American Journal of Artificial Intelligence
2020; 4(1): 20-29
http://www.sciencepublishinggroup.com/j/ajai
doi: 10.11648/j.ajai.20200401.12
ISSN: 2639-9717 (Print); ISSN: 2639-9733 (Online)

SciencePG
Science Publishing Group

A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques

Lamido Yahaya[1], Nathaniel David Oye[2], Etemi Joshua Garba[2]

[1]Department of Computer Science, Faculty of Science, Gombe State University, Gombe, Nigeria

[2]Department of Computer Science, School of Physical Sciences, Modibbo Adama University of Technology, Yola, Nigeria

Email address:
yahaya.lmd@gmail.com (L. Yahaya)

To cite this article:
Lamido Yahaya, Nathaniel David Oye, Etemi Joshua Garba. A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques. *American Journal of Artificial Intelligence*. Vol. 4, No. 1, 2020, pp. 20-29. doi: 10.11648/j.ajai.20200401.12

Received: March 12, 2020; **Accepted**: April 2, 2020; **Published**: April 23, 2020

Abstract: Heart disease is one of the major causes of life complicacies and subsequently leading to death. The heart disease diagnosis and treatment are very complex, especially in the developing countries, due to the rare availability of efficient diagnostic tools and shortage of medical professionals and other resources which affect proper prediction and treatment of patients. Inadequate preventive measures, lack of experienced or unskilled medical professionals in the field are the leading contributing factors. Although, large proportion of heart diseases is preventable but they continue to rise mainly because preventive measures are inadequate. In today's digital world, several clinical decision support systems on heart disease prediction have been developed by different scholars to simplify and ensure efficient diagnosis. This paper investigates the state of the art of various clinical decision support systems for heart disease prediction, proposed by various researchers using data mining and machine learning techniques. Classification algorithms such as the Naïve Bayes (NB), Decision Tree (DT), and Artificial Neural Network (ANN) have been widely employed to predict heart diseases, where various accuracies were obtained. Hence, only a marginal success is achieved in the creation of such predictive models for heart disease patients therefore, there is need for more complex models that incorporate multiple geographically diverse data sources to increase the accuracy of predicting the early onset of the disease.

Keywords: Data Mining, Machine Learning, Heart Disease, Classification, Prediction

1. Introduction

The heart is one of the most essential organs in humans. It is a kind of muscular organ which pumps blood into the body and is the central part of the body's cardiovascular system [25]. The cardiovascular system is composed of all blood vessels such as arteries, veins, and capillaries that form a complex network of blood circulation all over the body [15]. Any obstruction or abnormality in normal blood circulation or flow from the heart would result in several and severe complications of heart diseases. These are commonly called cardiovascular diseases (CVDs) and are among the deadliest diseases in the world. CVDs include diseases of the heart, vascular diseases of the brain, and diseases of blood vessels [48]. The World Health Organization (WHO) Report Global Atlas on Cardiovascular Disease Prevention and Control states that CVDs are the leading causes of deaths and disability in the world [22]. Although CVDs can be prevented through life style changes and other related measures but from all indications they are still on rise on daily basis, as stated in various reports by the WHO.

However, various reports by the WHO have indicated the rise of CVDs globally, which is very alarming. More people die from CVDs worldwide than from any other cause-an estimated 17.5million people in 2012 [49]. According to another WHO report, 17.9 million people die each year from CVDs, an estimated 31% of all deaths worldwide. Of these, 85% are due to heart attack and stroke [50]. The various reports by the WHO have indicated that deaths due to heart diseases have been on the increase, which are mainly attributed to inadequate preventive measures despite of the increasing risk factors.

Medical proofs have shown that there are certain risk factors that increase a person's chances of having a cardiovascular or more specifically a heart disease. Some of these factors as enumerated by [9] include family history of cardiovascular diseases, high level of LDL (bad) cholesterol, low level of HDL (good) cholesterol, hypertension, high fat diet, lack of regular exercise, and obesity. Other risk factors include smoking, diabetes, age and gender. With these factors and more, physicians generally make diagnoses by evaluating a patient's current health status and previous diagnoses made on other patients with the same status. Cardiovascular diseases are of many types, some of which were listed by [36]:

1. Coronary Heart Diseases: Damage or disease in the major blood vessels.
2. Cardiomyopathy: An acquired or hereditary disease of the heart muscles.
3. Ischemic Heart Disease: Heart problems caused by narrowed heart arteries, which causes less blood and oxygen to reach the heart muscles.
4. Heart Failure: A chronic condition in which the heart does not pump blood as well as required.
5. Hypertensive Heart Disease: Heart problems caused by high blood pressure.
6. Inflammatory Heart Disease: Heart problems or conditions caused by viral or bacterial infections.
7. Valvular Heart Disease: Damage or defect in one of the heart valves.

The increasing rate of heart diseases has become a global concern. Therefore, the healthcare industry needs to shape and intensify the way these diseases are handled in order to minimize the impact in the society. Huge data is available in the healthcare industry [42], more importantly the heart disease data, which needs to be efficiently analyzed for effective decision making. Based on data, statistics, clinical records and hospital management, it is claimed that in every 3 years, medical data doubles up and making health industry a multi-billion dollar domain [17]. Machine learning and data mining techniques play a very vital role in the medical data analysis and knowledge extraction. The increasing morbidity and mortality due heart diseases worldwide has attracted the attention of researchers to conduct many studies in their effort to minimize the rates. Data mining and machine learning techniques have been widely used in the implementation of clinical decision support systems for heart disease prediction. The data mining applications are used for better health policy-making and prevention of hospital errors, early detection, prevention of diseases and preventable hospital deaths [27].

2. Literature Review

Reference [3] presented a heart disease prediction framework using some supervised machine learning algorithms in R programming language. The algorithms used include Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Naïve Bayes (NB). The Cleveland datasets from the University of California, Irvine (UCI) machine learning repository consisting 303 instances and 76 features were used. The data was preprocessed due to missing values and the sample became 302 instances and only 14 heart disease features in size. The data was split into 70% and 30% for models training and testing respectively. It was a comparative analysis of the selected techniques in which the experimental results showed that the NB classifier performed the heart disease prediction better than the SVM and KNN, with an accuracy of 86.6%.

Reference [44] proposed a diagnostic system for predicting heart disease using Multi-Layer Perceptron Neural network (MLP) with back propagation as the training algorithm. The performance of the developed system was evaluated based on sensitivity, specificity, precision and accuracy. The Cleveland data of the UCI machine learning repository containing 303 instances and 76 features was employed for model training and testing. Data preprocessing was performed to remove 6 instances which contain missing values. Of the 76 features, only 14 were used as the most relevant to heart disease. Based on the experiments performed, the MLP-NN proposed model gave high accuracy of 93.39% for 5 neurons in hidden layer with running time of 3.86 seconds in the heart disease prediction.

Reference [28] proposed a logistic regression (LR) based approach of machine learning for heart disease prediction. Other algorithms such as NB, SVM, DT, and KNN were also explored using SK-Learn library for performance comparisons with the LR algorithm. According to them, the experimental results showed that the LR algorithm performed better at 86.89% accuracy. While other algorithms performed at 77.85% for KNN, 86% for NB, 78.69% for DT and 82% for SVM. Datasets used for model training and testing processes were not specified.

Reference [5] performed a comparative study between ANN and SVM classification algorithms based on Positive Predictive value (PPV) of cardiovascular diseases. Their data was obtained from three selected hospitals affiliated to AJA University of Medical Sciences, Iran. The sample is composed of 1324 instances and 25 features. The sample is a medical records of patients with coronary artery diseases who were hospitalized in the three mentioned hospitals between March 2016 and March 2017. The data was collected based on the variables used in the guideline of the Cleveland heart disease data policy in UCI machine learning repository. The collected data were controlled using different methods, such data preparation, integration, cleaning, normalization and reduction. The data was fed SPSS (v23.0) and Microsoft Excel 2013, then R 3.3.2 was used for statistical computing. The sample was divided into 70% and 30% for algorithm training and testing respectively. Results of their experiments showed that SVM algorithm presented higher accuracy and better performance than the ANN model, and was characterized by higher power and sensitivity.

Reference [26] studied the case of predicting the risk of cardiovascular diseases (CVDs) by comparing Auto Machine Learning techniques against a graduate student using several important metrics, including the total amount of time required for building machine learning models and the final classification accuracies on unseen test datasets. They proposed Auto-SKlearn model, which was motivated by

Scikit-Learn, a popular generic machine learning toolbox. Their model utilizes a large number of machine learning classifiers and preprocessing steps in the Scikit-learn toolbox. The classifiers used include LR, SVM, RF, Boosting, NN, and KNN. Given the training data, Auto-SKlearn first selects an appropriate set of data preprocessing steps, such as imputation of missing values. It then passes the processed data to feature processing block which further normalizes the data or reduces their dimensions using standard techniques, principal component analysis. Finally, the datasets are passed to the estimator block which selects and trains machine learning algorithms to predict desirable outputs from input data samples. Training, testing and evaluation were performed on two different cardiovascular disease datasets. The Cleveland heart disease datasets from the UCI machine learning repository, comprising 303 instances with 76 features from which only 13 were used. The other datasets were CVD data (source not specified) comprising 70,000 instances with 11 features. In the experiments, each dataset category was split into three: training set, validation set and testing set. The 303 instances of the UCI data was divided into 100 for testing, and 203 for training and validation. K-fold cross validation method was adopted. The 70,000 records of the CVD data was divided into 14,000 for testing and 56,000 for training and cross validation. According to the results of comparative analysis performed on the two different datasets, the Auto-ML model took only 30 minutes to build a competitive classifier for each dataset, compared to long periods of time (432 hours for UCI data and 360 hours for CVD datasets) taken by the graduate student to develop similar classifiers. The Auto-ML model is too slow and inefficient, taking 30 minutes to build a classifier.

Reference [21] presented a machine learning-based technique for detection of heart disease using sampling techniques to handle unbalanced datasets. The sampling techniques used include Random Over-Sampling, Synthetic Minority Over-Sampling (SMOTE) and Adaptive Synthetic Sampling Approach (ADASYN). Framingham datasets from the Kaggle website, which contains 4239 instances with 15 features were used for the algorithm training and testing. Based on the features, the aim was to predict whether a patient had a 10-year risk of future coronary heart disease. The machine learning techniques used include LR, KNN, AdaBoost, DT, NB, and RF. The performances of these classification algorithms were measured and evaluation based on precision, recall, and accuracy. Each of these parameters varies according to the sampling technique used. From their experimental results, SVM classifier with Random Over-Sampling technique appeared the best in the heart disease prediction with an accuracy of 99%. However, RF performed better with SMOTE technique at 91.3% accuracy while DT classifier and RF again performed better with ADASYN technique at 90.3% accuracy. Therefore, the classification accuracy of this approach was solely based on the sampling techniques, which are not always necessary in all types of datasets.

Reference [33] implemented a machine learning-based approach for heart disease prediction using comparative analysis of DT and SVM classification algorithms in Python. Age, chest pain, blood pressure, cholesterol level were among the heart disease features considered in the unmentioned datasets. The unspecified sample was divided into 75% and 25% for model training and testing respectively, using cross validation method. Data preprocessing was carried out to remove inconsistencies and missing values using PANDAS algorithm and Mat Plot Lib was used for data visualization. Experimental results showed that DT classifier performed much better than the SVM. The DT classifier had an accuracy of 100% while that of SVM was 55%. Their conclusion was that the performance of a classifier depends on the type of heart disease datasets used, which showed that the DT classifier performance could not be generalized as the best model for heart disease prediction despite of the 100% classification accuracy.

Reference [4] proposed a heart disease prediction framework based on RF algorithm in machine learning using Python. They used the Cleveland heart disease datasets obtained from the UCI machine learning repository for the algorithm training and testing. This sample originally contains 303 instances with 76 features but after preprocessing and manual attribute selection of features, only 9 features were used. 75% of the sample was used for algorithm training while 25% was used for testing. A graphical user interface (GUI) was developed using Visual Studio Code for visualization of the experiments. The RF classifier was employed for the classification, where an accuracy of 97.56% was achieved. The heart disease diagnosis was divided into four (4) stages based on artery blockage, where an artery blockage greater than 50% indicates the presence of heart diseases. This model could not detect heart disease early, since 50% of artery blockage is still classified as normal or absence of heart disease.

Reference [43] proposed a heart disease prediction based on machine learning techniques using NB and DT algorithms in Python. The datasets used for training and testing of the model were obtained from the Kaggle website, which contain 13 heart disease features. Another dataset from the UCI machine learning repository was used for the simulation. The proposed model was implemented on the Scipy environment. Form their experiments, results showed that DT algorithm performed better than the NB in the prediction of heart diseases. Their study had a lot of shortcomings, which include unspecified datasets, unavailability of real experiments, imprecise results, and improper feature selection approach.

Reference [16] proposed a web-based application for predicting heart diseases using machine learning techniques. The algorithms used for classification were SVM, LR, and NB. Heart disease datasets from the UCI machine learning repository were proposed for algorithm training and testing, divided into 75% and 25% respectively. The proposed system would have user interface, through which heart disease patients enter their information and a database implemented using MySQL, which stores patients' medical history. Data preprocessing was carried out to remove inconsistencies and missing values. From the three classifiers selected, SVM performed better with an accuracy up to 64.4%, and was selected for the main application. The remaining two algorithms, which are NB and LR had classification accuracies of 60% and 61.45% respectively. Based on their proposed experiments, the system would give prediction if a patient had a heart disease risk greater than 60%, which is too high. This means that the system could not predict earlier heart disease risks in patients.

Reference [37] developed a model using combined descriptive and predictive techniques of data mining for predicting patients with coronary artery diseases (CAD). Datasets containing 282 instances with 58 features obtained from a clinic were used. The data was preprocessed to remove missing values and outliers. K-means algorithm was chosen as clustering method (descriptive) and for the predictive technique, various classification algorithms, which include CHAID, Quest, C5.0, C & RT-DT, and ANN were chosen. Their experimental results showed that the C & RT-DT algorithm appeared the best in predicting CAD with an error of 0.074, when the entire datasets were used. However, results obtained for the three clusters were different. In clusters 1 and 2, C & RT-DT performed better with 0.022 and 0.023 errors respectively. While in cluster 3, CHAID algorithm appeared the best performing classifier with zero error. Accuracy in prediction depends on the heart disease features and other factors, but too much features, such as the one here (58) might result in unreliable predictions.

Reference [45] proposed a heart disease prediction framework called “Hybridization” that combined several machine learning algorithms into a single model. The Cleveland datasets from the online machine learning repository of the UCI consisting of 303 instances and 14 features were used in the model training and testing processes. Data preprocessing was carried out to reduce the attributes from 14 to 12. The range of classification algorithms applied included the NB, SVM, KNN, NN, J48, RF, and GA, taking into account their accuracies, sensitivities and specificities in the heart disease prediction. They were applied on the same dataset and features one after the other. The results of the experiments showed that NB and SVM performed better in the heart disease prediction with the same accuracy of 89.2%.

Reference [1] performed a comparative study on heart disease classification and prediction using machine learning techniques. The algorithms used include NB, DT, RF, SVM, and LR in the Rapid-Miner. The common Cleveland heart disease datasets from the UCI machine learning repository consisting of 303 instances and 14 attributes were used. During learning and of the model, 10-fold cross validation technique was used. From the results of the experiments, DT algorithm appeared the highest in the heart disease prediction accuracy followed by SVM at 93.19% and 92.30% respectively.

Reference [18] performed a comparative analysis on some of the popular machine learning algorithms used for heart disease prediction. WEKA 3.6 version was used to study four classifiers including RIPPER, DT, ANN, and SVM. The usual UCI datasets for Cleveland containing 303 instances and 14 attributes were used for the model training and testing. Data preprocessing operation was performed which subsequently reduced the sample size to 296 instances. The performances of the selected algorithms were compared with other classifiers which include the KNN, NB and MLP. The experimental results showed that the selected algorithms performed better, with SVM having the performance of 90.00% accuracy.

Reference [29] performed a comparative study on some of the most popular classification models used in data mining. They include K-Nearest neighbor (KNN), Support Vector Machine (SVM) and Artificial Neural Network (ANN) using MATLAB multilayered feed forward back propagation. Cleveland heart disease data containing 303 instance with 76 features from the UCI machine learning repository was used. They performed data preprocessing to remove records with many missing values, where the data size became 270 instances with only 13 attributes. Half of the data was used for models training and the other half for testing. Their experimental results showed that SVM outperformed both the KNN and ANN based on the classification accuracy at 85% while KNN at 82% performed better than ANN at 73% approximately.

Reference [2] conducted a study to identify the most significant features in heart disease prediction. In their system framework, seven classification algorithms in the Rapid-Miner studio were used, which include the KNN, DT, NB, LR, SVM, NN, and Vote. The Cleveland data containing 303 instances with 76 features obtained from the UCI machine learning repository was used. They performed a cross validation on the data using 10 folds cross validation approach. One subset was used for training and the remaining for testing. From the results of their experiment, Vote classifier appeared the best in the heart disease prediction with an accuracy of 87.4%.

Reference [14] also carried out a comparative investigation on heart disease prediction using support vector machine, decision tree, and k-nearest neighbor algorithms. They used the VA Long Beach dataset obtained from the UCI machine learning repository, which comprises of 270 instances and 12 attributes for the algorithm training and testing purposes. The model was evaluated based on accuracy, sensitivity, and specificity using confusion matrix. Their experimental results showed that Support Vector Machine (SVM) performed better than KNN and DT in classifying the heart disease patients, with an accuracy of 92%, sensitivity of 100%, and specificity of 83%.

Reference [12] presented a heart disease prediction framework using Naïve Bayesian classifier and K-means algorithms. Their datasets were obtained from the UCI machine learning repository, but not clearly specified. They used 13 heart disease features such as age, gender, high blood pressure, cholesterol, etc. to estimate the possibility of heart disease for a person. The system allows patients to enter their information, which would be classified as either normal or heart disease stages 1, 2, or 3. Using the NB and improved K-means algorithms, the risk rate of heart disease was detected and accuracy level also provided obtained according to the number of heart disease features entered.

Reference [13] developed a machine learning based hybrid intelligent system framework for heart disease patients' diagnosis using seven of the popular classification algorithms using Python. They include KNN, ANN, DT, SVM, NB, LR and MLP. Cleveland dataset containing 303 instances with 76 features was used for model training and testing. They applied a 10 folds cross validation approach on the data. Feature selection algorithms which include Relief, Minimal-Redundancy-Maximal-Relevance (mRMR) and Least Absolute Shrinkage and Selection Operator (LASSO) were used to select the best heart disease correlated features. The data was preprocessed to remove the instances with large missing values. This reduced the data size to 297 instances with only 14 features. Applying the feature selection algorithms reduced the features to 6 only as heart disease related. They tested each of the classifiers with any of the feature selection algorithms in order to get the best performing model. Their experimental results showed that SVM with LASSO feature selection algorithm appeared the best performing combination, as compared with other feature selection algorithms and classifiers. Narrowing the heart disease features to only 6 would lead to unreliable classification accuracy, as more relevant features were excluded.

Reference [39] presented a heart disease prediction framework that uses a Convolutional Neural Network based Multimodal Disease Prediction (CNN-MDRP) algorithm which uses both structured and unstructured big data from a particular hospital. It was a comparative study with a Convolutional Neural Network based Uni-modal Disease Prediction (CNN-UDRP) algorithm which uses only structured data. They used the Naïve Bayes (NB) classifier for the classification process. In their model, automatic selection of characteristics from a large data improves the disease prediction accuracy. Their experimental results showed that the CNN-MDRP model performed well in heart disease classification with an accuracy of 94.80. No dataset was specified for the algorithm training and testing.

Reference [8] performed a comparative analysis using some data mining techniques to design a cardiovascular disease prediction model after analyzing some existing models. Data used was obtained from Transthoracic Echocardiography database, which contains 336 instances and 24 attributes. They used three of the popular machine learning models: DT-J48, Naïve Bayes (NB), and Neural Network (NN) for the analysis and classification processes. The performance measure was done based on False Negative, False Positive, True Negative, True Positive, Precision, Recall, and Accuracy. Three different experiments were conducted. Their experimental results showed that NN model performed much better in heart disease prediction with 97.91% accuracy.

Reference [32] proposed a model to predict patients with heart failure using a multi-structure dataset integrated from various sources. They extracted different important factors of heart diseases from King Saud Medical City (KSUMC) system, Riyadh, Saudi Arabia. The datasets obtained were in structured, semi-structure, and unstructured format, comprising 100 real patient records with many missing values and misidentified attributes, extracted from the KSUMC Electronic Health Record (EHR). Validation of the selected dataset was achieved by consolidating some cardiologists and data scientists. Data preprocessing operation was performed to remove missing values and misidentified attributes to enhance the parameters, which were integrated into the Hadoop Distributed File System (HDFS). Machine learning algorithms: Support Vector Machine (SVM) and Decision Tree (DT) in WEKA were used for the classification process, and Area Under the call (AUC) technique was used for the performance measure. Their main contribution was the use of structured datasets in the design of heart disease predictive model for better results.

Reference [40] proposed a hybrid model combining Genetic and Naïve Bayes algorithms in python for heart disease prediction. They used the popular UCI dataset of Cleveland comprising 303 instances and 14 heart disease features. Based on the comparative analysis performed with other data mining techniques such as Weighted Fuzzy Rules, Logistic Regression, as used in previous studies, their proposed model, GA_Fuzzy_Naive was found to be more accurate at 97.14%.

Reference [25] proposed a tentative design of a cloud-based heart disease prediction system using machine learning techniques. Two of the UCI datasets: Cleveland heart disease data consisting of 303 instances with 14 features and VA Long Beach data consisting of 270 instances with also 14 features were merged together making a bigger dataset. Five machine learning algorithms, including MLP, LR, NB, RF, and SVM in the Java-based open access platform (WEKA) were applied in the classification and prediction processes. Of the five algorithms, SVM appeared the best classifier with a classification accuracy of 97.53%.

Reference [11] proposed a framework that combined the popular Naïve Bayesian classifier and Particle Swarm Optimization (PSO) feature selection algorithm for efficient heart disease prediction. The UCI dataset of VA Long Beach consisting of 270 instances and 14 features was used for the model training and testing processes. Of the 14 features, only 7 were selected for the heart disease prediction. From the experimental results, the Naïve Bayes predictive model performance was 79.12% accurate but escalated to 87.91% when integrated with the PSO selection algorithm. It was concluded that the NB+PSO model improved the heart disease classification accuracy, which is 8.79% better than the original NB performance.

Reference [7] used three of the most popular data mining techniques: RF, NB and DT to develop a prediction system in order to analyze and predict the possibility of heart diseases. Their fundamental objective was to identify the best classification algorithm suitable for providing maximum accuracy when classification of normal and abnormal person was carried out. The UCI dataset of VA Long beach consisting of 270 instances and 13 heart disease features were used for models' training and testing processes. The dataset was split into 80% and 20% for models training and testing respectively. Their experimental results showed that RF classifier performed better than NB and DT in the heart disease prediction.

Reference [41] presented a framework based on neural network (NN) to develop an effective heart disease prediction system (EHDPS) for predicting the risk level of heart disease. The Multi-Layer Perceptron (MLP) neural network (NN) with back propagation was used as training algorithm. The UCI dataset of the Cleveland consisting of 303 instances and 15 heart disease features was used for the model training and testing processes. The data was divided into 40% and 60% for the training and testing respectively. Data preprocessing operation was carried out to remove noisy data and missing values. Their experimental results showed that the proposed model was able to predict heart diseases with 100% accuracy.

Reference [19] proposed a hybrid approach for heart disease prediction using machine learning algorithms optimized by Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) techniques. Fast Correlation based Feature selection (FCBF) method was used to remove redundant features from the datasets, for improved performance of the classifiers. The Cleveland heart disease datasets of the UCI machine learning repository were employed for algorithm training and testing. The sample size is commonly known with 303 instances and 76 features, form which only 7 features were left for heart disease prediction when the feature selection techniques were applied. The classification experiments were carried out in WEKA environment using five algorithms, which include RF, KNN, SVM, NB, and MLP using 10-fold cross validation method. The effectiveness of the classifiers was evaluated based on time to build the model, correctly classified instances, incorrectly classifies instances, and accuracy. The evaluation was performed in the first instance without optimization, then with FCBF optimization and finally with the FCBF, PSO, and ACO optimizations. Confusion matrices were created to represent the evaluation scenario. Accuracies of the classifiers were measured based on precision, recall, TP rate and FP values. Comparative analysis of the result obtained showed that the proposed optimized model with FCBF, PSO and ACO techniques achieved an accuracy of 99.65% in predicting patients with heart diseases, with the KNN classifier and 99.60% with RF classifier.

Reference [23] proposed a "vertical system integration of a sensor node and toolkit of machine learning algorithm for predicting heart diseases in patients". The system could be used for both heart disease monitoring and diagnosis. The pulse rate sensor (AMPED) and Bluetooth were integrated to monitor the Heart Rate Variability (HRV) and send the data to mobile application. The graph of the heart rate and heart disease prediction could be seen through the mobile application. The prediction is done based on the increase or decrease of the HRV value. They used the common Cleveland heart disease datasets from the UCI machine learning repository for algorithm training and testing. The sample size is commonly known with 303 instances and 76 features. Half of the sample was used for training and building of the classification model. The machine learning algorithms used were DT, NB, SVM, KNN, LR, and RF. Experimentation with these algorithms yielded a result with an accuracy of 89% in predicting the heart disease using the RF classifier. For the monitoring purposes, two experiments were carried out. In the first instance, 20 healthy individuals were used, from which 5 were asked to hold the model and play sorts, no alarm was raised, because they were all healthy. In the second instance, also 20 but unhealthy individuals were used and alarm was raised, an SMS was sent to the nearest hospital. Their results showed that in both cases an accuracy of 100% was achieved. Their model was built based on the assumption that mobile phones were connected to internet at all times, which is not guaranteed.

Reference [35] presented a heart disease prediction framework that shows how synthetic data would be used to address privacy concerns and overcome constraints inherent in small medical research datasets. They examined the used of surrogate datasets comprising synthetic observations for modelling the system. The data was generated based on the characteristics of original observations and compared the prediction accuracy results obtained from LR, DT and RF. The UCI dataset of the Cleveland heart disease data containing 303 instances with 76 features, which was preprocessed to become 279 instances and 14 features was used as the original data. The experiment was divided into three stages. In the first stage, the base line models and their results were established, which fundamental objective was to validate and compare the accuracy and stability of the results of the proposed models to those in the previous studies. In the second stage, the same original data (Cleveland: 279 instances and 14 features) was used to generate 50,000 records, and it was used to train and test the previous LR, DT, and RF models. In the last stage, 60,000 records were generated from the same original dataset, and was used to train and test the ANN model of the perceptron forward and backward propagation algorithm type. From their experiments using the traditional models (LR, DT, and RF) with the surrogate data, they achieved an improved prediction stability within 2% variance at around 81% using 10-fold cross validation. While using the ANN with the surrogate data, they improved the heart disease prediction accuracy by nearly 16% to 96.7% while maintaining stability at 1%.

Reference [38] developed a heart disease prediction system which they compared with NB algorithm performance. The Cleveland data obtained from the UCI machine learning repository was employed during model training and testing scenarios. The sample size consists of 303 instances and 76 features. The proposed system receives patients information which would be classified a 0 absent

"Absent" or 1 which means "Present" of heart disease. Results of the comparative analysis with the NB classifier showed that their proposed model performed better with classification accuracy of 97%.

Reference [36] performed a comparative study on various machine learning algorithms for predicting patients with heart disease through graphical representation of results. For the algorithm training and testing, the Cleveland heart disease datasets from the UCI machine learning repository were used. The sample is originally composed of 303 instances with 76 features, from which 14 were used for the classification. Classifiers in WEKA, which include NB, SVM, DT, and KNN were used. Experimental results were represented graphically, which showed that NB classifier performed better than the other classifiers in predicting patients with heart diseases correctly.

Reference [30] proposed a heart attack risk prediction using smartphones and data mining methods. They developed an android application by incorporating clinical data obtained from patients who were admitted with chest pain in a cardiac hospital. Datasets collected from a particular hospital containing 917 instances and 70 attributes. Of the 917 instances, 636 were collected from a cardiac hospital while 281 instances were collected from health camps irrespective of their symptoms and presence of heart disease. They Chi-square test, Fisher's Exact Test, Probability, Percentage and Ratios to calculate the risk score. The data and the generated risk score were integrated to the android application, which they named Predict-Risk. In the android application the risk was categorized as per score generated for variables of risk factors but if the user gives an input of having one or more symptoms, the risk level ascends up by one. Accuracy of prediction depends on the heart disease features used, but 70 is too much and could lead to unreliable results.

Reference [20] proposed a Neural Network –based prediction of coronary heart disease risk using feature correlation analysis (NN-FCA) using two stages, feature selection and feature correlation analysis. In the first of the system process, KNHANES-V1 dataset was selected and in the second step, statistical analysis was performed to identify features related to coronary heart disease risk. In the third step, predictors of coronary heart disease risk were selected using feature sensitivity- based feature selection. In the fourth step, Neural Network (NN)-based coronary heart disease risk predictors were trained using feature correlation analysis of features. In the fourth step, performance measures were made to validate NN-based coronary heart disease risk predictions using feature correlation analysis. The KNHANES-V1 was conducted by the Korean Centre for Disease Control and Prevention to obtain the datasets. The sample size contains 8108 instances from which 3324 were excluded due to uncertainty. And 630 instances were below the age of 30 years. So, the resulting sample for coronary heart disease related was 4146 instances. The input variables for the model training were age, sex, cholesterol, blood pressure, and other related features. The output variables were high blood pressure, dyslipidemia, stroke, myocardial infarction, and angina. When these 5 are not present, coronary heart disease of low risk. But when 1 of the 5 is present, coronary heart disease is of high risk. The statistical analysis was performed using IBM SPSS version 22.0. Confusion matrix and Receiver Operating Characteristics (ROC) were used for performance comparison of the classifiers. The experimental results showed that the NN-FCA model was as good as FRS model in terms of the coronary heart disease prediction. Compared to the validation of the FRS for the Korean population, the NN-FCA model resulted in a large ROC curve and more accurate coronary heart disease risk prediction. There are other significant indicators, such as diabetes, cholesterol level, etc. that might be a sign of high risk of heart disease but not incorporated, which may lead to late prediction.

Reference [47] designed a framework for heart disease prediction using data mining techniques. One of the UCI datasets was used to train and test the system using 10-fold cross validation method. SVM, NB, KNN, C4.5, Back Propagation classifiers were used and performances were compared. SVM classification algorithm appeared the best in terms of accuracy, sensitivity, precision, low specificity, mean absolute error and low computing times in all feature combinations. The SVM classifier accuracies at 13, 12, 11, 10, 9, 8 and 7 feature combinations were 83.70%, 84.00%, 84.00%, 84.10%, 84.40%, 84.80%, and 85.90% respectively. Datasets employed for algorithm training and testing were not clearly specified.

Reference [24] performed a comparative study between two heart disease prediction techniques. The compared techniques were Framingham Risk Score (FRS) and Quantum Neural Network (QNN) algorithms. They used heart disease datasets consisting 689 instances for model training and 5,209 datasets of the Framingham study conducted on patients, and was taken from the University of Washington, Seattle, WA, USA, for the validation. During training process of the QNN, the best possible weights were identified for each of every layer by conducting different experiments. The QNN architecture consists of 7 input nodes, 85 hidden nodes, and 1 output node. The numbers of hidden were identified after several experiments. The QNN experimental results were compared with that of the Framingham Risk Score (FRS) using the same parameters, where it achieved an accuracy up to 98.57%.

Reference [27] compared the performances of J48, Logistic Model Tree (LMT), and Random Forest (RF) algorithms in WEKA for heart disease prediction task. The Cleveland datasets from the UCI, which commonly consist 303 instances and 76 features were used for the training and testing using the 10-fold cross validation method. The evaluation was based on accuracy, sensitivity, and specificity. Experimentation was carried out on Core i3 with 2.4GHz CPU and 4GB RAM. Results showed that J48 appeared with the highest classification accuracy of 56.76% followed by LMT at 55.77%, then RF.

Reference [46] proposed an SVM based approach with Framingham health parameters for risk prediction of cardiovascular diseases to ensure high sensitivity and accuracy. They used datasets obtained from the Blue Mountain Eye Study (BMES) database, which was created from a population based cohort study, where eye and other health outcomes in an urban Australian population for patients greater than 49 years of age. The study was carried out under the approval of the Western Sydney Area Health Service Human Research Ethnic Committee. The sample size consists of 2406 people with 1450 females and 956 males. The data was divided into two as 80% and 20% for model training and testing respectively. The SVM performance was compared with LR and Framingham Risk equation (FRE) on 104 cardiovascular cases. From the experimental results, it was observed that the correct prediction using FRE was 40, using LR was 50, and using SVM was 71. A confusion matrix was created for each of the three models. The number of false positives when the prediction was performed using FRE model was 108, using LR was 68, and using SVM was 57. From the results obtained, SVM classifier performed better than LR and FRE in correct prediction of cardiovascular cases. The model could not predict patients with CVD cases who are below 50 years of age.

Reference [34] proposed an intelligent system framework for heart disease prediction using NB classifier, which was implemented on java platform. The study was carried out in comparison with DT algorithm prediction performance. In their implemented system, patients were to enroll their required information which would be stored in the system database, and the classification would be done automatically during enrollment. Upon entering the information, patients would be classified as either heart disease or normal. The information is viewed by a medical professional.

Reference [10] presented a heart disease prediction model using hybridization of data mining techniques to help medical practitioners in detecting the heart disease status based on the patient clinical data. Each of the popular algorithms selected which include NB, SVM, and NN was analyzed in isolation. Subsequently, the three classifiers were combined into a single hybrid model to obtain different performance. In their conclusion, they were of the view that the accuracy of each of the algorithms used (NB, SVM and NN) could be enhanced through hybridization.

Reference [6] presented a framework called "Heart Attack Prediction using Data mining Techniques" using Fuzzy C Means classifier to predict the risk of heart attack in patients. The Fuzzy C means is an unsupervised machine learning clustering algorithm that allows one piece of data to belong to two or more clusters. The datasets used for the model training and testing were obtained from the University of California, Irvine (UCI) machine learning repository. The sample size contains 270 instances and 73 heart disease features, in which only 13 were used for the heart attack prediction. Data preprocessing was carried out to remove missing values. The results of classification experiment performed showed that the proposed classifier (Fuzzy C Means) achieved better accuracy than most of the existing classification algorithms.

3. Conclusion & Future Work

From the vast literature reviewed, it was observed that most of the studies used Cleveland heart disease dataset, which contains only 303 instances with 14 features. The sample size which represents a particular geographical area is so small and restricted. Few studies that employed other data sources also used a single dataset with limited heart disease features. Therefore, the various classification accuracies obtained in heart disease prediction could not be generalized. To obtain a more generalized classification and prediction accuracy, other multiple heart disease datasets from geographically diverse sources with more features should be explored for developing more efficient machine learning models, and that is the fundamental intent of our future research, which is on progress. With this, more efficient classification and early prediction of heart diseases would be achieved, which in turn minimizes the escalating rates of morbidity and mortality due to CVDs.

References

[1] Alotaibi, F. S. (2019). Implementation of machine learning model to predict heart failure disease. *International Journal of Advanced Computer Science and Applications, 10* (6), 261-268.

[2] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2018). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*. doi: 10.1016/J.TELE.2018.11.007.

[3] Anitha, S., & Sridevi, N. (2019). Heart disease prediction using data mining techniques. *Journal of Analysis and Computation, 8* (2), 48-55.

[4] Annepu, D., & Gowtham, G. (2019). Cardiovascular disease prediction using machine learning techniques. *International Research Journal of Engineering and Technology, 6* (4), 3963-3971.

[5] Ayatollahi, H., Gholamhosseini, L., & Salehi, M. (2019). Predicting coronary artery disease: a comparison between two data mining algorithms. *BMC Public Health*. doi: 10.1186/S12889-019-6721-5.

[6] Banu, G. R., & Jamala, J. H. (2015). Heart attack prediction using data mining technique. *International Journal of Modern Trends in Engineering and Research, 2* (5), 428-432.

[7] Benjamin, H., David, F., & Belcy, S. A. (2018). Heart disease prediction using data mining techniques. *ICTACT Journal of Soft Computing, 9* (1), 1824-1830.

[8] Chaithra, N., & Madhu, B. (2018). Classification models on cardiovascular disease prediction using data mining techniques. *Journal of Cardiovascular Diseases and Diagnosis*. doi: 10.4172/2329-9517.1000348.

[9] D'Souza, A. (2015). Heart disease prediction using data mining techniques. *International Journal of Research in Engineering and Science, 3* (3), 74-77.

[10] Devi, S. K. (2016). Prediction of heart disease using data mining techniques. *Indian Journal of Science and Technology*. doi: 10.17485/ijst/2016/v9i39/102078.

[11] Dulhare, U. N. (2018). Prediction system for heart disease using naïve bayes and particle swarm optimization. *Biomedical Research, 29* (12), 2646-2649.

[12] Gawali, M., & Shirwalkar, N. (2018). Heart disease prediction system using data mining techniques. *International Journal of Pure and Applied mathematics, 120* (6), 499-506.

[13] Haq, A. U., Li, J.-P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Hindawi Mobile Information System*. doi: 10.1155/2018/3860146.

[14] Hariharan, K., Vigneshwar, W. S., Sivaramakrishnan, N., & Subramaniyaswamy, V. (2018). A comparative study on heart disease analysis using classification techniques. *International Journal of Pure and Applied Mathematics, 119* (12), 13357-13366.

[15] Hussein, M. U. (2017, October 29). *Physics and the Cardiovascular System*. Retrieved from ResearchGate: https://www.researchgate.net.

[16] Jagtap, A., Malewadkar, P., Baswat, O., & Rambade, H. (2019). Heart disease prediction using machine learning. *International Journal of Research in Engineering, Science and Management, 2* (2), 352-355.

[17] Kashyap, A. (2018). Artificial intelligence and medical diagnosis. *Scholars Journal of Applied Medical Sciences*, 4982-4985. doi: 10.21276/sjams.2018.6.12.61.

[18] Khan, S. N., Nawi, N. M., Shahzad, A., Ullah, A., & Mushtaq, M. F. (2019). Comparative analysis for heart disease prediction. *International Journal on Informatics Visualization, 1* (4-2), 227-231.

[19] Khourdifi, Y., & Bahaj, M. (2018). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International Journal of Intelligent Engineering and Systems, 12* (1).

[20] Kim, J. K., & Kang, S. (2017). Neural network-based coronary heart disease risk prediction using feature correlation analysis. *Hindawi Journal of Healthcare Engineering*. doi: 10.1155/2017/2780501.

[21] Lakshmanarao, A., Swathi, Y., Sri, P., & Sundareswar, S. (2019). Machine learning techniques for heart disease prediction. *International Journal of Science and Technology Research, 8* (11), 374-377.

[22] Nagendra, K. V., & Ussenaiah, M. (2018). A study on various data mining techniques used for heart diseases. *International Journal of Recent Scientific Research*, 24350- 24354.

[23] Nandhini, S., Debnath, M., Sharma, A., & Pushkar. (2018). Heart disease prediction using machine learning. *International Journal of Recent Engineering Research and Development, 3* (10), 39-46.

[24] Narain, R., Saxena, S., & Goyal, A. K. (2016). Cardiovascular risk prediction: a comparative study of Framingham and quantum neural network based approach. *Dovepress Journal: Patient Preference and Adherence, 10*, 1259-1270.

[25] Nashif, S., Raiban, M., Islam, M., & Imam, M. H. (2018). Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. *World Journal of Engineering and Technology, 6*, 854-873.

[26] Padmanabhan, M., Yuan, P., Chada, G., & Nguyen, H. V. (2019). Physician-friendly machine learning: a case study with cardiovascular disease risk prediction. *Journal of Clinical Medicine*. doi: 10.3390/jcm8071050.

[27] Patel, J., Upadhyay, T., & Patel, S. (2016). Heart disease prediction using machine learning and data mining techniques. *IJCSC, 7* (1), 129-137.

[28] Prasad, R., Anjali, P., Adil, S., & Deepa, N. (2019). Heart disease prediction using logistic regression algorithm using machine learning. *International journal of Engineering and Advanced Technology, 8* (3S), 659-662.

[29] Rabbi, M. F., Uddin, M. P., Ali, M. A., & Kibria, M. F. (2018). Performance evaluation of data mining classification techniques for heart disease prediction. *American Journal of Engineering Research, 7* (2), 278-283.

[30] Raihan, M., Mondal, S., More, A., Boni, P. K., & Sagor, M. F. (2017). Smartphone based heart attack risk prediction system with statistical analysis and data mining approaches. *Advances in Science, Technology and Engineering Systems Journal, 2* (3), 1815-1822.

[31] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning algorithms: a survey. *International Journal of Engineering and Technology, 7* (2.8), 684-687.

[32] Rammal, H., & Emam, A. Z. (2018). Toward robust heart failure prediction models using big data techniques. *In Proceedings of the Tenth International Conference on e-Health, Telemedicine and Social Medicine*, 85-91.

[33] Reddy, P. K., Reddy, T. S., Balakrishnan, S., Basha, S. M., & Poluru, R. K. (2019). Heart disease prediction using machine learning algorithm. *International Journal of Innovative Technology and Exploring Engineering, 8* (10), 2603-2606.

[34] Ritesh, T., Gauri, B., Ashwini, D., & Priyanka, S. (2016). Heart attack prediction system using data mining. *International Journal of Innovative Research in Computer and Communication Engineering, 4* (8), 15582-15585.

[35] Sabay, A., Harris, L., Bejugama, V., & Jaceldo-Siegl, K. (2018, December 24). *Overcoming small data limitations in heart disease prediction by using surrogate data*. Retrieved from SMU Data Science Review: https://scholar.smu.edu/datasciencereview/vol1/iss3/12.

[36] Sen, S. K. (2017). Prediction and diagnosis of heart disease using machine learning algorithms. *International Journal of Engineering and Computer Science, 6* (6), 21623-21631.

[37] Shamsollahi, M., Badiee, A., & Ghazanfari, M. (2019). Using combined descriptive and predictive methods of data mining for coronary artery disease prediction: a case study approach. *Journal of Artificial Intelligence and Data Mining, 7* (1), 47-58.

[38] Sharmila, S., & Gandhi, M. P. (2017). Analysis of heart disease prediction using data mining techniques. *International Journal of Advanced Networking and Applications, 8* (5), 93-95.

[39] Shirsath, S. S., & Patil, S. (2018). Disease prediction using machine learning over big data. *International Journal of Innovative Research in Science, Engineering and Technology, 7* (6), 6752-6757.

[40] Singh, N., & Jindal, S. (2018). Heart disease prediction system using hybrid technique of data mining algorithms. *International Journal of Advanced Research, Ideas and Innovations in Technology, 4* (2), 982-987.

[41] Singh, P., Singh, S., & Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. *International Journal of Nanomedicine*. doi: 10.2147IJN.S124998.

[42] Solanki, A., & Barot, M. P. (2019). Study of heart disease diagnosis by comparing various classification algorithms. *International Journal of Engineering and Advanced Technology, 8* (2S2), 40-42.

[43] Sridhar, A., & Kapardhi, A. (2018). Predicting heart disease using machine learning algorithm. *International Research Journal of Engineering and technology, 6* (4), 36-38.

[44] Subhadra, K., & Vikas, B. (2019). Neural network based intelligent system for predicting heart disease. *International Journal of Innovative Technology and Exploring Engineering, 8* (5), 484-487.

[45] Tarawneh, M., & Embarak, O. (2019). Hybrid approach for heart disease prediction using data mining techniques. *Acta Scientific Nutritional Health, 3* (7), 147-151.

[46] Unnikrishnan, P., Kumar, D. K., Arjunan, S. P., Kumar, H., Mitchell, P., & Kawasaki, R. (2016). Development of health parameter model for risk prediction of CVD using SVM. *Computational and Mathematical Methods in Medicine*. doi: 10.1155/2016/3016245.

[47] Voleti, S. R., & Reddi, K. K. (2016). Design of an optimal method for disease prediction using data mining techniques. *International Journal of Advanced Research in Computer Science and Software Engineering, 6* (12), 328-337.

[48] WHO. (2011). *Global Atlas on cardiovascular disease prevention and control.* Geneva: WHO Library Cataloguing.

[49] WHO. (2016). *Technical package for cardiovascular disease management in primary health care.* Geneva: WHO Library Cataloguing.

[50] WHO. (2017). *Global action plan for the prevention and control of noncommunicable diseases.* Geneva: WHO Library Cataloguing.