# A SURVEY OF PERFORMANCES OF SOME SELECTED MACHINE LEARNING ALGORITHMS FOR CARDIOVASCULAR DISEASE PREDICTION

LAMIDO YAHAYA[1*], IBRAHIM HASSAN[2] AND ABBAS MUHAMMAD RABIU[3]

[1]Department of Computer Science, Faculty of Science, Gombe State University, Gombe, Gombe State
[2]Department of Computer Science, School of Science, Gombe State Polytechnic, Bajoga, Gombe State
[3]Department of Computer Science, Faculty of Computing, Federal University Dutse, Jigawa State
Corresponding Author: yahaya.lmd@gmail.com

## ABSTRACT

Cardio-Vascular Diseases (CVDs) are the leading causes of early deaths in the world. Middle- and low-income countries suffer the biggest challenge of effective diagnosis and treatment due to the inadequacy of efficient diagnostic tools and physicians. This affects the proper prediction and treatment of patients. Though, large proportion of CVDs could be prevented but they continue to escalate mainly because preventive measures put in place are inadequate. Huge CVD data is available in the healthcare sector which led to several researches. The University of California, Irvine (UCI) heart disease data has been used extensively by machine learning researchers in trying to come up with a more efficient predictive model. Previously, the focus was on investigating the performances of some selected machine learning algorithms on the UCI data. These algorithms include Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT-J48), Random Forest (RF), K-Nearest Neighbor (KNN) and Artificial Neural Network (ANN). There are few researchers who used CVD datasets other than that of the UCI. This paper carries out a survey on the performances of these algorithms on CVD prediction using 12various datasets other than the UCI data. From our investigation on the 18 researches conducted, most of them in 2018 and 2019, we found that DT-J48, NB and SVM gained much attention than any other algorithm, where J48 was used 11 times and appeared the most used algorithm for developing clinical decision support systems. NB and SVM appeared 10 and 9 times respectively, ANN was employed 8 times, while KNN and LR were considered 3 times each. RF appeared with the least frequency of 1 only. Finally, it has been discovered that no single algorithm would be generalized as the best in CVD prediction based on the data in which it was used.

**Keywords:** Machine Learning, Algorithms, Heart Disease, Classification, Prediction

## INTRODUCTION

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 (Hussein, 2017). Any abnormal condition that obstructs normal blood circulation or flow from the heart would result in several and severe complications of heart diseases. These are commonly called Cardio-

Vascular Diseases (CVDs), and remained among the deadliest diseases in the world (Umasankar & Thiagarasu, 2019). The world Health Organization (WHO) has reported severally on the trends of CVDs worldwide. The most recent report came in the year 2017, which showed that the deadliest disease is still escalating. The report stated that 17.9 million people die each year from CVDs, an estimated 31% of all deaths worldwide, from which 85% are due to heart attack and stroke (WHO, 2017).However, there are certain risk factors that increase a person's chances of having a CVD. They include family history of CVDs, high level of bad cholesterol, low level of good cholesterol, high blood pressure, high fat diet, physical inactivity, and obesity. Other risk factors include smoking, diabetes, age and gender. Cardiovascular diseases are of various types, some of which were listed by Nagendra & Ussenaiah (2018):

1. Coronary heart disease;
2. Angina pectoris;
3. Congestive heart failure;
4. Cardiomyopathy;
5. Congenital heart disease;
6. Arrhythmias; and
7. Myocarditis

It has become a global concern that CVD cases with high morbidity and mortality rates are increasing globally. Machine learning techniques play a very vital role in the medical data analysis and knowledge extraction. The increasing morbidity and mortality due to CVDs worldwide has attracted the attention of researchers to conduct many studies in their effort to minimize the rates (Yahaya, et al., 2020; Ashraf, et al., 2019). Machine learning techniques have been widely used in the implementation of clinical decision support systems for CVD prediction. Some of the recent studies conducted using these techniques include Sridhar & Kapardhi (2018); Jagtap et al.(2019); Annepu & Gowtham (2019); and Subhadra & Vikas (2019).

This study is fundamentally focused on analyzing the performances of these algorithms in the prediction of CVDs. Algorithms such as Naïve Bayes (NB), Artificial Neural Network (ANN), Decision Tress (DT), K-Nearest Neighbor (KNN), Logistic Regression (LR), and Support Vector Machine (SVM) are among the most popular techniques used in CVD data analysis and prediction. These algorithms can be used to enhance the data storage for practical and legal purposes (Nikhar & Karandikar, 2016). The theoretical background of these algorithms was briefly presented by Haq et al.(2018) as follows:

I. **Logistic Regression (LR)**: For binary classification problem, in order to predict the value of predictive variable *y* when $y \in [0, 1]$, 0 is negative class and 1 is positive class. It also uses multi-classification to predict the value of y when $y \in [0, 1, 2, 3]$.

II. **Support Vector Machine (SVM)**: This is a machine learning algorithm which has been mostly used for classification problems, such as the CVD prediction. It has been used extensively on the heart disease data for CVD classification. SVM used a maximum margin strategy that transformed into solving a complex quadratic programming problem.

III. **Naïve Bayes (NB):** The NB is a classification supervised learning algorithm. It is based on conditional probability theorem to determine the class of a new feature vector. The NB uses the training dataset to find out the conditional probability value of vectors for a given class. After computing the probability conditional value of each vector, the new vectors class is computed based on its conditionality probability.

IV. **Artificial Neural Network (ANN):** The ANN is a mathematical model that integrates neurons that pass messages. The ANN has three components including inputs, outputs, and transfer functions. The input units take extraordinary values and weights, which are modified during the training process of the network. The output of the artificial neural network is calculated for the known class; the weight is recomputed using the error margin between the output of predicted and actual class.

V. **Decision Tree (DT):** A decision tree shape is just a tree where every node is a leaf node or decision node. The techniques of the decision tree are simple and easily understandable for how to take the decision. A decision tree contained internal and external nodes linked with each other. The internal nodes are the decision-making part that makes a decision and the child node to visit the next nodes. The leaf node on the other hand has no child nodes and is associated with a label.

VI. **K-Nearest Neighbor (K-NN):** KNN is a machine learning classification algorithm that predicts the class label of a new input; K-NN utilizes the similarity of new input to its input's samples in the training set if the new input is same the samples in the training set. Let *(x, y)* be the training observations and the learning function h: $X \longrightarrow Y$, so that given an observation *x, h(x)* can determine *y* value.

VII. **Random Forest (RF):** RF is also a popular machine learning algorithm that can be used for both regression and classification tasks but generally performs better in classification tasks. As the name suggests, RF technique considers multiple decision trees before giving an output. So, it is basically an ensemble of decision trees. For classification, it uses a voting system and then decides the class whereas in regression it takes the mean of all the outputs of each of the decision trees (Ramalingam et al., 2018). It works well with large datasets with high dimensionality.

According to Sharma & Rizvi (2017), each of the algorithms has its capability for instance, NB uses probability for predicting the heart disease, whereas DT is used to provide a classified report for the heart disease, while ANN provides opportunities to minimize the error in the prediction of heart diseases. In fact, there exists a wide gap between the accuracy of traditional heart disease prediction done by medical professionals and that of the modern techniques. Kirubha & Priya (2016) stated that according to the latest survey conducted by WHO, the medical professional is able to correctly predict only 67% of heart disease. Lashari et al. (2018)

stated that machine learning applications might benefit the healthcare industry immensely but this depends on how clean the data is.

Therefore, the fundamental intent of this paper is to check the performances of the selected algorithms, so that best ones in the prediction of CVDs would be figured out and considered for developing baseline predictive models. This will help to ensure more accurate and early prediction of patients, which can eventually help them take preventive measures to minimize the morbidity and mortality rates due to CVDs. However, for the contribution of our paper, we discovered that no single algorithm could be generalized as the best for all types of CVD datasets, but rather based on a particular data due to variability in the number of instances, dimensionality, and presence of noise.

The remaining part of this paper is organized as follows: in section 2.0, we presented some previous researches conducted based on the 12 different CVD datasets, wherein stated the performances of each of the selected algorithms. In section 3.0, we presented the various datasets for heart disease or CVD predictions used by several researchers. There are 12 different datasets, where some of them were clearly described while some were not. In section 4.0, we presented a comparison table showing the references and the datasets employed, the algorithms used, the best performing models and their respective prediction accuracies. In section 5.0, we presented a summarized discussion on the selected algorithms as well as a table indicating their cumulative frequencies of use in various references. In section 6.0, we gave the concluding remark about the major findings of our paper.

## Literature Review

This section of the paper gives a survey on various research works conducted based on various heart disease or CVD datasets, obtained from various medical institutions in the world. The articles are 18 in number, extracted from a previous study, which is a comprehensive review on heart disease prediction, comprising all types of heart disease datasets. The researches were conducted in the year 2019, 2018, 2017, 2016 and only 1 in 2013. The extracted research articles are as follows:

Prasad et al. (2019) proposed a Logistic Regression (LR) based approach of machine learning for heart disease prediction. Other algorithms such as NB, SVM, DT-J48, and KNN were also explored using SK-Learn library in Python software for performance comparisons with the LR algorithm. But the dataset used was not specified. According to them, the experimental results showed that the LR algorithm performed better at 86.89% accuracy. While the remaining algorithms performed at 77.85% for KNN, 86% for NB, 78.69% for DT-J48 and 82% for SVM.

Ayatollahi et al. (2019) 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 record 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. It provided a better classification for the prediction of cardiovascular diseases.

Lakshmanarao et al. (2019) 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-J48, 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-J48 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.

Reddy et al.(2019) implemented a machine learning-based approach for heart disease prediction using comparative analysis of DT-J48 and SVM classification algorithms in Python. Age, chest pain, blood pressure, cholesterol level was 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 library and Mat Plot Lib was used for data visualization. Experimental results showed that DT-J48 classifier performed much better than the SVM. The DT-J48 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-J48 classifier performance could not be generalized as the best model for heart disease prediction despite of the 100% classification accuracy.

Shamsollahi et al. (2019) 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.

Kutrani & Elthalhi (2019) carried out a study to predict whether a patient needs a cardiac catheterization procedure or not. Five popular classification algorithms including SVM, DT-J48, KNN, ANN, and NB were used in the prediction of the catheterization procedure based on the prediction accuracy. The study was carried out using a home dataset obtained from Benghazi Heart Disease Centre, which is a real data of patients who underwent cardiac catheterization from December 2003 to May 2007. The datasets consist of 1770 instances and 11 features. Data preprocessing was carried out to remove missing values and the sample size became 1427 instances and only 9 attributes. Results of the experiments for the five classification algorithms showed that NB, ANN, and DT-J48 had the highest prediction accuracies, but in general J48 without the *smoker* attribute was the best to predict whether a patient needs a cardiac catheterization procedure with an accuracy of 89%.

Rammal & Emam (2018) 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: SVM and DT-J48 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.

Sridhar & Kapardhi (2018) proposed a heart disease prediction based on machine learning techniques using NB and DT-J48 algorithms in Python. The datasets used for training and testing of the model were obtained from the Kaggle website, which contains 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 in Python. Form their experiments, results showed that DT-J48 algorithm performed better than the NB in the prediction of heart diseases.

Shirsath & Patil (2018) 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%.

Meda & Bhogapathi (2018) proposed a heart disease prediction framework called Fuzzy Neural Genetic Algorithm (FNGA). Datasets used for their model training and testing purposes, were obtained from Andhra Pradesh population, where they used attributes such as sex, age, fasting blood sugar (FBS), chest pain, etc. for the classification. A fuzzy technique was employed as the preprocessing advance to order the information into lower, medium, and higher classes. The FNGA model was evaluated against some of the most popular classification algorithms, which include SVM, NB, DT-J48, and Fuzzy C-means classifiers. The evaluation was done based on the execution parameters such as Average Running Time, Average Execution Time, Execution Time, False Negative, False Positive, True Negative, True Positive, Precision, Recall, and Accuracy. The FNGA model appeared the best with a classification accuracy of 98.6%.

Chaithra & Madhu (2018) 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.

Raihan et al. (2017) 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.

Kim & Kang (2017) 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.

Narain et al. (2016) 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%.

Unnikrishnan et al. (2016) 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.

Ritesh et al. (2016) 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. The contribution of this study was the system ability to predict heart disease at early stage as stated by the developers.

Devi et al.(2016) 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.

Taneja (2013) used a “knowledge discovery in database (KDD)” methodology to develop a heart disease prediction model that can predict heart disease cases based on the measurements taken from transthoracic echocardiography examination. The datasets used were obtained from PGI, Chandigarh, which contains transthoracic echocardiography examination report of 7008 patients for the period of 2008 to the first quarter of 2010. The datasets contain 20 attributes which were reduced to 15 after expert consultations. Some of the popular machine learning algorithms including DT-J48, NN, and NB were used for classification and prediction processes. 10-fold cross validation technique was employed during training and testing of the model. Of the three selected classification algorithms, experimental results showed that DT-J48 performed better than NN and NB with a prediction accuracy of 95.56%, though implemented on some selected attributes.

From the various researches investigated, it was discovered that not much attention was paid to using other CVD datasets by machine learning researchers for developing clinical decision support systems. More than 60% of such systems were developed based on the UCI heart disease data, which is too narrow in scope for generalization. However, it was observed that from the eighteen researches, six were conducted in 2019, five were conducted in 2018, while two in 2017, four in 2016 and one in 2013.

### The CVD Datasets

Previous investigations showed that more than 60% of researches conducted in heart disease or CVD prediction in general, employed the online available UCI data. Few studies considered using other datasets apart from that of the UCI. These datasets were obtained from various sources. They are as follows:

i. AJA University of Medical Sciences, Iran; the sample is composed of 1324 instances and 25 features (Ayatollahi et al., 2019).

ii. Framingham datasets from the Kaggle website, which contains 4239 instances with 15 features (Lakshmanarao et al., 2019).

iii. King Saud Medical City (KSUMC) system, Riyadh, Saudi Arabia. The datasets obtained were in structured, semi-structured, and unstructured format, comprising 100 real patient records with many missing values and misidentified attributes, extracted from the KSUMC Electronic Health Record (EHR)(Rammal & Emam, 2018).

iv. Benghazi Heart Disease Centre, Libya, which is a real data of patients who underwent cardiac catheterization from December 2003 to May 2007. The datasets consist of 1770 instances and 11 features (Kutrani & Elthalhi, 2019).

v. Datasets containing 282 instances with 58 features obtained from a clinic (Shamsollahi et al., 2019). The source of this dataset was not clearly specified.

vi. Andhra Pradesh population, India; the sample size was not clearly described, as the number of instances and features were not specified (Meda & Bhogapathi, 2018).

vii. Transthoracic Echocardiography database, which contains 336 instances and 24 attributes (Chaithra & Madhu, 2018). The source of this dataset was not clearly specified.

viii. Korean Centre for Disease Control and Prevention datasets. The sample size contains 8108 instances from which 3324 were excluded due to uncertainty (Kim & Kang, 2017).

ix. 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 (Raihan et al., 2017). The source was not clearly specified.

x. The 5,209 datasets of the Framingham study conducted on patients, and was taken from the University of Washington, Seattle, WA, USA (Narain et al., 2016).

xi. 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 were considered. 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 (Unnikrishnan et al., 2016).

xii. The datasets from PGI, Chandigarh, which contains transthoracic echocardiography examination report of 7008 patients for the period of 2008 to the first quarter of 2010, which contains 20 attributes which were reduced to 15 after expert consultations (Taneja, 2013).

These are the various datasets used by researchers for CVD risk prediction other than the well-known UCI data. The machine learning algorithms were applied to classify CVD patients from those that are normal, where various prediction accuracies were obtained. The datasets have varying numbers of instances and features, which in fact affect the prediction accuracies of the used algorithms.

## MATERIALS AND METHOD

This section presents a comparison table showing the algorithms used in each reference, the best performing algorithms and their performance accuracies in various datasets has been illustrated in Table 1.

### Algorithms' Performance Analysis

In either of the presented CVD datasets, one or more of the selected machine learning algorithms were used for the prediction. The datasets were used to train, test, and evaluate the machine learning models in predicting CVD risk cases for clinical decision support purposes. The algorithms include LR, ANN, KNN, SVM, DT-J48, NB, and RF. From the 18 different researches investigated that used the 12different CVD datasets, DT-J48 was the most used algorithm, up to 11 times followed by SVM and ANN with 9 and 8 respectively. Though there are few researches which did not clearly specify the dataset used. NB was employed for the CVD prediction up to 10 times. The remaining 3 algorithms are LR, KNN and RF, where LR and KNN were employed 3 times each, while RF has the least frequency, which employed once only. Detail of the frequencies together with the references in which each algorithm appeared was presented in table 1 in section 4.0 of this paper. These algorithms with their corresponding reference numbers in which they were used as indicated in Table 1 and frequencies of use are shown in Table 2.

The datasets mentioned are entirely different in number of instances, features or attributes as well as format, and each of the selected algorithms has been used on one or more datasets for the CVD prediction. Therefore, the performance of an algorithm could not be ascertained and generalized based on the dataset in which it was used. This is because its performance is highly dependent upon the type of the data. Datasets with high dimensionality and noise tend to degrade algorithm performance, where an efficient algorithm could perform poorly. Each of the researches used one or more of the selected algorithms on one of the mentioned datasets, where performances were compared. According to the findings of this study, an algorithm might appear the best in the prediction using a particular dataset, and the same algorithm could be the worst on another dataset due to variability. Therefore, the algorithm cannot be generalized as the best against others.

**Table 1:** Comparison Table of Algorithms and their Performances

| S/N | Reference | Dataset Used | Algorithms Used | Best Algorithm | Accuracy |
|---|---|---|---|---|---|
| 1. | Prasad et al. (2019) | Not specified | LR, NB, SVM, DT-J48, and KNN | LR | 86.89% |
| 2. | Ayatollahi et al. (2019) | AJA University of Medical Sciences, Iran | ANN and SVM | SVM | Not specified |
| 3. | Lakshmanarao et al. (2019) | Framingham datasets from the Kaggle website | SVM, LR, KNN, AdaBoost, DT-J48, NB, and RF | SVM | 90.3% |
| 4. | Reddy et al.(2019) | Not specified | DT-J48and SVM | DT-J48 | 100% |
| 5. | Shamsollahi et al. (2019) | Not specified | CHAID, Quest, C5.0, C & DT-J48, and ANN | CHAID | Not specified |
| 6. | Kutrani & Elthalhi (2019) | Benghazi Heart Disease Centre, Libya | SVM, DT-J48, KNN, ANN, and NB | DT-J48 | 89% |
| 7. | Rammal & Emam (2018) | King Saud Medical City (KSUMC) system, Riyadh, Saudi Arabia | SVM and DT-J48 | Not specified | Not specified |
| 8. | Sridhar & Kapardhi (2018) | Kaggle website | NB and DT-J48 | DT-J48 | Not specified |
| 9. | Shirsath & Patil (2018) | Not specified | NB, CNN-UDRP CNN-MDRP | CNN-MDRP | 94.80% |
| 10. | Meda & Bhogapathi (2018) | Andhra Pradesh population, India | FNGA, SVM, NB, DT-J48, and Fuzzy C-means | FNGA | 98.6% |
| 11. | Chaithra & Madhu (2018) | Transthoracic Echocardiography database | DT-J48, NB, and ANN | ANN | 97.91% |
| 12. | Raihan et al. (2017) | Not specified | Not specified | Not specified | Not specified |
| 13. | Kim & Kang (2017) | Korean Centre for Disease Control and Prevention | ANN | ANN | Not specified |
| 14. | Narain et al. (2016) | University of Washington, Seattle, WA, U.S.A. | ANN and Framingham Risk Score (FRS) | ANN | 98.57% |
| 15. | Unnikrishnan et al. (2016) | Blue Mountain Eye Study (BMES) database, Australia | LR, FRE, and SVM | SVM | Not specified |
| 16. | Ritesh et al. (2016) | Not specified | NB and DT-J48 | Not specified | Not specified |
| 17. | Devi et al.(2016) | Not specified | NB, SVM, and ANN | Not specified | Not specified |
| 18. | Taneja (2013) | PGI, Chandigarh, India | DT-J48, ANN, and NB | DT-J48 | 95.56% |

**Table 2:** Algorithms' Frequency of Use

| S/N | Reference No. in Table1 | Algorithm | Frequency of Use |
|---|---|---|---|
| 1. | 1, 3, 4, 5, 6, 7, 8, 10, 11, 16, 18 | DT-J48 | 11 |
| 2. | 1, 3, 6 | KNN | 3 |
| 3. | 1, 3, 15 | LR | 3 |
| 4. | 1, 3, 6, 8, 9, 10, 11, 16, 17, 18 | NB | 10 |
| 5. | 2, 5, 6, 11, 13, 14, 17, 18 | ANN | 8 |
| 6. | 3 | RF | 1 |
| 7. | 1, 2, 3, 4, 6, 7, 10, 15, 17 | SVM | 9 |

## CONCLUSION AND FUTURE WORK

Though most of the clinical decision support systems developed for CVD prediction are based on the UCI data but there are other researchers who employed different datasets for the same purpose using various machine learning algorithms. Various prediction accuracies were analyzed but it was not possible to generalize the performance of a single algorithm as the best. This is because each of the algorithms has its own uniqueness. For instance, RF usually performs better on a dataset with high dimensionality while some perform poorly on the same data. Here, the best algorithm could be attributed to the higher frequency of use, and from our survey it was discovered that some of these algorithms were the most frequently used for the CVD prediction using the mentioned datasets. From the 18 researches investigated on 12 different datasets (though there are few datasets that were not clearly specified), DT-J48 appeared with the highest frequency, which is up to 11 times. It was succeeded by NB and SVM with 10and 9 respectively. RF and CHAID algorithms did not get much attention from those researchers because they were employed once each. Moreover, hybrid algorithms were also not much considered. FNGA and CNN-MDRP were the hybrid approaches considered once each. Therefore, our study discovered that no single algorithm could be generalized as the best for all types of CVD datasets but rather based on a particular data due to variability in the number of instances, dimensionality, and presence of noise.

To obtain a more generalized prediction accuracy, three different datasets would be employed to train, test and evaluate the three most frequently used algorithms (DT-J48, NB and SVM). The best predictive model would be considered for implementation as a clinical decision support system on CVD prediction, and that is our future research which is currently on progress.

## REFERENCES

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

Ashraf, M., Rizvi, M. A. & Sharma, H., 2019. Improved heart disease prediction using deep neural network. *Asian Journal of Computer Science and Technology,* 8(2), pp. 49-54.

Ayatollahi, H., Gholamhosseini, L. & Salehi, M., 2019. Predicting coronary artery disease: a comparison between two data mining algorithms. *BMC Public Health.*

Chaithra, N. & Madhu, B., 2018. Classification models on cardiovascular disease prediction using data mining techniques. *Journal of Cardiovascular Diseases and Diagnosis.*

Devi, S. K., Krishnapriya, S. & Kalita, D., 2016. Prediction of heart disease using data mining techniques. *Indian Journal of Science and Technology.*

Haq, A. U. et al., 2018. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Hindawi Mobile Information System.*

Hussein, M. U., 2017. *Physics and the Cardiovascular System.*

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), pp. 352-355.

Kim, J. K. & Kang, S., 2017. Neural network-based coronary heart disease risk prediction using feature correlation analysis. *Hindawi Journal of Healthcare Engineering.*

Kirubha, V. & Priya, S. M., (2016. Survey on data mining algorithms in disease prediction. *International of Journal of Computer Trends and technology,* 38(3), pp. 24-128.

Kutrani, H. & Elthalhi, S., 2019. Cardiac catheterization procedure prediction using machine learning and data mining techniques. *IOSR Journal of Computer Science,* 21(1), pp. 86-92.

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), pp. 374-377.

Lashari, S. A., Ibrahim, R., Senan, N. & Taujuddin, N. S. A. M., 2018. Applications of data mining techniques for medical data classification: a review. *MATEC Web of Conferences.*

Meda, S. & Bhogapathi, R. B., 2018. Identification of heart disease using fuzzy neural genetic algorithm with data mining techniques. *Advances in Modelling and Analysis B,* 61(2), pp. 99-105.

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

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,* Volume 10, pp. 1259-1270.

Nikhar, S. & Karandikar, A. M., 2016. Prediction of heart disease using machine learning algorithms. *International Journal of Advanced Engineering, Management and Science,* 2(6), pp. 617-621.

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), pp. 659-662.

Raihan, M. et al., 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), pp. 1815-1822.

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,* pp. 85-91.

Reddy, P. K. M. et al., 2019. Heart disease prediction using machine learning algorithm. *International Journal of Innovative Technology and Exploring Engineering,* 8(10), pp. 2603-2606.

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), pp. 15582-15585.

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), pp. 47-58.

Sharma, H. & Rizvi, M. A., 2017. Prediction of heart disease using machine learning algorithms: a survey. *International Journal on Recent and Innovation Trends in computing and Communication,* 5(8), pp. 99-104.

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), pp. 6752-6757.

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

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

Taneja, A., 2013. Heart disease prediction system using data mining techniques. *Oriental Journal of Computer Science and Technology,* 6(4), pp. 457-466.

Umasankar, P. & Thiagarasu, V., 2019. Data mining for prediction of heart disease: A Literature Survey. *Asian Journal of Computer Science and Technology,* pp. 1-6.

Unnikrishnan, P. et al., 2016. Development of health parameter model for risk prediction of CVD using

SVM. *Computational and Mathematical Methods in Medicine.*

WHO, 2017. *Global action plan for the prevention and control of noncommunicable diseases,* Geneva: WHO Library Cataloguing.

Yahaya, L., Oye, N. D. & Garba, E. J., 2020. A comprehensive review on heart disease prediction using data mining and machine learning techniques. *American Journal of Artificial Intelligence,* 4(1), pp. 20-29.