An Efficient and Intelligent Machine Learning Model for Early Heart Disease Assessment Using Significant Risk Attributes

[1]Sami Alshmrany and [2]Syed Immamul Ansarullah
[1]Faculty of Computer Science and Information Systems
Islamic University of Madinah, Saudi Arabia
[2]Computer Science and Information System
Maulana Azad National Urdu University, India

Abstract
Heart disease is emerging as the single most critical reason of mortality worldwide and is one of the costliest chronic conditions. Regardless of heart diseases damaging complications, it is the most preventable and controllable disease; therefore, it is important to predict it ahead of time. Considering the mortality rate of heart disorder and its rising health care costs, it is important to predict this malignant disease at its earliest. There are existing cardiac disorder risk assessment models however they are costly and operationally complex that restrains their use in rural areas and at public-level screening evaluations. To overcome these drawbacks of the prevailing heart disease risk systems, we develop a heart disease risk assessment model that can be utilized for public-standard screening to recognize patients at a high risk of heart disease and produce knowledge to facilitate initial intervention and enhance patient's health. The developed risk assessment model uses random forest, support vector machine and decision tree machine learning techniques. The developed risk model's efficiency is checked using various model and medical metrics. Experimental results show that the random forest risk assessment model outperforms other proposed risk models with the highest recognition rate, precision, sensitivity and AUROC score of 90.42%, 91.97%, 89.75% and 94%. As per our knowledge the experimental results obtained are highest than the published values in the literature. The developed risk model is applicable in where people lack the facilities of the integrated primary medical care technologies.

Key words:
Machine Learning Techniques, Data Mining Approaches, Classification Techniques, Heart Disease Prediction

1. Introduction

Heart disease is a type of cardiovascular disease (CVD) that is the chronic condition ascribed to the heart and blood circulation system [1]. The heart disease conditions are one of the foremost health and socioeconomic challenges at the present times because it drags the world population towards the high mortality rate and cause the immense damage on the world economy [2]. Heart disease is the leading contributors to global mortality rate which causes 17.9 million deaths per year [3]. The primary factors that drag the world to this outbreak are the fundamental heart disease risk attributes. Healthcare domain experts suggest that if there is no cessation of heart disease risk attributes, the fatality rates and economic burden associated with this disease may grow and exceed all other conditions in the forthcoming years. Predicting heart disease based on the association and causal relationship between its parameters is an intricate process and comprises a myriad of influences. To predict this malignant disease at its earliest and to scale down its health and economic burden researchers apply machine learning algorithms on an inundated data of healthcare industry. Machine learning techniques evaluate the unbridled healthcare data to identify useful patterns and observations that allow us to make nontrivial predictions on new data. Considering the damaging consequences of cardiac disease, we developed a risk examination model that would be applicable to predict cardiac disorder patients at its initial stages with optimal accuracy.

2. Research Background

In recent times, researchers made decisive contributions to heart disease identification using various machine learning technique which are explained as follows:

Palaniappan and Awang (2008) developed a risk evaluation model using the basic machine learning classification techniques [5]. This model extract useful hidden patterns related to the heart disease. The model answers complicated and intricate questions which traditional risk model fail to answer. The developed risk model uses the Cleveland heart disorder dataset that consists of 909 patient records and 15 risk characteristics. The model is developed on the .NET platform, and for communication, the Data Mining Extension query language and functions are used. The model performance is checked through lift chart and classification matrix. Experimental results show that the Naive Bayes risk evaluation model outperforms the Neural Network and Decision Tree models.

Patil and Kumaraswamy (2009) developed an efficient approach for cardiac disorder prediction using a K-Means clustering algorithm [6]. The risk model is developed in Java language on the benchmark Cleveland cardiac disorder dataset. Researchers use the Maximal Frequent

Manuscript received January 5, 2020
Manuscript revised January 20, 2020

Itemset algorithm to obtain the significant data insights that are most appropriate for heart disease. After deriving the frequent heart disease risk patterns, the weights assigned to the patterns are calculated, and the pattern with a significant weight higher than a predefined threshold value is utilized for the early risk prediction. The significant weighted patterns are pruned and verified by medical domain experts.

Anbarasi, Anupriya, and Iyengar (2010) developed a cardiac disorder risk evaluation system to accurately predict the risk using only the significant subset of risk features [7]. Researchers use the genetic machine learning algorithm to check the significant attributes that contribute more towards heart disease risk prediction. The risk model is developed using the Classification by Clustering, Naive Bayes and Decision tree techniques. The experimental results describe that the decision tree model outperform other risk assessment models. Researchers use the genetic algorithm to select the significant subset of risk features; the genetic algorithm begins with searching zero attributes, and then an initial population and finally ends with significant randomly generated heart disease risk rules. Setting the values of cross over probability to 0.6 and mutation probabilities to 0.033, the genetic search algorithm come up with six significant risk features that contribute high for the initial cardiac disorder prediction.

Shouman, Turner, and Stocker (2012) developed a K Nearest Neighbor risk evaluation model using the Cleveland heart disease dataset to detect cardiac disorder patients well in advance with optimal accuracy [8]. Initially, the value of K is set to 1 and then iteratively incremented till the upper limit of 13 and when k=7 the highest accuracy and specificity of 97.4% and 99% are achieved respectively. In this work, researchers discovered that applying the voting technique did not show any progress in the precision even after estimating different values of parameter k.

Al-Milli (2013) developed a cardiac disorder risk prediction model by applying the back-Propagation Neural Network algorithm [9]. The researcher uses the Cleveland benchmark dataset consisting of 13 medical attributes and MATLAB tool to build the risk model. After parameter settings, the experiments were run 10000 iterations. In the MATLAB tool, the risk evaluation model is executed 11 times; however, each run provided varying results. From the experimental results, it is found that when the model was executed 10th time, the highest variance from the training and testing process was achieved. The researcher used the box plot representation to illustrate the distribution of solution quality for training and testing datasets. In both cases, there is less dispersion of the output data, which demonstrates that it is a robust algorithm. The experiments conducted showed optimal performance compared to similar approaches of state of the art.

Masethe Hlaudi and Masethe Mosima (2014) designed a model to predict and classify heart attacks by using J48, Naive Bayes, Simple CART, REPTREE, and Bayes Net data mining algorithms [10]. The patient dataset used to build the heart disease model is collected from the health care professionals in South Africa that have 490 instances and 11 attributes. They use the WEKA tool for the prediction of heart disease. Researchers applied the stratified 10-fold cross-validation on the dataset for estimating the unbiased results. From the experimental results, it is found that the results did not provide any remarkable differences in heart disease prediction when different classification algorithms were applied.

Jabbar, Deekshatulu, and Chandra (2015) proposed a new approach that combines K- Nearest Neighbour and the Genetic classifier for effective classification to predict cardiac disorder victims [11]. Genetic search is applied as a goodness measure to reduce the insignificant and inappropriate risk features and to grade the features which contribute more towards classification. The less significant features are excluded, and the classifier is designed based on the classified features. The performance of the developed method is verified with six medical and one non-medical dataset. Among these seven heart disease datasets, one dataset is collected from different hospitals in Andhra Pradesh, INDIA, and the rest of the sex datasets are obtained from the UCI machine learning repository. The Experimental outcomes show that the classifier increases the efficiency of heart disease diagnosis.

Ngueilbaye, Lei, and Wang (2016) used Naive Bayes and Support Vector Machine classification algorithms for the initial prediction of cardiac disorder patients [12]. To check the performance of the applied classifiers, researchers used various measures like probability and classification accuracy. Experimental results show that the Naive Bayes algorithm outperforms the SVM model. The small dataset of 315 instances was collected from different hospital databases.

After reviewing the literature we found some research gaps in the existing models which are described as follows

- Most of the existing heart disease risk evaluation models lack generalization capability.
- The existing risk evaluation tools help in classifying victims at risk of heart disease; however, there is not known performance accuracy for them.

To overcome these research limitations, we develop an effective, low-cost heart disease evaluation model using significant risk attributes that would predict heart disease at its earliest with optimal generalization capability and accuracy.

3. Research Methodology

The knowledge discovery from data (KDD) research methodology is followed to build a heart disease risk evaluation model. We use the Jupyter notebook web application, pandas, Scikit-learn, Matplotlib, and Numpy libraries to build the risk model. Below given Figure 1shows the research methodology followed for the initial prediction of heart disease [13].

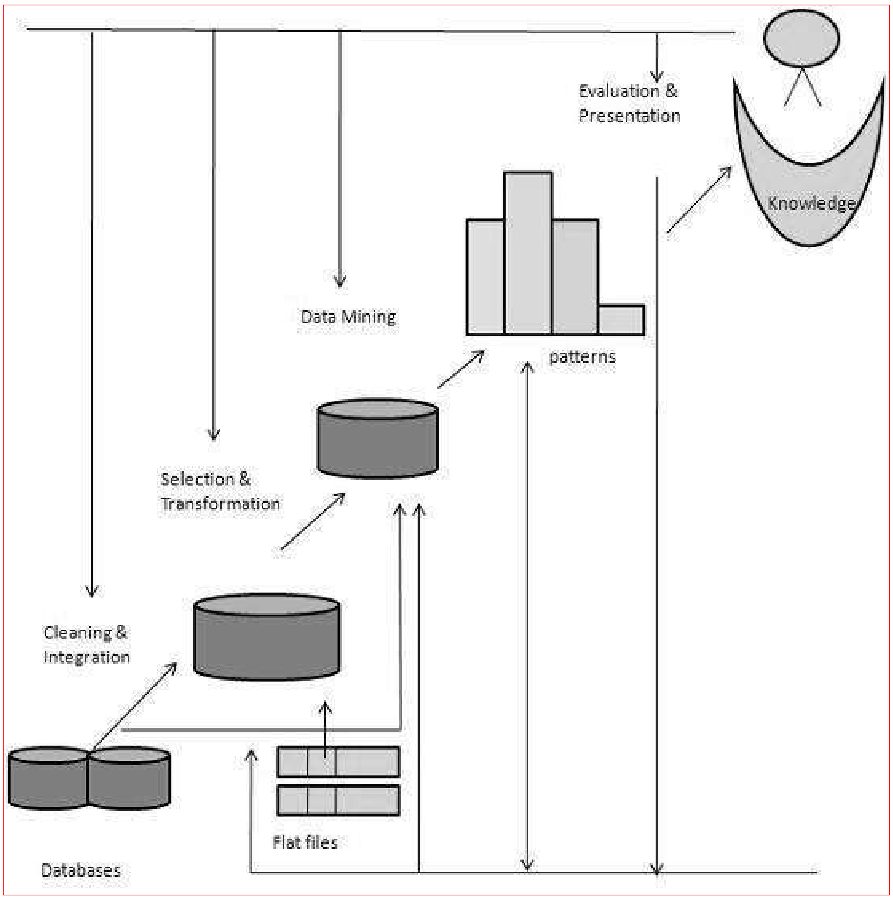


Fig. 1 The Process of Knowledge Discovery in Data [Source: Han and Kamber]

The model is developed on primary data that is obtained from various heterogeneous data resources like govt hospitals, and private clinics of Hyderabad (INDIA) through quantitative data collection methods (interviews). This heart disease dataset comprises 7300 patient records having fourteen (14) risk attributes as described in the below-given table 1. After Exploratory Data Analysis (EDA) we found that the heart disease dataset is noisy and includes numerous missing features signified with null values. The null values are filled with data imputation methods. The heart disease dataset attributes are organized into two types nominal and numeric. For example, the Gender attribute values as “male” and “female” represent the nominal attribute, and the “age” attribute values as 70 years represent a numeric attribute [13].

Table 1: Description of the Heart Disease Risk Features

Features	Data Type	Features with subsequent values and explanation
Age	Numeric	Represents the age of a patient in the number of years
Sex	Nominal	Represents sex of a patient, 0= Female, and 1= Male
BMI	Numeric	Represents the total Body Mass Index
Blood Pressure	Numeric	Represents the Systolic & Diastolic BP of the patient in mmHg
Healthy Diet	Nominal	Is the patient taking a nutritious diet? It is represented as 0=Following, 1=Occasionally and 2= Not Following
Physical Activity	Nominal	Whether the patient is doing exercise or not. 0= No Exercise, 1= Regular Exercise and 2= Occasionally
Alcohol Consumption	Nominal	How often the patient drinks alcohol, and it is represented as 0= Non-Alcoholic, 1= Occasionally and 2= Alcoholic
Smoking	Nominal	Whether the patient is smoking or not 0= Non-Smoker, 1= Regular and 2= Occasionally smoking
Diagnosis	Nominal	0= No and 1= Yes

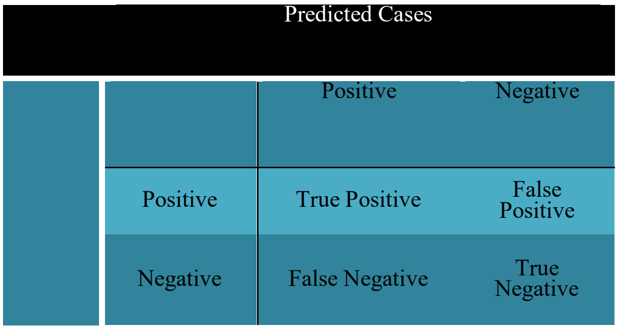
During EDA it is found that the independent attributes of the heart disease dataset are loosely correlated with one another. This loose correlation among independent heart disease attributes is a good sign to improve the performance of the model. However, if the attributes in a dataset are tightly correlated (called multicollinearity), then change in one variable can lead to change to another variable that can deteriorate the performance of an algorithm [144]. Correlation among the attributes does not mean causation hence, the strong relationship among attributes should be evaluated significantly.

4. Experimental Results of the Proposed Heart Disease Risk Evaluation Model

The real progress of any model is evaluated through its performance evaluation measurements. In this research work we check the overall accuracy of classifiers using 10-Fold cross validation, confusion matrix and AUROC measures. The accuracy evaluators are discussed as follows:

- Confusion Matrix: The error matrix is a principal source of performance measurements in classification problems. Below given table 2 shows the two class classification confusion matrix. The different values in the confusion matrix are as follows: True Positive, False Positive, True Negative and False Negative. By using these values we evaluate the performance of a developed model.

Table 2: contingency matrix for two class classification



		Predicted Cases	
		Positive	Negative
	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

- Accuracy: Accuracy is the total percentage cases that are correctly classified by an algorithm [15].

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

- Sensitivity: Sensitivity is the ratio of diseased cases that are correctly identified. It is also called as true positive or recognition or recall rate [16].

$$\boldsymbol{Sensitivity} = \frac{(\boldsymbol{TP})}{(\boldsymbol{TP}+\boldsymbol{FN})} \quad (2)$$

- Specificity: Specificity is the ratio of patients without the disease that are correctly identified [16].

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

- Precision: precision means if an algorithm predicts yes and how often is it correct [17].

$$\boldsymbol{Precision} = \frac{(\boldsymbol{TP})}{(\boldsymbol{TP}+\boldsymbol{FP})} \quad (4)$$

- AUROC curve is used to check the measure of separability between the true positive rate and false positive rate.

In this research, the heart disease dataset is mined through Decision Tree, Random Forest, and Support Vector Machine techniques using 10-fold cross-validation to get unbiased results. Various medical domain performance metrics sensitivity, specificity, accuracy, precision, AUROC score, and misclassification rates, and the model measures like computational complexity and comprehensibility are calculated to obtain the optimal and accurate results. The experimental results obtained by different heart disease risk evaluation models are as follows:

4.1 Decision Tree

Decision tree is a type of supervised learning technique that is used for solving the classification and regression problems [18]. The main aim of using decision tree is to create a model that can predict value of a target variables or a class by learning simple decision rules inferred from training data [19] [20]. On this heart disease data set we applied the decision tree and the experimental results of the decision tree model are shown in the confusion matrix figure 2. From the decision tree model's confusion matrix the sensitivity, specificity, accuracy, precision and error rates are derived that are described as follows:

Putting the derived sensitivity values of the confusion matrix figure 2 in equation (2) the sensitivity of 82% is obtained. The closer the value for this measure is to 1, the better the rules are at identifying those patients who have heart disease. Similarly, putting the derived specificity values of confusion matrix figure 2 in equation(3) the specificity of 0.8092% is obtained which means the decision tree model can recognize the healthy cases with an accuracy of 80%. The overall accuracy of the decision tree model is obtained by using the equation (1) in figure 2 which is equivalent to 0.8185% which represents that the decision tree heart disease model's overall performance (in diagnosing both the diseased and non-diseased heart disease cases) , the higher the accuracy percentage, the more accurate the model is.

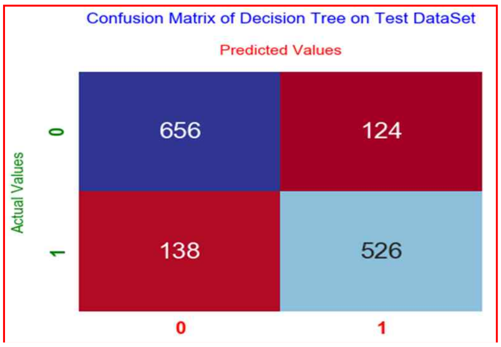


Fig 2 Decision Tree Model Confusion Matrix

Similarly, putting the values of the confusion matrix figure 2 in equation (4), a precision of 0.8410% is obtained. If a high precision rate of the decision tree model is obtained, then it means that the model will obtain a low false-positive rate.

4.2 Support Vector Machine

SVM is a discriminative classifier that takes labelled training data and produces optimal hyperplanes as output which categorizes the unknown examples [21]. In this research, the Support Vector Machine is used to develop a risk model that can predict heart disease at its early stages. The performance results of the Support Vector Machine model on the heart disease dataset are shown in confusion matrix figure 3, and the sensitivity, specificity, accuracy and precision are derived which are described as follows: Using equation (2) in confusion matrix figure 3 the sensitivity of the Support Vector Machine (SVM) model is calculated as 0.825%. Similarly, by using equation (3) in figure 3 the specificity of 0.8152% is obtained. The overall accuracy of 0.8213%, is obtained by putting the values of the figure 3 in equation (1), this means that the SVM heart disease model's overall performance (in diagnosing both the diseased and non-diseased heart disease cases) is 82%. Similarly, the precision of 0.8473% is obtained from figure 3 using equation (4), which means that the SVM model obtained the low false-positive rate.

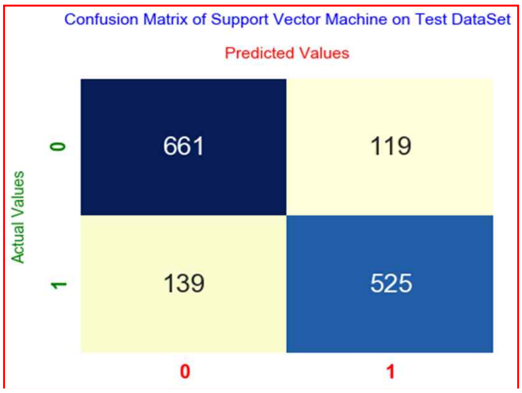


Fig. 3 SVM Confusion Matrix on Test Dataset

4.3 Random Forests

Random forest algorithm creates the forest with a number of decision trees from randomly selected training set. In random forests classification process each decision tree votes and the aggregated votes decide the final class of the test object and in regression process mean prediction or regression of the individual trees is calculated [17]. The predictive results of the Random Forest model on heart disease dataset are shown in the confusion matrix figure 4. The sensitivity, specificity, accuracy, and precision are calculated as follows:

The random forest model can recognize the positive heart disease cases with a sensitivity of 85% by putting values of figure 4 in equation (2). Similarly, the number of patients that were diagnosed healthy is equivalent to 0.8338% that is obtained by using equation (3) in figure 4. The total accuracy of 0.8462% is achieved by using the equation (1) in figure 4. Similarly, the precision of 0.8589% is obtained using equation (4) in figure 4.

We simulate the accomplished experimental results of the developed decision tree, SVM and random forest heart disease models with the prevailing research; the results obtained are to the best of our knowledge greater than the published results in the literature. Hence the proposed risk evaluation model is used for the initial prediction of the heart disease patients.

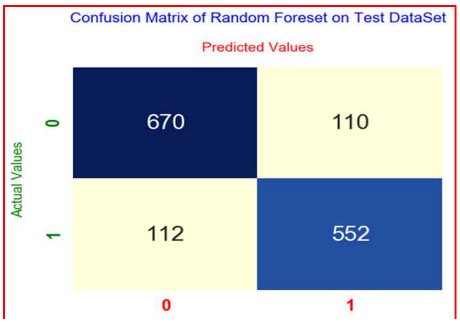


Fig. 4 Random Forest Model Confusion Matrix on Test Dataset

5. Performance Comparison of The Developed Heart Disease Models

This section presents performance and comparison of the Decision Tree, Support Vector Machine, and Random Forest risk prediction models through different measures as described in the following table 3. However we have not compared the proposed heart disease risk models based on speed, robustness, scalability and interpretability aspects.

Table 3: Performance Measures of Developed Heart Disease Models

	Performance Measures			
Models	Sensitivity	Specificity	Accuracy	Precision
Decision Tree	82%	80%	81%	84%
Support Vector Machine	82%	81%	82%	84%
Random Forest	85%	83%	84%	85%

Experimental results demonstrate that the random forest model performs most excellent in comparison to other risk models. The results of the developed heart disease risk evaluation model are checked with the prevailing risk tools which demonstrate that the results are exceptionally encouraging with outstanding predictive accuracy. After the basic assessment of experimental results, it is imperative to cautiously check and assess the data to mine important insights, build best models, and establish most favourable attribute settings. The results show that the random forest model outperforms other risk evaluation models with an optimal accuracy of 85%, the specificity of 83%, and the sensitivity of 85% and precision of 85%. The accuracy obtained by the random forest is highest for predicting heart disease and as per our knowledge is not achieved by prevailing heart disease risk models.

6. Conclusion and future work

The Medical field is immersed with data and to mine useful insights from this raw data the machine learning techniques are used. Stimulated by the rising heart disease death rates and its overall burden on the world, we develop a risk model that would help in an initial prediction of heart disease victims. Even though the new prediction technologies are now used, however these technologies are expensive which hold back their application at public level screening testing. In this research work we develop the heart disease risk evaluation model using the Jupyter notebook web application. The heart disease dataset is mined using Random Forest, Decision Tree, and Support Vector Machine techniques to check out if a person bearing different personal attributes and features, would have heart disease risk or not.

We compute the sensitivity, specificity, precision and accuracy for each model using 10-fold cross validation to check how accurately our heart disease risk assessment model performs. We simulate the accomplished outcomes against the prevailing research; the outcomes obtained are the best as per our source of knowledge and are greater than published values in the literature. In future we will improve the proposed research work using large heart disease datasets with varying number of risk features and will develop a one-size-fits-all heart disease risk model that could successfully prescribe a treatment plan for the heart disease.

References

[1] World Health Organization (2010). Global status report on noncommunicable diseases 2010. https://www.who.int/nmh/publications/ncd_report_full_en.pdf

[2] World Health Organization (2011a). The Top Ten Causes Of Death. Accessed 18 August 2017, from http://www.who.int/mediacentre/factsheets/fs310_2008.pdf

[3] Center for Disease Control and Prevention (2014). Heart Disease and Family History. http://www.cdc.gov/genomics/resources/diseases/heart.htm

[4] World Bank Disease Control Priorities Project (2013). Health Priority Setting in the Southern Cone: Action Needed on Lifestyle Risk Factors. http://www.dcp2.org/file/80/

[5] S. Palaniappan and R. Awang (2008). Intelligent heart disease prediction system using data mining techniques. IEEE/ACS Int. Conf. Comput. Syst. Appl., vol. 8, no. 8, pp. 343–350.

[6] Patil, S. B., and Kumaraswamy, Y. (2009). Extraction of significant patterns from heart disease warehouses for heart attack prediction. IJCSNS, 9(2), 228-235.

[7] M. Anbarasi, E. Anupriya, and N. C. S. N. Iyengar (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm," Int. J. Eng. Sci. Technol., vol. 2, no. 10, pp. 5370–5376, 2010.

[8] M. Shouman, T. Turner, and R. Stocker (2012). Using data mining techniques in heart disease diagnosis and treatment. Japan-Egypt Conf. Electron. Commun. Comput., pp. 173–177, 2012.

[9] N. Al-Milli (2013). Back-propagation Neural Network for Prediction of Heart Disease. J. Theory. Appl. Inf. Technol., vol. 56, no. 1, pp. 131–135, 2013.

[10] D. Masethe Hlaudi and A. Masethe Mosima (2014). Prediction of Heart Disease using Classification Algorithms. Proc. World Congr. Eng. Comput. Sci., vol. II, pp. 22–24.

[11] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra (2015). Computational Intelligence Technique for Early Diagnosis of Heart Disease. IEEE Int. Conf. Eng. Technol., Vol No, March, pp. 1–7, 2015.

[12] A. Ngueilbaye, L. Lei, and H. Wang (2016). Comparative Study of Data Mining Techniques on Heart Disease Prediction System: a case study for the Republic of Chad. Int. J. Sci. Res., vol. 5, no. 5, pp. 1564–1571, 2016.

[13] Liao, S.-C., and Lee, I.-N. (2002). Appropriate medical data categorization for data mining classification techniques. Informatics for Health and Social Care, 27(1), 59-67.

[14] Wonsuk Yoo, Robert Mayberry, Sejong Bae, Karan Singh, Qinghua, and James W. Lillard, Jr.(2015). A Study of Effects of MultiCollinearity in the Multivariable Analysis. Int J Appl Sci Technol.

[15] Jiawei Han, Micheline Kamber, and Jian Pei (2006). Data Mining Concepts and Techniques (Third Edition): Morgan Kaufmann Publishers.

[16] Mudasir M Kirmani and Syed Immamul Ansarullah (2016) Classification models on cardiovascular disease detection using Neural Networks, Naive Bayes and J48 Data Mining Techniques". International Journal of Advanced Research in Computer Science. Volume 7, No. 5, September-October 2016.

[17] Ian H. Witten, Eibe Frank, and Mark A. Hall (2011). Data Mining Practical Machine Learning Tools and Techniques (Third Edition): Morgan Kaufmann Publishers.

[18] National Center for Chronic Disease Prevention and Health Promotion (2013). Know the facts about heart disease. http://www.cdc.gov/heartdisease/docs/consumered_heartdisease.pdf

[19] Heart and Circulatory Disease Statistics (2019). British Heart Foundation. https://www.bhf.org.uk

[20] Australian Bureau of Statistics (2013). Accessed 12 March 2016, from http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/3303.0Media%20Release12011?opendocument&tabname=Summary&prodno=3303.0&issue=2011&num=&view=

[21] American Heart Association (2013). What is Cardiovascular Disease (Heart Disease)?

[22] http://www.heart.org/HEARTORG/Caregiver/Resources/WhatisCardiovascularDis ease/What-is-Cardiovascular-Disease_UCM_301852_Article.jsp

[23] Atherosclerosis. National Heart, Lung, and Blood Institute (NHLBI). https://www.nhlbi.nih.gov/health-topics/atherosclerosis