Data Mining Approach to Coronary Artery Disease Diagnosis

Afeni B. O[1]., Lawal O.O[1]. and Akinyemi S. G[2]
[1]Department of Computer Science, Joseph Ayo Babalola University (JABU), Ikeji Arakeji, Nigeria
[2]Department of Statistics, Auchi Polytechnic, Auchi, Nigeria
Corresponding Author: boafeni@jabu.edu.ng

ABSTRACT

The prevalence of coronary artery disease (CAD) is on high increase all over the world. The disease often results to death and has been categorized as one of the world's most predominant cause of death. The high mortality rate from CAD is as a result of many factors which include lack of accurate diagnosis, shortage of medical specialists and services, poor interpretation of laboratory results and late diagnosis. These inadequacies have prompted the development of computer aided diagnostic systems for CAD using data mining approach. Data mining is an advanced technology, which is used to analyze large volumes of datasets and extracts patterns that can be converted to useful knowledge. Two filter-based feature selection methods namely Information Gain and Chi-Square methods were used to identify the most relevant features for the diagnosis. After which two data mining techniques - K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) were used. The implementation of the models was carried out in Python environment. The highest accuracy obtained from the resulting models on test dataset was 88.2% for SVM as against the highest accuracy of KNN which is 85.3%. The result shows that the proposed system performs well on the test dataset. Thus, the system is good for diagnosis of CAD and could be of immense benefits to the health sector and every individual.

Keywords: *Coronary Artery Disease, Data Mining, Diagnosis, Predictive Model*

INTRODUCTION

In every fields of human endeavor, data are being generated and collected and accumulated at a huge pace. There is a pressing necessity for a new generation of computational tools that can be of assistance in extracting relevant information from the rapidly increasing volumes of data. The essential part of this process is the application of explicit data mining methods for discovery and extraction of patterns from data (Bhatla and Jyoti, 2015). Data mining is an interdisciplinary field of computer science which involves the computational process of discovering patterns in large dataset. It cut across artificial intelligence, machine learning, statistics and database systems (Kumar and Sonia, 2017). The main objective of data mining process is to extract relevant information from a dataset and then transform it into a better and understandable form for further use. In healthcare, data mining entail steps that allows the extraction of patterns from preprocessed data by the application of specific algorithms.(or by basically applying specific algorithms).

The most vital and hardest - working muscle in human body is the heart. The heart with blood vessels makes up the cardiovascular

system. The basic function of the heart is to pumps blood into every cell of the human body. It can be stated that heart muscle is the engine of the human body (Palaniappan and Awang, 2008). One of the most predominant health issues is coronary artery disease. This disease develops when the major blood vessels that supply human heart with blood, oxygen and nutrients (coronary arteries) become diseased or damaged. The arteries usually get blocked with cholesterol-containing deposits (plaque). Hence, it is very essential to predict such diseases through appropriate symptoms.

There are many types of algorithms presently being used for disease prediction which includes Decision Trees, Naïve Bayes, Multilayer perceptron and Support vector machine. Unfortunately, all medical personnel do not possess proficiency in every area of specialty and likewise there is a shortage of specialist especially in other area of cardiology. Therefore, an automatic medical diagnosis system would possibly be remarkably beneficial for bringing the efficient and accurate result. Appropriate computer-based information and decision support systems can help to minimize the negative effect of heart disease. In this work a performance comparison of heart disease diagnosis is executed with the help of two feature selection methods, K- Nearest Neighbor (KNN) and Support Vector Machine (SVM).

REVIEW OF RELATED WORKS

Several researchers have been exploring the use of data mining techniques to diagnose heart disease. Some factors related to heart disease includes and not limited age, sex, blood sugar, chest pain, blood pressure, cholesterol. These are some of the factors used in the diagnosis of heart disease in patients. Soni, (2011) did a survey of current techniques of extraction from databases using data mining techniques for diagnosis of heart disease. The authors used Naive Bayes, Decision Trees and K-Nearest Neighbor. The limitation of the work was that the classification based on clustering did not perform well. Deepika (2011) used pruning classification association rule (PCAR) derived from Apriori algorithm. The proposed method deletes minimum frequency item with minimum frequency item sets and removes infrequent instances from sets of instances then the frequent item set was revealed and used.

Jabbar *et al.* (2012) presented an efficient associative classification based genetic algorithm for heart disease prediction. The goal for using genetic algorithm to predict disease from large dataset was to get the best attribute set. As revealed in the work, there are limitations in the prediction of the disease using data mining approach. Reducing in the set of attributes (features) can make it less complex with better accuracy. Anooj *et al.*, (2012) worked on diagnosis of heart disease using a weighted fuzzy rule-based system. The system was designed to automatically retrieve knowledge from the patient's data. Mamdani fuzzy inference system was used to build the weighted fuzzy rules. Bhatla *et al.*, (2012) in their work also proposed to evaluate various data mining techniques used for heart disease prediction. Findings from the work revealed that SVM outperformed other data mining techniques used. Another observation from the work is that decision tree also gave a

fairly good accuracy with the help of genetic algorithm and feature subset selection. (Anbarasi, *et al.*, 2010). Sethukkarasi and Kannan (2012) designed a novel neuro fuzzy technique with genetic algorithm for feature extraction. A radial basic function neural network was constructed in the work with five input, training and normalization in hidden layer and output layer with one node.

Khaleel *et al.* (2013) presented an approach to diagnose heart diseases with Apriori data mining technique. A graphical representation was also used to visualize the techniques. A prototype was developed to validate the efficiency of the approach. It was observed that the prototype can be suitable in real world. Methaila *et al.*, (2014) in their work applied three data mining classification modeling techniques in addition to weighted association Apriori algorithm and MAFIA algorithm for heart disease prediction. However, the performance evaluation of the work was not explicitly stated. Goyal and Chhillar (2015) present a work on heart disease prediction using K-means and apriori algorithm. The work also presented the issues in diagnosing the diseases and analyzed the results. Bahrami and Shirvani (2015) also studied different classification techniques for diagnosis heart disease. Classifiers such as Decision Tree, KNN, and Naive Bayes were used for the classification. The result revealed that Decision tree outperformed others in terms of classification accuracy.

METHODOLOGY

The main objective of this work is to design predictive models for coronary artery disease using data mining technique. The predictive models were designed using two different data mining algorithms which are K-Nearest Neighbor (KNN) and Support Vector Machine (SVM). It was implemented in Python environment. Dataset used in this work was obtained from University of California, Irvine (UCI) online repository (https://archive.ics.uci.edu).

Two experiments were conducted in this study and for the two experiments, three scenarios were considered per experiment. The first scenario contains all the attributes in the dataset, the second scenario contained attributes selected with information gain feature selection method, while the third scenario contained attributes selected with chi-square feature selection method. With two experiments and three different scenarios a total of six predictive models were designed. The selected features were fed to the classifiers. The models were evaluated based on five criteria described below

I. Accuracy: Accuracy is the ratio of correct predictions to the total predictions given as

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

ii. Precision: The ratio of True positives to the Overall positive predictions. It is otherwise referred to as precision or positive predictive value given as:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

iii. False Alarm Rate (FAR): This is simply the ratio of false positives (false alarms) to the total negatives. It is otherwise known as the false positive rate given as:

$$FAR = \frac{FP}{TN + FP} \quad (3)$$

iv. Recall: Also known as the Sensitivity or true positive rate is the ratio of True

positives to the total Positives i.e.

$$Recall = \frac{TP}{FN + TP} \quad (4)$$

In order to evaluate the performance of the algorithms used, there was need to plot the results of the classification on a confusion matrix (Figure 1).

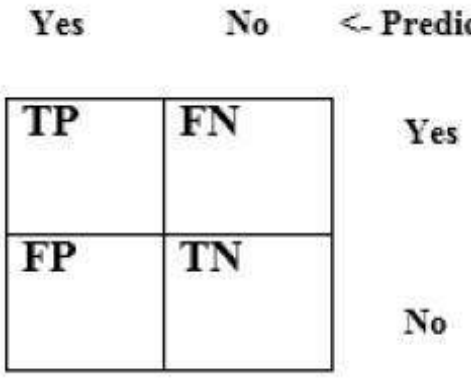

Figure 1: Diagram of a Confusion Matrix

True positives (TP) are the correctly classified Yes cases, False positives (FP) are incorrectly classified No cases, True negatives (TN) are correctly classified No cases; and False negatives (FN) are incorrectly classified Yes cases. The true positive/negative and false positive/negative values recorded from the confusion matrix can then be used to evaluate the performance of the prediction model. To determine most relevant features of the dataset, the ranked features were grouped into 3 groups. The first group contain the first top 5 features, the second group contain the first top 10, and the last group contain all 13 features (without feature selection). The grouped features passed to each classifier for classification, and the group whose performance is best during classification was chosen as most relevant features. The Machine Learning Algorithms used in this work are discussed below:

(a) K-Nearest Neighbor (KNN)

KNN is a learning classifier that classifies unlabeled examples based on their resemblance to examples in the training set. The records were represented as a point in a *i*-dimensional space where *i* is the number of attributes. To find a class label for a test data sample, the nearest CAD training data point from the test sample in the *i*-dimensional space is located using a proximity measure and the target class label of the nearest training data point is assigned as the predicted target class for the test data point. However, when the data points are in between the boundaries of two different CAD instances, an algorithm with a majority voting method is used to measure the nearness of the data points with more than two dimensions and assign a class label. The number of neighbors used to assign the target class of the instance query is identified by the value of *n*. KNN is based on Euclidean distance between the training set and the testing set. Given that is the instance to be classified ranging from 1 to *n*, is the total number of instances in a data set ranging from 1 to *k* with same number of features. The Euclidean distance between two tuples, $X_1 = (x_{11}, x_{12}, \ldots, x_{1n})$ and $X_2 = (x_{2i}, x_{22}, \ldots, x_{2n})$ is defined as:

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \quad (5)$$

(b) Support Vector Machines (SVM)

SVM is a one of the data mining methods. It recognizes patters and data in a classification

task. SVM classifies and separates similar data by finding the best hyper plane that separates all data points of one class from other class. From the perspective of statistical learning theory, the motivation for considering a binary classifier SVM comes from the theoretical bounds on the generalization error. These generalization bounds have two important features: upper bound is independent of size of the input space, and the bound is minimized by maximizing the margin between the hyperplane separating the two classes and the closest data point to each class – called support vectors. Closest points are called support vectors because they support where the hyperplane should be located. That is, moving the non-support vectors will not shift the hyperplane, whereas moving the support vectors will shift the hyperplane. Given a training dataset: θ, containing data feature vectors x_i and the corresponding data labels y_i, in the form of

$$\theta = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n) \quad (6)$$

where $x_1 \in$ ^{m}, m is a dimension of the feature (real) vector, $y_1 \in \{0,1\}$ and n is the number of instances in the dataset. We assume (x) is some unknown function to classify the feature vector xi.

$$g(x): \ ^m \rightarrow \{0,1\} \quad (7)$$

In SVM method, optimal margin classification for linearly separable patterns is achieved by finding a hyperplane in m dimensional space. The linear classifier is based on a linear discriminant function of the form,

$$f(x) = \sum_i wi\, xi + b \quad (8)$$

RESULTS AND DISCUSSION

The predictive models were designed using two different data mining techniques (KNN and SVM) and the performance was evaluated. Three scenarios were considered as earlier mentioned. The intention here is to investigate the effect of feature selection on the performance of the models. In the first scenario, the algorithms were run on a full training set containing 802 instances with 13 attributes. In the second scenario, the two algorithms were run with features selected with information gain feature selection method. While the third scenario, the algorithms were run with attributes selected via chi-square feature selection method. The confusion matrixes and the detailed performance measures of the developed model is presented in Table 1 and Table 2 respectively. Figure 3 shows the performance evaluation chart of the developed models.

Table 1: Confusion Matrixes for Experiments

Models	Actual	Predicted	
		Yes	No
KNN with all attributes	Yes	318 (39.7%)	78 (9.7%)
	No	79 (9.8%)	327 (40.8%)
KNN with Information Gain Feature Selection	Yes	332 (41.4%)	64 (8%)
	No	62 (7.7%)	344 (42.9%)
KNN with Chi-Square Feature Selection	Yes	335 (41.8%)	61 (7.6%)
	No	57 (7.1%)	349 (43.5%)
SVM with all attributes	Yes	318 (39.7%)	78 (9.7%)
	No	72 (9%)	334 (41.7%)
SVM with Information Gain Feature Selection	Yes	344 (42.9%)	52 (6.5%)
	No	50 (6.2%)	356 (44.4%)
SVM with Chi - Square Feature Selection	Yes	347 (43.4%)	49 (6.1%)
	No	46 (5.7%)	360 (44.9%)

Table 2 Detailed Performance Measures for Experiments

MODEL	Correct Classification	Accuracy (%)	Precision (%)	TP Rate (%)	FP Rate (%)	FAR (%)
KNN with all attributes	645 (80.4%)	80.4	80.1	80.3	80.7	19.3
KNN with Information Gain Feature Selection	676 (84.3%)	84.3	84.3	83.8	84.3	15.7
KNN with Chi-Square Feature Selection	684 (85.3%)	85.3	85.5	84.6	85.1	14.9
SVM with all attributes	653 (81.4%)	81.3	81.5	80.3	81.1	18.9
SVM with Information Gain Feature Selection	700 (87.3%)	87.3	87.3	86.9	87.3	12.7
SVM with Chi-Square Feature Selection	707 (88.2%)	88.2	88.3	87.6	88	12

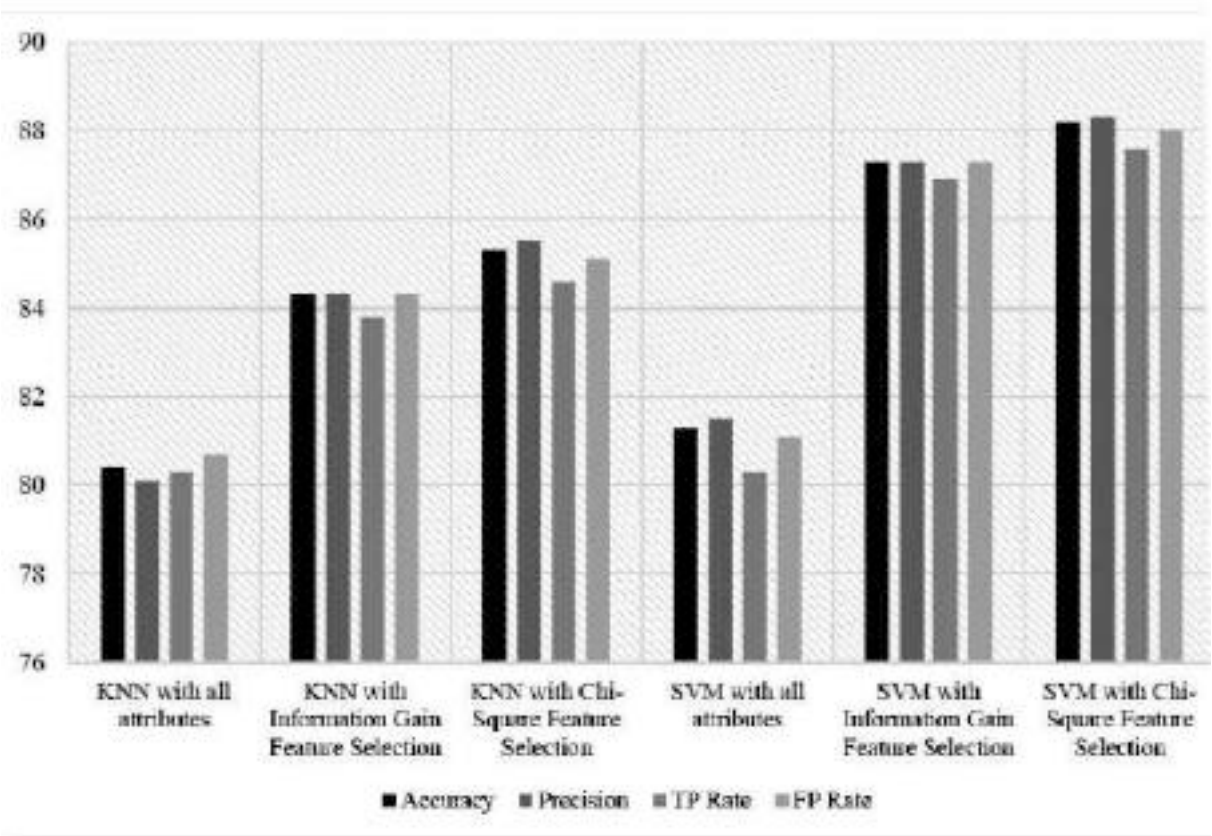

Figure 3: Performance evaluation Chart of the Predictive Models

From the performance evaluation of the developed model, it was revealed that Chi-Square features selection method is better with SVM model gave the highest accuracy and least False Alarm Rate (FAR) of 88.2% and 12% as against that of KNN models which was 85.3% and 14.9%. The variables identified by the predictive model using the feature selection algorithms can help to provide and handle the relationship that exist between the attributes with respect to the diagnosis of CAD.

CONCLUSION

The models trained and classified well on testing dataset with SVM model attaining a very good accuracy as well as low false alarm rate of roughly (7.3%). The features selected via two features selection techniques were considered as the most relevant features when tested with SVM model as the model produced a prediction accuracy of 88.2 % on Chi- Square and 87.3% on information gain. Conclusively, SVM model performed better on the dataset compared to that of KNN. Sequel to the performance of the prediction models for CAD diagnosis, a better perception of the relationship between the variables relevant to CAD diagnosis was estimated. The models can be integrated into the available Health Information System (HIS) which covers and manages clinical information that can be fed to the diagnosis classification model thus improving clinical decisions. It is recommended that a continual evaluation of attributes monitored during diagnosis of CAD be made in order to increase the number of information relevant to developing an improved prediction model for the disease using the feature selection method and data mining approach.

REFERENCES

Anbarasi, M. Anupriya, E. and Iyengar, N, (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. International Journal of Engineering Science and Technology, vol. 2(10)

Anooj, P. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University – Computer and Information Sciences* vol. 24, page 27 – 40.

Bahrami, B. and Shirvani, M. (2015). Prediction and Diagnosis of Heart Disease by Data Mining Techniques. *Journal of Multidisciplinary Engineering Science and Technology* (JMEST), vol. 2(2).

Bhatla, N. and Jyoti, K. (2012). An Analysis of Heart Disease Prediction Using Data Mining Technique. *International Journal of Engineering Research and Technology (IJERT),* vol 1 (8).

Deepika, N. (2011). Association Rules for Classification of Heart Attack Patients. *International Journal of Advanced Engineering Sciences and Technologies (IJAEST)*, vol. 11(2), pp 253-257.

Goyal, S. and Chhillar, R. (2015). A Literature Survey on Applications of Data Mining Techniques to Predict Heart Diseases, International Journal of Engineering Sciences Paradigms and Researches (IJESPR), vol. 20(01)

Jabbar, M., Chandra, P. and Deekshatulu, P. (2012). Heart Disease Prediction System using Associative Classification and Genetic Algorithm. *International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies.*

Methaila, A. Kansal, P. Arya, H. and Kumar, P. (2014). Early Heart Disease Prediction Using Data Mining Techniques. *Journal of Computer Science and Information Technology,* pp. 53–59.

Palaniappan, S. and Awang, R. (2008). Intelligent Heart Disease Prediction System Using Data Mining Techniques. *International Journal of*

Computer Science and Network Security, vol.8(8).

Sethukkarasi, R. and Kannan, P. (2012). An Intelligent System for Mining Temporal rules in Clinical database using Fuzzy neural network. *European Journal of Scientific Research*, vol. 70(3), pp. 386 – 395.

Soni, J. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*, vol. 17 (8),