# Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction

**Isreal Ufumaka**[*]

[*] Computer Science, University of Benin.

DOI: 10.29322/IJSRP.11.01.2021.p10936
http://dx.doi.org/10.29322/IJSRP.11.01.2021.p10936

***Abstract-*** Machine learning has become popular today as so many of its algorithms are now commonly used in different data science projects in various industries especially in the health care sector. It is imperative for researchers and medical professionals to be assisted by machine learning methods in early detection of diseases such as heart disease which is one major killer of humans in our world today. Machine learning algorithms are excellent at learning from data, and since healthcare providers generate huge amount of data on a daily basis, these algorithms can thrive in this field. In this research study, a comparative analytical approach was taken in the determination of which algorithm performs better under the given condition. Various experiments were carried out using cross validation of 5 and 10 folds, to ensure that models created can generalize well enough. This study makes use of data from University of California, Irvine (UCI) machine learning database containing 303 instances with 14 attributes. The collected data is scaled using Min-Max normalization technique. Different popular models are built using supervised machine learning classification algorithms on the scaled data such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), and Gradient Boosting ensemble method. These algorithms are also evaluated using standard performance metrics such as precision, recall, and F1-score. From the experiments carried out, it can be concluded that SVM performs better as it out performs the other algorithms.

***Index Terms***- Classification Algorithms, Gradient Boosting, Logistic Regression (LR), Machine Learning, Support Vector Machine (SVM).

## I. Introduction

The human heart is a vital organ of the human body system. It can be seen as a mechanical device that works by circulating oxygen rich blood to other body organs such as the brain, kidney, lungs, etc. The heart works day and night ensuring that other organs receive their fair share of oxygen rich blood, and a disruption in its activities will affect proper functioning of other organs which could be fatal. Heart disease also known as cardiovascular disease is most times a life threatening medical condition a person suffers from as a result of the inability of the heart to function well enough in its circulatory duties. Some examples of heart diseases are coronary disease, rheumatic disease, and congenital disease to a host of others that plague both the developed and developing world. The World Health Organization (WHO) estimated that heart disease was the top cause of death with 7 million lives lost in 2015 of which greater than 75% of them were in developing countries. This estimate shows that in 2030 about 23.6 million people will die due to heart disease [12].

Various persons engage in unhealthy living routines such as unhealthy diet, smoking of tobacco, heavy drinking, stress and anxiety which can lead to the development of a heart disease. Early detection is the key to reducing the risk of a heart disease although heart disease has been difficult to diagnose [7]. Better decision making based on the available information gotten from health care providers such as in hospitals and clinics could help improve disease prediction as hospitals and clinics create a large pool of bio medical data.

Machine learning provides various computer aided means to predict for the likely hood of a heart disease from heterogeneous medical data. Machine learning is the modern science of getting computers to act without explicitly been programmed. Various machine learning techniques have proven to be useful in the prediction and treatment of diseases such as Alzheimer, Hepatitis, Diabetes, etc. with a high level of accuracy. Machine learning also provides a means of manipulating data in a dire bid to find insight by providing various architectural approach for doing so. Various machine learning algorithms exist that can be used for classification and regression problems or a combination of algorithms like Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Neural Network (NN), etc.

Using machine learning algorithms for heart disease prediction improves accuracy of prediction by exploiting complex interactions between various existing risk factors. The aim of this study is to compare different machine learning models for early prediction of heart diseases.

## II. BACKGROUND AND RELATED WORK

In Arubpradeep and Niranjana [1], the human heart is a very important part of the body. In the study they carried out, the objective was to develop and compare the accuracy between different algorithms such as KNN, SVM, DT, and RF. They compared their developed models using standard performance metric such as accuracy, precision, and recall which showed SVM preforming better than the other models. UCI heart disease dataset was used and it was split into 80% (training) and 20% (testing) and this gave no room for the validation of their model by either creating a validation set or using cross validation as used in our research.

Obiri and Sarku [6], stated that data mining techniques give an alternative approach for an efficient and timely detection of heart disease at quite an early point of the progression of the disease. Their study objective was to perform prediction and comparison on the UCI dataset using different performance matrices on DT, LR, NB algorithms. They made use of 10 fold cross validation alongside single value decomposition feature reduction techniques. In their study, LR performed the best and NB was the worst. They suggested that DT could perform better with more data. Our research study goes a bit further by using an ensemble method which performed impressively.

Kavitha et al. [7], evaluated different machine learning algorithms using standard performance matrices. The algorithms whose performance were evaluated are the SVM, NB, RF, and KNN. This evaluation was done using data obtained from UCI database. They evaluated the performance of their models using only accuracy, precision, and recall but for our research, whereas we considered the harmonic mean (f1-score) of precision and recall.

Raghunath *et al.* [11], opined that the unavailability of proper timely diagnosis of heart disease conditions has caused the death of millions of people. In their study, the objective was to develop and compare different machine learning diagnostic model that is computationally efficient, accurate, and correct for the prediction of heart disease. This objective was achieved by using the KNN classification model to fit their training and prediction data, the same process was repeated using SVM, DT, NB, LR and RF. They opined that their study predicted if given certain characteristic symptoms, a person will get a heart disease or not even in absence of heart disease experts. Their research focus on using the training and testing method of training and evaluating their but in our research study, we utilized the k fold cross validation approach.

According to Kabirirad *et al.* [4], Artificial Neural Netwokr is an effective way of predicting heart disease. A comparison of developed models for high accuracy was the main objective of their study. This was achieved by using neural networks, SVM, principal component analysis networks, Jordan/Elman Network, Radia Based Kernel (RBF) which was implemented on NeuroSolution 7.0 software. It was revealed by their study that MLP, RBF had the best result with 98% of accuracy. In their study 1215 medical instances were used to implement and compare between various algorithms. They divided their data set manually into a training (80%) testing (15%) and validation set (5%), whereas this research uses a cross validation method to ensure validation is achieved with every fold of the data.

According to Ufumaka [3], using machine learning for heart disease prediction provides a way of gaining meaningful insight from data. The study objective was to create a prediction models for heart disease that make use of tuned hyper parameters for heart disease prediction. In the study, this objective was achieved using the MLP and SVM algorithm and grid search cross validation for tuning of hyper parameters used by the models. It was opined that a model using hyper parameter provides a better results than a trial and error approach. However, the tuned models were only experimented with a cross validation of 10 folds, and this alone cannot tell the true performance of a ML model and this is shown in this research study.

Mukherjee and Sharma [5], suggested that there is a need to make health solutions more clear-cut and sound in the modern era of increasing lingering disease such as heart disease. The objective of their study is to design an end to end analytical model for heart disease diagnosis that is both robust and scalable. In their study, they made use of the deep neural network of 10 input layer nodes, 10 hidden layer nodes and 2 output layer to achieve their objective. They opined that their study contributed to the development of an analytical model for the detection of various types of heart related disease. Their study was centered on 30 real time Electrocardiography (ECG) data with 10 attributes. They focused on evaluating their models performance with accuracy alone, and did not consider other metrics which can be miss leading sometimes.

According to Miao and Miao [8], heart disease is one of the leading cause of death in developed countries such as the United States killing over 630,000 persons yearly. Their study objective was to develop a classification and diagnosis model for early and accurate diagnoses of coronary heart disease in patients. The objective of their study was achieved by building Deep Neural Network (DNN) classification and prediction model based on a Deep Learning algorithm. The developed DNN model was based on the deep MLP with linear and non-linear transfer functions, regularization and dropout and a binary sigmoid classification. In their proposed model the hidden layers consisted of 147, neurons. They opined that based on the high level of accuracy they achieved, medical applications of DNN can be relied on and can also be clinically useful in diagnosis of patients with chest pain and other forms of heart disease. Their study analyzed 228 dataset with the deep architecture, splitting the data into training and testing set. Considering their small data set, an architecture of such would have benefited more if a cross validation approach was used instead of a training and split as the model would have properly been validated.

In Priya and Kumar [7], the ability of the kidney to keep our body healthy will be reduced by chronic kidney disease. In their study, the objective is to develop a model that is capable of making prediction of chronic kidney disease from clinical data. In other to achieve their goal, DT and NB classifier methods are used alongside the 10-fold cross validation. Their study revealed that the classifiers used are capable of aiding medical professionals in making timely diagnosis. In their work, they made use of 400 instances and 25 chronic kidney disease attribute. Comparing the models they used and other sophisticated models such as SVM, LR or others would have given a better insight to their work.

Various papers have been focused at using one or a few ML techniques for heart disease prediction on a small data set and considering performance evaluation of their model from a particular angle. However, this research paper focus on six ML techniques that have popularly been used over the years including an ensemble technique. This paper also employs multiple standard performance evaluation methods taking into account dataset. Most papers also split their dataset into just the train and test set neglecting the validation set, while others make use of the cross validation method using either 5 or 10 folds unlike this research study that looks at it from both angles by using multiple folds.

## III. DATA COLLECTION AND RESEARCH DESIGN

### *A. Data Collection*

The data for heart disease prediction in this study was collected from the University of California, Iverine (UCI) online machine learning repository available at https://archive.ics.uci.edu/ml/datasets/Heart+Disease . The UCI dataset contains patient's data concerning heart disease diagnosis that was collected at several locations around the world. It contains 76 attributes including age, resting blood pressure, sex, cholesterol levels, echocardiogram data, exercise habits, and many other. More specifically, the Cleveland database which comprises of 14 attributes are used in this study. With the goal of referring to the presence of heart disease using integer values of 0 (absence) to 1 (presence). The dataset taken include 303 patients made up of 241 males and 62 females. Table 1 gives a description of the 13 parameters utilized in this study.

Table 1: Description of 13 parameters used

| S/N | Attribute | Description | Values |
|---|---|---|---|
| 1 | Age | Age in years | Continuous |
| 2 | Sex | Male or female | 1 = male<br>0 = female |
| 3 | Cp | Chest pain type | 1 = typical type 1<br>2 = typical type angina<br>3 = non-angina pain<br>4 = asymptomatic |
| 4 | Thestbps | Resting blood pressure | Continuous value in mm hg |
| 5 | Chol | Serum cholesterol | Continuous value in mm/dl |
| 6 | Restecg | Resting electrographic results | 0 = normal<br>1 = having_ST_T wave abnormal<br>2 = left ventricular hypertrophy |
| 7 | Fbs | Fasting blood sugar | 1 ≥ 120 mg/dl<br>0 ≤ 120 mg/dl |
| 8 | Thalach | Maximum heart rate achieved | Continuous value |
| 9 | Exang | Exercise including angina | 0 = no<br>1 = yes |
| 10 | Oldpeak | ST depression induced by exercise relative to rest | Continuous value |
| 11 | Slope | Slope of the peak exercise ST segment | 1 = unsloping |

| | | | |
|---|---|---|---|
| | | | 2 = flat<br>3 = down sloping |
| 12 | Ca | Number of major vessels colored by fluoroscopy | 0 – 3 value |
| 13 | Thal | Defect type | 3 = normal<br>6 = fixed<br>7 = reversible defect |
| 14 | Target | Presence of heart disease | 0 = no<br>1 = yes |

The overall goal of machine learning is to create a model with a high capability to generalize well enough. It can be argued that how we split data affects how well our machine learning model is able to generalize. In this study, the dataset collected from UCI is split using cross validation which is a resampling procedure into both 5 and 10 folds. The aim of using cross validation is to create a model that can generalize well enough considering our data size. With cross validation, we need not worry about over fitting or under fitting as the dataset is split into k folds and the model is trained and validated on these folds.

### *B. Data Preprocessing*

Data preprocessing is the process of ensuring that our data is in a right form. This typically involves removing missing values and taking care of other thing within out dataset that many cause either distortion or noise. The collected dataset from UCI was already preprocessed as it contained no missing value. However, our dataset was scaled using the Min-Max normalization technique. Originally, the heart disease dataset consists of various attribute which have values on different scale which may lead to bad scaling. The Min-Max normalization technique typically performs linear transformation on data. It was used to scale the data between 0 and 1 which ensured that features have the same scale. As per Min-Max normalization technique,

$$V' = \frac{V - min}{\max - min} \qquad 1$$

### *C. Algorithms*

i. K-Nearest Neighbors (KNN): It is used for both classification and regression. The KNN algorithm is used mainly in finding values of the factors of heart disease by using K values which makes it possible to find values for the factors of heart disease prediction by making boundaries for each class of attributes. This algorithm simply calculates the distances between training samples, using the Euclidean formula given below.

$$dist(x, y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \qquad 2$$

ii. Support Vector Machine (SVM): It is also a supervised learning algorithm for classification and regression, it classifies heart disease datasets into hyperplanes.

$$f(x) = B_0 + \sum_i (a_i \times (x, x_i)) \qquad 3$$

Where x is the new input vector, $B_0$ is the bias, and $a_i$ is the weight which must be obtained from the training data
SVM has some highly sophisticated properties. SVM consist of a small subset of data points that are extracted from training sample by the learning algorithm itself. SVM provides an analytic approach for determining the optimum size of the feature (hidden) space thereby guaranteeing optimality of the classification task [2]. In [2], the basic idea of SVM can be summarized by the illustration in Figure 1 into:

- Nonlinear mapping of an input vector into a high-dimensional feature space that is hidden from both the input and output;
- Construction of an optimal hyperplane for separating the features discovered.

The number of features constituting the hidden space in Figure 1 is determined by the number of support vectors.

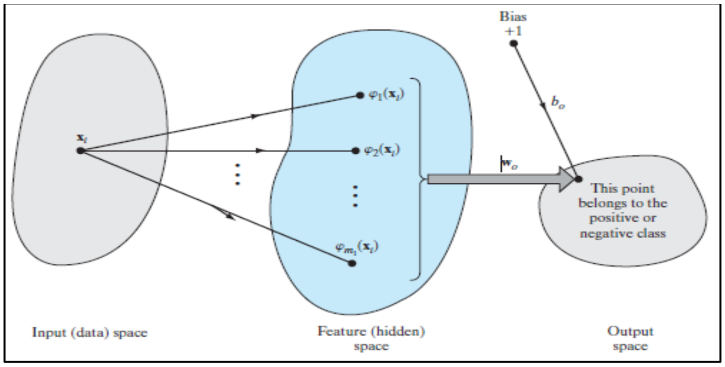
Bias
+1
$b_o$
$\mathbf{w}_o$
$\mathbf{x}_i$
$\varphi_1(\mathbf{x}_i)$
$\varphi_2(\mathbf{x}_i)$
$\varphi_{m_1}(\mathbf{x}_i)$
This point belongs to the positive or negative class
Input (data) space
Feature (hidden) space
Output space

Figure 1: Illustrating the mappings in a support vector machine for pattern classification [2].

Where x is an input vector, $\varphi$(x) is the *feature vector*, $w_o$ is the *weight vector, $b_o$ the bias.*

iii. Naïve Bayes (NB): This algorithm works on the principle of independent features. It represents training data as points in space separated into categories by a clear gap [11]. Naïve Bayes categorize dataset in a distinct manner for heart disease prediction and it can also be used in other real world application such as document classification. Each feature is implicitly assumed to be independent and self-sufficient in some way, thereby contributing individually to the training data point probability of belonging to a particular class.

$$P(c|x) = \frac{P(x|c)P(c)}{P(c)} \qquad 4$$

Where $P(c|x)$ *is the posterior probability of class given predicator,* $P(c)$ *is the prior probability of class, and* $P(x|c)$ *is the likelihood probability of predicator given class*

iv. Random Forest (RF): This algorithm can be used for both classifications on regression problem. This algorithm creates different decision trees for attributes, and corrects over fitting of their training during its process. It is simple to use as it provides flexibility.

v. Logistic Regression (LR): This is one very popular, simple, and efficient algorithm used in classification today. When it is used for predicting one outcome i.e. 0 or 1 it is call binary, when it is a more than one it is called multinomial, and when it is used for a multiple category problem it is known as ordinal. The logistic function is given as:

$$p(x) = \frac{1}{(1+e^{-x})} \qquad 5$$

Where x is the input.

vi. Gradient Boosting (GB): Boosting is an ensemble learning method, it combines multiple algorithms to give a more robust accuracy. Logically, it works by new models predicting the errors of previous models using a gradient decent approach for loss minimization.

### *C. Research Design*

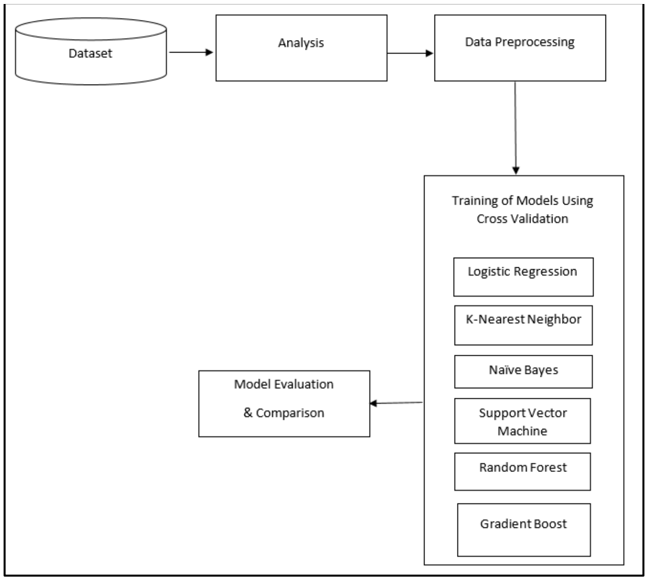

Figure 2: The Proposed System

Figure 2 shows the clear works of the proposed system. The steps taken in Figure 2 is explained as follows:

a. The records of patient are inputted into the system.
b. An analysis is done on the dataset; the analysis helps explains the data.
c. After the analysis, the data set is preprocessed. For this study the dataset was scaled using MinMax Normalization technique.
d. Various algorithms are trained using different k fold cross validation on the scaled dataset.
e. Model evaluation and comparison of results from the different algorithm was carried out using standard performance metrics.

### *D. Performance Evaluation*

Performance evaluation is an assessment of how well our models perform. The following performance evaluation techniques are used to describe how well how model performs:

i. TP (True Positive): It is the number of heart disease occurrence that are classified as true and are actually true.
ii. TN (True Negative): It is the number of heart disease occurrence that are classified as false and are actually false.
iii. FN (False Negative): It is the number of heart disease occurrence that are classified as false and are actually true.
iv. FP (False Positive): It is the number of heart disease occurrence that are classified as true and are actually false.

Some metrics can be calculated and used to further evaluate the models performance. Some of these metrics used in this study are as follows:

i. Sensitivity (Recall/True Positive Rate): It tells the actual positive that were correctly classified as been positive, and it is given as:

$$Sensitivity = \frac{TP}{(TP+FN)} \qquad 6$$

ii. Precision: precision is the ratio of significant instance against the retrieved instance and is given as:

$$Percision = \frac{TP}{(TP+FP)} \qquad 7$$

iii. Accuracy: It is a measure of proportion of the correct prediction that was made by the model, and it given as:

$$Accuracy = \frac{TP+TN}{(TP+FP+FN+TN)} \qquad 8$$

iv. F1 Score: It is a metric that tell use the harmonic mean between our recall and precision. It is given as:

$$F1\ Score = \frac{2*Precision*Recall}{Precision+Recall} \qquad 9$$

## IV. Result

Table 2: Model Performance Using Different Standard K Folds

| **Model** | **Folds** | **Performance Metrics** | | | |
|---|---|---|---|---|---|
| | | **Accuracy (%)** | **Precision (%)** | **Sensitivity/Recall (%)** | **F1 Score (%)** |
| KNN | 5 | 80.530 | 80.614 | 84.849 | 82.630 |
| | 10 | 82.785 | 83.941 | 85.478 | 84.507 |
| LR | 5 | 82.508 | 80.518 | 89.697 | 84.832 |
| | 10 | 82.483 | 81.684 | 88.493 | 84.711 |
| SVM | 5 | 83.180 | 80.919 | 90.909 | 85.524 |
| | 10 | 83.807 | 81.796 | 91.507 | 86.166 |
| NB | 5 | 82.514 | 82.379 | 86.667 | 84.345 |
| | 10 | 80.871 | 81.949 | 83.677 | 82.465 |
| RF | 5 | 82.836 | 82.065 | 87.879 | 84.755 |

| | | | | | |
|---|---|---|---|---|---|
| | 10 | 81.118 | 81.543 | 84.889 | 82.933 |
| GB | 5 | 82.828 | 81.409 | 89.816 | 85.140 |
| | 10 | 79.194 | 79.265 | 84.191 | 81.363 |

Table 2 clearly shows the result achieved by the models using different folds during cross validation. Various experiments were carried out using both 5 and 10 fold, and in both cases, SVM performed better than the rest models as the F1 score shows this. On major advantage of using cross validation over its variant percentage split as used by many is the reduction of over fitting, as cross validation (K fold) splits up the given data set into k number of fold. It starts by using the first fold for testing and the rest for training. It continues in this fashion as it swaps various folds and each time using one fold for test and the rest for training. This ensures that the model generalizes well on the dataset especially on a small dataset as used in this study.

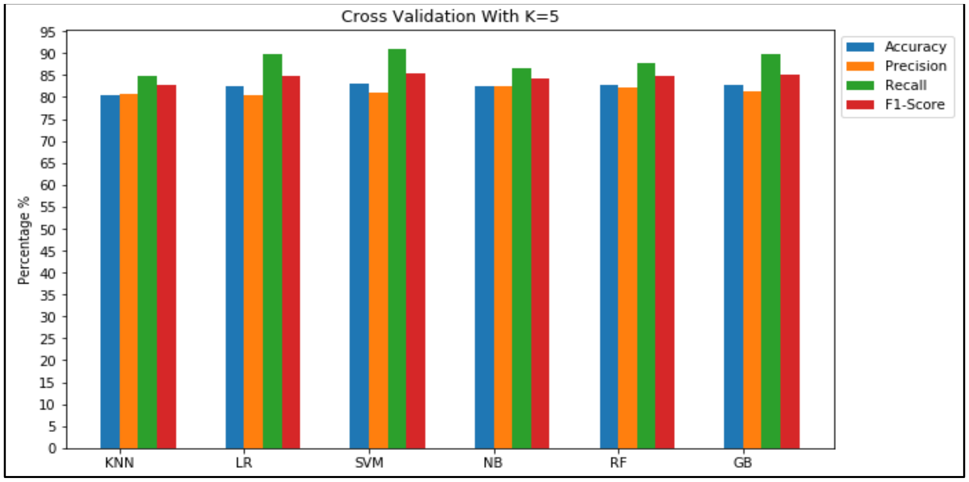

Figure 3: Comparison of models performance with cross validation when k=5

In figure 3, the cross validation is used and k is set to 5. The result from various models show SVM leading in it accuracy, it has a F1 score of 85.524% followed by GB, then RF and KNN performing the least.

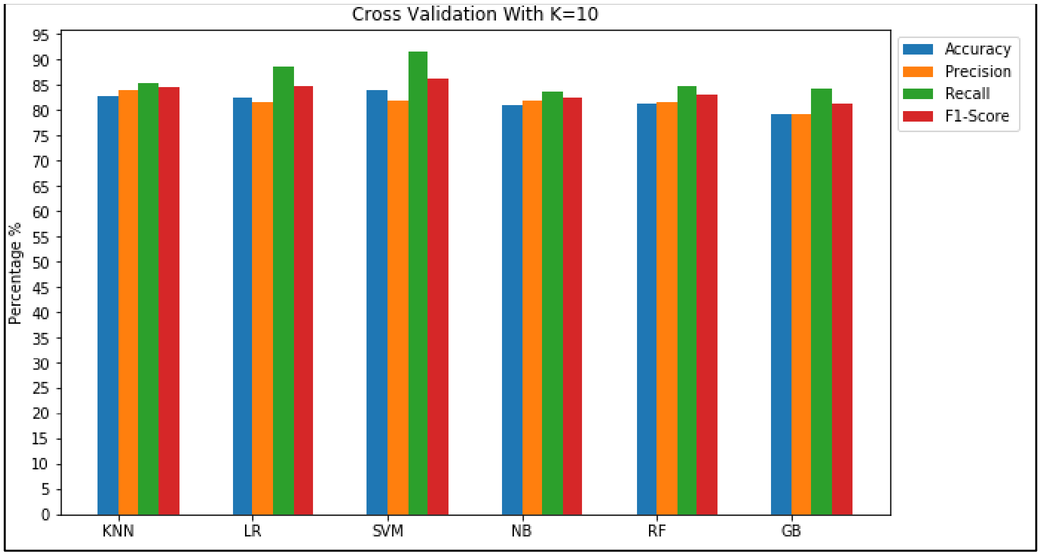

Figure 4: Comparison of models performance with cross validation when k=10

Figure 4 give a graphical representation of the accuracy or various models after applying cross validation when k=10. SVM still maintained the top performance with a F1 score of 86.166%. The performance of RF and GB may have improved if the dataset was larger.

## V. CONCLUSION

In this research study, six (6) popular machine learning model have been implemented for heart disease diagnosis using data from UCI. Different experiments was conducted on these algorithms using K fold cross validation of 5 and 10 folds, after which standard performance metrics were used to evaluate the performance of these models. The objective of this study was to build different machine learning model and compare their performance of predicting for heart disease. From the results obtained from each fold on the models, the sophisticated SVM came out performing the best even out performing Gradient Boosting ensemble method. The experiments showed that the number of folds plays an important role in improving the accuracy of a model. This study can aid researcher and medical professionals in understanding theses algorithms better for heart disease prediction. However, it should be noted that each of the algorithms used can outperform the others under different circumstance.

Future work of this research study can focus more on improving the accuracy of these models by performing hyper parameter tuning, and using larger datasets on other ensemble techniques. It is believed that using optimum parameters on these models can significantly improve their performance.

## ACKNOWLEDGMENT

I would like to thank my undergraduate project supervisor Mr. Efosa I. for the initial stir he gave down the Artificial Intelligence path. I would also like to express my special appreciation to my parents for their support morally and otherwise.

## REFERENCES

[1] Arubpradeep N., and Niranjana G. (2020), Different Machine Learning Models Based Heart Disease Prediction. International Journal of Recent technology and Engineering, 8(6).

[2] Haykin, S. (2010). Neural Networks and Learning Machines. Pearson Education, New Jersey:NJ.

[3] Ufumaka, I. (2020). Michine Learning Approach For Heart Disease Prediction, Research Gate, https://doi.org/10.13140/RG.2.2.21393.86888/3

[4] Kabirirada, S., Kardanmoghaddamb, H., and Afshin, V. (2016). Heart disease Prediction by Using Artificial Neural Networks, International Journal of Computer Science and Information Security, 14 (1).

[5] Kavitha C., Chetan B. A., Ahmed N., Mayi J. S. R., and krishnaveni K. (2020). Disease prediction Evaluation using Machine Learning Gaining Knowledge Strategies. Internationsl Journal For Engineering and Science and Computing, 10(3).

[6] Miao, K. H. and Miao, J. H., (2018). Conary Heart Disease Diagnosis Using Deep Neural Networks, International Journal of Advanced Computer Science and Applications, 9(10), 1-8

[7] Mukherjee, S., and Sharma, A. (2019) Intelligent Heart Disease Prediction using Neural Network. International Journal of Recent Technology and Engineering, 30(5), ISSN: 2277-3878.

[8] Obiri A. D., and Sarku E. (2020). Predicting The Presence of Heart Disease Usung Comparative Data Mining and Machine Learning Algorithms. International Journals of Computer Applications, 176(11).

[9] Patel, J., Upadhyay, T., and Patel, S. (2016). "Heart Disease Prediction Using Machine Learning and Data Mining Technique" International Journal of Computing Science and Communication. vol (7), 129 – 137.

[10] Priya S. S., and Kumar, S. (2018). Chronic Kidney Disease Prediction Using Machine Learning. International Journal of Computer Science and Information Security (IJCSIS), 16(4).

[11] Raghunath, D., Usha, C., Veera, K., and Manoj, V. (2019). PREDICTING HEART DISEASE USING MACHINE LEARNING TECHNIQUES. IRJCS: International Research Journal of Computer Science, 6, 149-153.

[12] Who Health Organization (2016). "Health Topics: Cardiovascular Disease." Retrieved from: https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1 Accessed [8-9-19] .

## AUTHORS

**First Author** – Isreal Ufumaka, BSc., University of Benin and isrealufumaka@gmail.com

**Correspondence Author** – Isreal Ufumaka, isrealufumaka@gmail.com, +2347062974329.