

[27] N. Absar *et al.*, "The efficacy of machine-learning-supported smart system for heart disease

prediction," in *Healthcare*, 2022, vol. 10, no. 6, p. 1137. https://doi.org/10.3390%2Fhealthcare10061137.

[28] M. A. Hambali and R. G. Jimoh, "Performance Evaluation of Principal Component Analysis and Independent Component Analysis Algorithms for Facial Recognition," *J. Adv. Sci. Res. Its Appl.*, vol. 2, pp. 47 – 62, 2015.

[29] S. O. Abdulsalam, Y. K. Saheed, M. A. Hambali, T. T. Salau-Ibrahim, and N. B. Akinbowale, "Student's Performance Analysis Using Decision Tree Algorithms," *Ann. Comput. Sci. Ser. Ser.*, vol. 15, no. 1, pp. 55–62, 2017.

[30] S. O. Abdulsalam, A. N. Babatunde, R. S. Babatunde, and M. A. Hambali, "Comparative Analysis of Decision Tree Algorithms for Predicting Undergraduate Students' Performance in Computer Programming," vol. 2, no. 20, pp. 79–92, 2015.