Transition from Observation To Knowledge To Intelligence (TOKI)

Editors
Dr. Victor ODUMUYIWA, Dr. Olufade ONIFADE,
Prof. Amos DAVID & Prof. Charles UWADIA

Victor ODUMUYIWA
Department of Computer Sciences,
University of Lagos
Nigeria

ISBN: 978-978-976-000-8

Transition from Observation to Knowledge to Intelligence

3rd Biennial International Conference on Transition from Observation to Knowledge to Intelligence (TOKI)
15-16 August 2019
University of Lagos, Nigeria

Editors

Dr. Victor ODUMUYIWA
Dr. Olufade ONIFADE
Prof. Amos DAVID
Prof. Charles UWADIA

Comparative Analysis of Hybridized Data Mining Models for Heart Diseases

AYOADE Akintayo M.
Department of Computer Science,
Lead City University, Ibadan. Nigeria.

OYEKUNLE Victoria B.
Department of Computer Science,
Lead City University, Ibadan. Nigeria.

AFE Oluwaseyi F.
Department of Computer Science,
Lead City University, Ibadan. Nigeria.

OKESOLA Kikelomo I.
Department of Computer Science,
Lead City University, Ibadan. Nigeria.

Abstract. One of the main sources of global increase in mortality rate could be linked to heart related diseases according to recent studies. The diagnosis process involved in the analysis of heart diseases data is cumbersome; hence the need for a data mining tool to ease the process. This paper, therefore, aims to compare the accuracy of different classification, ensemble and hybrids techniques in predicting heart disease. The dataset used is the heart statlog from UCI Machine Learning Repository dataset for heart diseases, containing 207 instances and 13 attributes/features. 10-Fold cross-validation method was used to minimize experimental result variance and to increase the amount of data made available to the classification techniques, which would otherwise have been limited given the number of instances. The different classifiers used are Naïve Bayes (NB), Radial Basis Function (RBF) Network, Decision Tree, Sequential Minimal Optimization (SMO) and K-Nearest Neighbor (K-NN), while some ensemble techniques used are AdaBoost, Bagging and Random Forest (RF). Moreover, we developed three novel hybridized techniques of SMO & RBF Network; SMO, RBF Network and NB; and SMO, RBF Network, NB & DT referred to as hybrid 1, hybrid 2, and hybrid 3 respectively. The results of the experiments show that our hybrid 3 technique outperforms the other aforementioned techniques with accuracy of 85.56%.

Keywords: data mining, classification techniques, heart disease, ensemble

1. Introduction

One of the main sources of global increase in mortality rate could be linked to heart related diseases, which cannot be clearly categorized based on age, gender, family history, life style or race (Pouriyeh, S., et al, 2017). According to recent studies, cardiovascular diseases account for more death around the world. In 2016, according to World Health Organization, people in excess of 17.7million died of heart disease which represents 31% of all deaths registered worldwide (WHO, 2016). Cardiovascular diseases (CVDs) refer to a collection of heart and blood vessels disorders that include rheumatic heart disease, coronary heart disease, cerebrovascular disease, deep vein thrombosis, pulmonary embolism, peripheral arterial disease and congenital heart disease, (WHO, 2016). On the other hand, heart disease is a medical term for a large number of medical conditions that relates to and affects the heart and all its parts (Sharma et al., 2017). Medical diagnosis is an exceptionally essential. However, it is a complex activity which must be performed precisely and competently within a minimum processing time (Takore & Shelke, 2013). Hence, the need for a reliable and cost effective means for early detection and accurate prediction.

Typically, the treatment of any disease is carried out by relying on the knowledge, experience, and instinct of the expert, while the valuable information gained from relevant databases which could aid the diagnosis process in terms of accuracy, time and cost are being ignored. In developing countries, the dearth of medical experts, wrong diagnosis, and cumbersome diagnosis process call for a tool to analyze data of heart diseases (Ayoade et al., 2018; Chadha & Mayank, 2016). These data include large volume of medical details and disease diagnosis of several patients over a period of time and their history. Medical datasets comprise a whole lots of concealed evidence that could be vital in decision making. Consequently, there is urgent need to apply data mining techniques in the medical domain that is capable of discovering hidden patterns in the Heart Disease data (Chadha & Mayank, 2016). Therefore, data mining is the drawing out of implied and hypothetically valuable information from these massive and dynamic data. The

technique is important to anticipate patient's future behavior or predict future health status of patients based on the given history (Takore & Shelke, 2013).

Many researchers have worked on identification and prediction of heart related diseases with varying techniques resulting in different level of accuracy and recommendation of best techniques. Nevertheless, supporting healthcare specialists in the diagnosis of cardiovascular diseases still demands for greater accuracy. As a result, improvement of accuracy and reduction in time complexity in detecting heart disease is the focus of research for to the data scientists around the world.

This work aims at carrying out a comparative analysis of some of the algorithms in each category of the techniques to know their level of accuracy and time complexity. Existing categories considered are Decision tree, Bayes, SVM.

3 Related Works

Data mining techniques are used generally for diseases analysis, including cardiovascular conditions. Patil and Kinariwala (2017) proposed an improved random forest classification algorithm in which number of base classifiers, a tuning parameter, will be automatically determined that will result to an ensemble with maximum accuracy and minimum correlation, thereby enhancing performance of the traditional random forest. The algorithm performance is not based on the nature of data, selection of the ensemble member is dynamic, and classification outcome is based on the collective performance decided by the combine strategy. In addition, missing values and noisy data can be well taken care of by random forest classifier, while the classifier does not need human intervention in determining the tuning parameter. The fitting procedure on the accuracy, correlation and fitted curves is done in iteration until the differences in the curves meets a specific criterion. Authors concluded by strengthening the suitability of random forest for diagnosis and prediction with recommendation since it is undependable on the peculiarities of the training set, and its reduction of bias by multiple classifiers.

In (Palechor et al., 2017), the algorithms used in the proposed method are decision trees, Bayesian networks, support vector machines, and k-nearest neighbors. For a better result, the authors used clustering method for data segmentation according to their diagnosis. These techniques were compared and SVM algorithm gave the best result. The results showed that proposed method is efficient and accurate for the diagnosis of CVDs.

Chadha and Mayank (2016) had a comparative analysis of data mining techniques for predicting heart disease. Techniques considered are: neural network, Naïve Bayes and decision tree, in varying combinations and specifically, with artificial neural network. It reveals that same classifier can perform differently on dissimilar data mining methods, number of attributes, and type of algorithm used. It concluded with the view that artificial neural network perform better than Naïve Bayes and decision tree based on appropriate attributes. Some of the strength of these classifiers are: suitability of decision tree for classification problems, Naïve Bayes ease of implementation, the accuracy of ANN and the learning vector quantization (LVQ) algorithm in ANN are easy to interpret. While the limitations are, DT is very sensitive to small perturbation in the dataset, and the issue of the black box of the ANN.

In the work of (Aro et al., 2017), data mining techniques were grouped into four different categories, namely, classification, prediction, association, and clustering methods. While heart disease was grouped into three classes: coronary heart disease, Angina, and Heart Attack. Some of the review techniques focused on prediction NN, fuzzy rules, Naïve Bayes, and evolutionary computing. A few were targeted towards classification and implemented with techniques like K-nearest neighbor, genetic algorithm, and K-means clustering. The need to identify performance of each classification technique, based on accuracy and evaluation of performance using more than one dataset and data mining software tool, to generate a viable comparative evaluation was established as well as the need to include method to handle reduction of memory usage.

Pouriyeh et al., (2017) investigated the accuracy of seven classifiers which includes: decision trees, Naïve Bayes, K-nearest neighbor, Multilayer Perception (MLP), Support Vector Machine (SVM), single conjunctive rule learner and radial basis function. These combine to form an ensemble classifier. The ensemble methods of bagging, boosting, and stacking were used for these classifiers, and the evaluation was based on recall, precision, F-measure and ROC. SVM produced the highest accuracy of 84.15% while SCRL has the lowest accuracy 69.96% with 10-fold cross validation on individual evaluation. With bagging method, DT has the worst accuracy of 78.54%. Evaluating with boosting method, SVM still has the highest accuracy of 84.81% and SCRL with the worst accuracy of 81.18%. Stacking approach using MLP and SVM generated highest accuracy of 84.15%. While the combination of K-NN, NB, RBF, MLP, and SVM produced the worst accuracy of 78.54%. Outcome of the work reviles that with boosting method, SVM outperformed the other classifiers, and accuracy of weak classifiers involved was improved.

Sharma et al., (2017), proposed a method in which they included two additional attributes (to the existing ones) which are smoking and heart diseases history so as to obtain more suitable results. They demonstrated three forms of attributes in their dataset. These are input, key and prediction attributes. The input attributes are age, gender, blood pressure, pulse rate, and cholesterol, where age (continuous and dynamic) and gender (static and constant) are the non-modifiable attributes. The other parameters have a continuous and random value. Smoking and history of heart disease (which are also the modifiable attributes) use constant values to predict the risk rate of heart disease. The key attribute is the patient's ID because it is unique for each patient. The prediction attribute yielded the likelihood (risk level) of having the disease. The risk level is classified into three categories namely: low, high, and normal risk. This was indicated as <50%, >50% and 0 respectively. The authors compared the results of the data mining techniques used (i.e. Naïve Bayes, Decision Tree, Neural Networks) for both 13 attributes and that of 15 attributes. In the results, decision tree

and neural networks performed better with 15 attributes than with 13 attributes.

In (Chaurasia & Pal, 2013), the authors performed experiments to discover the finest classifier out of the three classifiers selected for their study. The approach is meant to predict patients with heart diseases using methods of data mining. The classifiers selected for the diagnosis were ID3, CART and DT. As compared with the other two classifiers, CART was observed to outperform others using patient's data with an accuracy of 83.49% and with the least error of 0.3 on the average when compared with others. However, in terms of time required model building, CART built in 0.23 seconds, ID3 in 0.02 seconds while DT in 0.03 seconds. The results obtained imply that of all the machine learning algorithms tried, CART classifier will likely improve the prediction accuracy considerably. The results of their experiment also demonstrates that the vital attributes for heart diseases are Chest pain *(cp)*, The slope of the peak exercise segment *(slope)*, Exercise induced angina *(Exang)*, and Resting electrocardiographic *(Restecg)*. These observed important attributes were discovered using Chi-square test, Gain Ratio test and Info Gain test for the assessment of input variables.

4 Methodology

4.2 Dataset

For this study, the dataset selected is "Statlog (Heart) Data Set" and it is located in the Machine Learning Repository UCI. This dataset has 13 attributes and 270 instances.

4.3 Data Preprocessing

Real-world data is most of the time imperfect, unreliable, and deficient in certain tendencies, and usually error-prone. Data preprocessing is an established method of overcoming such challenges and involves transforming raw data into more process-able format for further processing. The different steps involved are data cleaning, transformation, integration, reduction and discretization.

Although, for this paper, the data preprocessing stage was not carried out because the dataset had already been pre-processed from

which 13 attributes were extracted out of a larger set of 75. However, the overall framework adopted is shown in Figure 1 below.

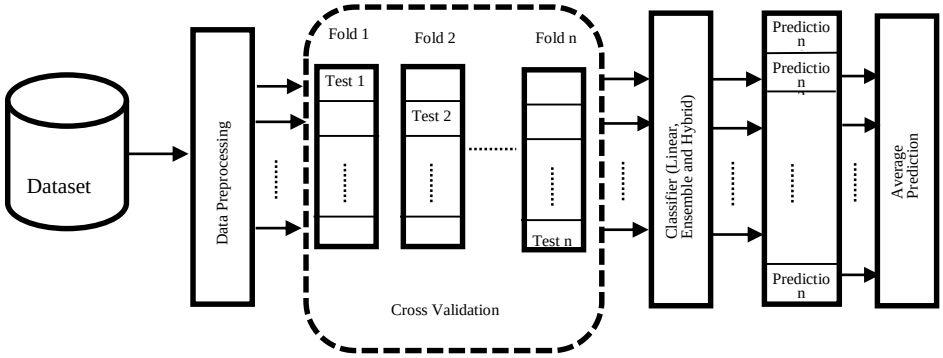


Figure 1: The Prediction System Framework

4.4 Algorithm Selection

It is a meta-algorithmic method to pick an algorithm from a collection on a case-by-case basis. It is inspired by the observation that on many real-world problems, algorithms performances differ. That is, although an algorithm may perform well in some cases, it may perform poorly on some others and the same could be said for other algorithms as well. The only criterion for algorithm selection application is that there must be a set of complementary algorithms. Our choice of learning algorithm is based on prediction accuracy and time complexity.

3.3.1 Artificial Neural Network (ANN)

It is a sequence of algorithms that tries to identify fundamental connections in a set of data in such a manner that imitate the operation of human brain i.e., it mimics the network of brain tissues and nerves as a base to develop algorithms which model multifarious formations and extrapolation problems. The mathematical model is given as follows:
The weights W denote the "strength" of connection between neurons, Y_i represents the output.

Each neuron located in the hidden layer receives weighted inputs plus bias from each neuron in the preceding layer, as modelled by equation (1)

$$Z_i = (\sum_{k=1}^{N_{j-1}} X_k^{j-1} W_{k,i} - b_k) \qquad (1)$$

Where:

X_k^{j-1} stands for the input from *k*-th node in the *j*-th layer,

$W_{k,j}$ represents the weight of the link between node and all the nodes in the previous layers, and

b_i denotes the bias to the node, N_{j-1}, is the number of nodes in the layer $j-1$.

The sigmoidal function is the most commonly used activation function, defined as:

$$f(Z_i) = {}^{1}/_{(1+e^{-Z_i})} \qquad (2)$$

This sum is propagated into to an activation function, to yield the output of the node, modelled as:

$$Y_i = f(Z_i) \qquad (3)$$

The activation function serves to model nonlinear behaviours.

In this paper, we used the variation of the algorithm referred to as Radial Basic Function (RBF) Network. RBF network can be used for classification, regression and function estimation.

3.3.2 Naïve Bayes Classifier

Naïve Bayes (NB) classifier is a modest probabilistic classifier built on Bayes' theorem with strong (naive) independence assumptions. NB classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. For instance, a fruit could be regarded as an apple if it is red and round with about 4 inches in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. The naive assumption of class conditional independence is often made to reduce the computational cost. Bayes theorem is represented mathematically as follows:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad (1)$$

where P(A|B) is a conditional probability, the likelihood of an event A occurring given that B is true.

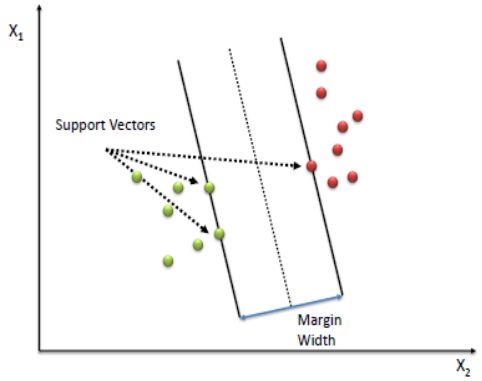
P(B|A) is also a conditional probability: the likelihood of event B occurring given that A is true

P(A) and P(B) are the probabilities of observing A and B independently of each other.

3.3.3 Support Vector Machine (SVM)

Support Vector Machines (SVMs), is a discriminative classifier that have been extensively used to model both classification challenges and nonlinear regressions, though it is mostly used in classification problems

SVM achieves classification by discovering the hyperplane that makes the most of the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors.



For a 2 dimensional space, the hyperplane form is;

$\beta_0 + \beta_1 * x_1 + \beta_2 * x_2 = 0$

The dots above the line;

$\beta_0 + \beta_1 * x_1 + \beta_2 * x_2 > 0$

The dots below the line;

$\beta_0 + \beta_1 * x_1 + \beta_2 * x_2 < 0$

where β is the vector normal to the surface

Algorithm

1) *Define an optimal hyperplane: maximize margin*
2) *Extend the above definition for non-linearly separable problems: have a penalty term for misclassifications.*
3) *Map data to high dimensional space where it is easier to classify with linear decision surfaces: reformulate problem so that data is mapped implicitly to this space.*

Conversely, it requires higher training time for large datasets; therefore, it does not perform well for large datasets and datasets with much noise. This drawback can be taken care of by some of the offshoot of SVM. An example is Sequential Minimal Optimization (SMO).

3.3.4 Sequential Minimal Optimization (SMO)

Sequential Minimal Optimization (SMO) algorithm is used for training a support vector classifier using polynomial or RBF kernels. It replaces the missing values and transforms nominal attributes into binary attributes. It is the fastest quadratic programming algorithm for training SVM particularly for linear SVM and sparse data performance.

3.3.5 Ensemble

Ensemble learning is an approach developed to increase the predictive performance of base classifiers. This is achieved by the coming together of different base classifiers in such a way that the same problem is presented to each of these classifiers and the classification of a new problem will depend on collective recommendation of the majority of the classifiers in the ensemble. For this paper, we employed common ensemble methods of bagging, randomization, boosting as well as development of three (3) novel hybrid classifiers namely hybrid1, hybrid 2 and hybrid3. Hybrid1 was formed from SMO and RBF Network algorithm, Hybrid2 from SMO, RBF Network and Naïve Bayes algorithm and Hybrid3 from SMO, RBF Network, Naïve Bayes and Decision Tree algorithms as depicted in Figure 2.

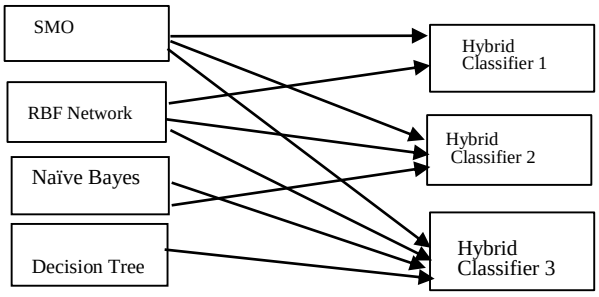


Figure 2: Hybrid Classifier Formation

The algorithm for the hybrid ensembles is as follow:

1. *Input: Training data* $D = \{x_i, y_i\}_{1=1}^{m}$
2. *Output: ensemble H*
3. *Step 1: learn base classifier*
4. ***for*** *t=1 to T* ***do***
5. *learn* h_t *based on D*
6. ***end for***
7. *Step 2: Build new data set of predictions*
8. ***for*** *i=1 to m* ***do***
9. $D_i = \{x_i', y_i\}, where\ x_i' = \{h_1(x_i), \dots, h_T(x_i)\}$
10. ***end for***
11. *Step 3: learn a meta-classifier*
12. *Learn H based on* D_i
13. *Return H*

5 Result and Discussion

In this section, we present and discuss the results of the implementation of different algorithms described in section 3 above. Ten (10) fold cross validation technique is used to minimize experimental result variance and to increase the amount of data made available to the classification models, which would otherwise have been limited given the number of instances in the dataset.

Table 1: Accuracy versus Time for single classifiers

Classifier	Accuracy (%)	Time (s)
Naïve Bayes	83.7037	0.01
RBF Network	84.0741	0.24
SMO	84.0741	0.09
KNN	84.0741	0
Decision Tree (DT)	83.3333	0.11

Table 2: Accuracy versus Time for ensemble classifiers

Classifier	Accuracy (%)	Time (s)
AdaBoost	80	0.13
Bagging	79.2593	0.11
RandomForest	81.4815	0.27
Hybrid 1	84.0741	0.04
Hybrid 2	85.1852	0.02
Hybrid 3	85.5556	0.03

RBF Network, SMO and KNN algorithms, all being single classifiers, are comparable in terms of accuracy with each having

percentage accuracy of 84.0741 implying that any of them could be adopted in practice, however, the KNN algorithm will be preferred in time-critical application because it takes the least time to build an optimal model as shown in Table 1. For the ensemble counterparts, Hybrid3, being an ensemble of SMO, RBF Network, Decision Tree and Naïve Bayes returned the highest accuracy of 85.5556 as depicted in Table 2. In practice however, hybrid2 and hybrid 3 models are comparable as both slightly outperform each other in either of the performance evaluation metrics of accuracy and time to build model but not in both. In applications where accuracy is preferred over time factor, hyrid 3 will be the model of choice but in time-critical application, hybrid 2 will be chosen over hybrid 3. Figures 3a, 3b, 4a and 4b show comparison among the classifiers and the time taken to build their respective models.

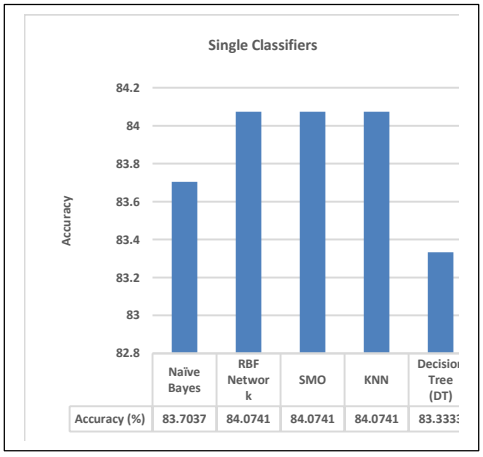


Figure 3a: Accuracy of single classifiers

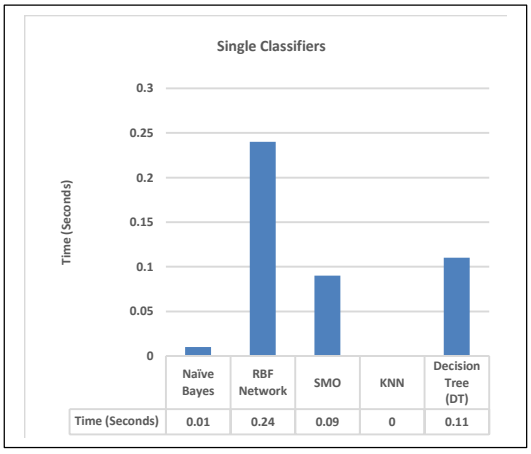


Figure 3b: Time to build single model

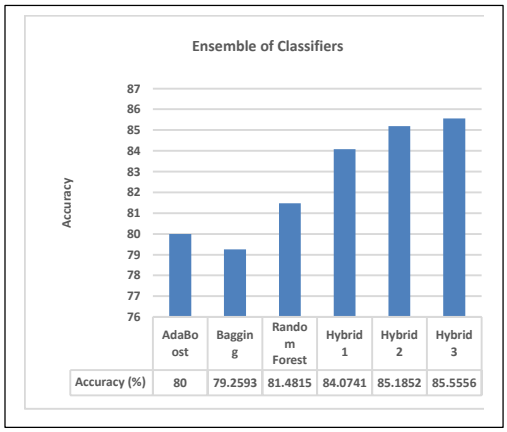


Figure 4a: Ensemble classifiers' accuracy

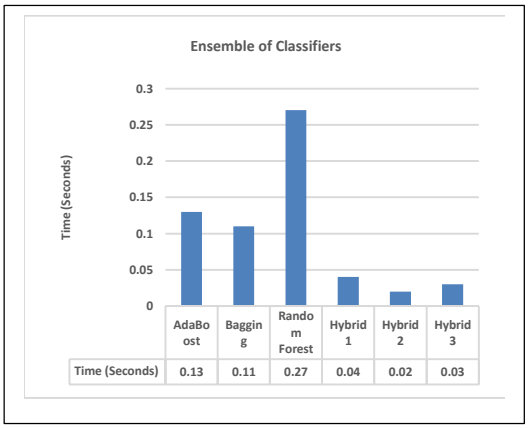


Figure 4b: Time to build ensemble model

6 Conclusion

In this paper, both single and ensemble classifiers were investigated. The different single classifiers used are Naïve Bayes (NB), Decision Tree, Sequential Minimal Optimization (SMO), Radial Basis Function (RBF) Network, and K-Nearest Neighbor (K-NN), while the ensemble techniques used are AdaBoost, Bagging and Random Forest (RF). We also developed three novel hybridized techniques referred to as hybrid1, hybrid2, and hybrid3. Our hybrid ensemble classifiers outperform all the individual classifier components from which they are developed in term of both the evaluation measures used except for KNN algorithm which is better in term of time. However, when KNN is used in the ensemble, the percentage accuracy of the hybridized algorithm dropped drastically. We intend to investigate this observation in our future work.

List of References

Ayoade, A.M., et al., A Framework for Patient Diagnosis System in Sub-Sahara African Primary Health Care Centers. decision-making, 2018. 7(02).

Aro, T., et al., A Review On Data Mining Techniques For Heart Disease Prediction. Annals. Computer Science Series, 2017. 15(1).

Chadha, R. and S. Mayank, Prediction of heart disease using data mining techniques. CSI transactions on ICT, 2016. 4(2-4): p. 193-198.

Chaurasia, V. and S. Pal, Early prediction of heart diseases using data mining techniques. 2013.

Palechor, F.M., et al., Cardiovascular Disease Analysis Using Supervised and Unsupervised Data Mining Techniques. JSW, 2017. 12(2): p. 81-90.

Pouriyeh, S., et al. A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. in Computers and Communications (ISCC), 2017 IEEE Symposium on. 2017. IEEE.

Patil, P.R. and S. Kinariwala, Automated Diagnosis of Heart Disease using Data Mining Techniques. International Journal of Advance Research, Ideas and Innovations in Technology, 2017. 3(2).

Sharma, M., F. Khan, and V. Ravichandran, Comparing Data Mining Techniques Used For Heart Disease Prediction. 2017.

Takore, M. and R.R. Shelke, Data Mining Techniques to Find Out Heart Deisease: An Overview. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 2013. 4(III): p. 5.

World-Health-Organization, Cardiovascular diseases (CVDs). 2016.