Volume 1, Issue VII, July 2022, **pp. 69-84**

Final peer reviewed 5/7/2022

Online Publication 6/7/2022

Available Online at http://www.ijortacs.com

DEVELOPMENT OF A CARDIOVASCULAR HEART DISEASE CLASSIFICATION MODEL USING NEURAL NETWORK BASED DATA MINING TECHNIQUE

[1]Ezigbo L.I., [2]Okonkwo R.O., [3]Nwobodo L.O

[1,2,3] Enugu State University of Science and Technology

ifyeigbolucy@yahoo.com

Abstract

This paper presents the development of a cardiovascular heart disease classification model using neural network based data mining technique. The aim was to develop a classification model for the detection of cardiovascular heart disease using data mining technique. The research methods are data collection, data extraction, feed forward neural network, and the cardiovascular classification model. The study collected data of 18 heart disease attributes with updated features such as alcohol and kola-nut which is domicile in the African region and has contributed to cardiovascular heart problem in the region. The data was extracted using Best First Search Method and then trained with neural network model. The models were implemented with neural network toolbox in Simulink and evaluated. The result achieved a classification accuracy of 96.51%; Mean square error value of 0.0356810Mu and True positive classification value of 0.959. The result was compared with existing state of the model and a percentage improvement of 2.26% was achieved.

Keywords: Cardiovascular Heart Disease, Neural Network, Data Mining, Classification Model

I. INTRODUCTION

According to Carlos (2004) cardiovascular disease is a kind of heart disease which affects the heart organ, thus resulting to various symptoms such as chest pain, complications among other. Many risk factors have been attributed with this epidermis, which are age, gender, diabetes; high cholesterol, obesity among other and the heart disease type are dependent on the symptoms. Cindy (2008) classified heart disease as congenital, congestive, coronary, pulmonary and rheumatic heart diseases

respectively and all are very dangerous if not detected early and treated.

According to Richard (2009), heart disease remains the most cause of death and stroke in the last decade", with over 1.5million cases recorded in Nigeria alone per annum. This population is most common for babies less than two years and senior citizens from 60 years and above (Google Medical Information, 2021). According to Anitha and Sridevi (2019), in the United States, over 11 million citizens are estimated to have this problem by 2050 and 17.9 million in Europe by 2060.

Various approaches have been employed to tackle this problem ranging from traditional to scientific measures. However despite the effort, the disease still posed major threat to the existence of man, especially those with the risk factors. Therefore it is very important for physicians to detect this epidermis at a very tender stage and prefer measures for early diagnosis.

Today the conventional means to help cardiologist detect heart disease is based on electronic health recorded, however these data are not reliable to help detect these diseases on early stage. Over the year many literatures like (Sarangam et al., 2018; Zeinab, 2017; Gomath et al., 2016; Jaymin et al., 2015; Amato et al., 2015) among others have proposed the use of classification model based on artificial intelligence and data mining techniques for solve this problem of heart disease early detection.

However, from the review and also from discussions with domain expert, it was observed that attributes of heart disease are environmental related and some causes of heart disease in one region may not be the same in other regions. For instance the consumption of Kola nut which is rampart in the African region according to (Agatha et al., 1987) contributes to the rate of heart disease, but not always the case in the European environment. This as a result makes the conventional classification model, not reliable for the detection of heart disease as they lack some vital attributes like kola nut features content in the body to help predict the risk of heart abnormally. To this end, there is need for a system which considers this attributes alongside other heart disease related attributes, and then train with a data mining technique to predict heart diseases in time series.

Data Mining has attracted great attention from various fields due to wide and large data present in these fields. It is a means of

extracting meaningful and useful information from the large databases and is made up of series of algorithms for solving regression and classification based pattern recognition problems. This technique will be adopted and used to predict early heart disease issues in patients in this research.

II. LITERATURE REVIEW

Sarangam et al. (2018) used decision table and Naïve Bayesian technique for the prediction of heart disease. The result obtained is 83.70%, the author in another research used random forest algorithm for the same cause and achieved a detection accuracy of 81.85%. However despite the success achieved in the various algorithms used, there is need for improved performance. Zeinab (2017) used ANN with genetic algorithm for the detection and prediction of heart disease using 12 variables as input and achieved an accuracy of 93.85%, however despite the success achieved, the study was limited by the dataset used and kolanut was not considered.

Gomath et al. (2016) presented a paper which used Naives Bayesian, ANN and decision list for the detection and prediction of heart disease respectively. They study was designed and implemented for testing, and the result showed a respective prediction accuracies of 79%, 77% and 76% respectively. However despite the success achieved, there is need for improvement. Jayamin et al. (2016) detected and predicted heart disease using logistic model tree and achieved an accuracy of 55.8%. The result however still needs to be improved with a better accuracy.

Amator (2015) presented a paper on the ANN based device as a physical aid for physicians to analyzed body and detect heart beat rate. However the accuracy of the system is 89% and can be improved. Chaurasia et al. (2017) presented a paper which developed a knowledge based expert system using data mining technique to enable doctors and physicians to forecast and predict heart disease disorder early, however the success of the system was limited by the training dataset.

III. METHODLOGY

The methodology used for the development of the new system was guided by the Dynamic Systems Development Model (DSDM) approach which accommodates the use of structural and mathematical modeling inspired techniques for the development of the cardiovascular disease detection system. The methods used were data collection from physiobank (physiobank, 2021) repository as the primary source of data collection, while Parklane hospital, Enugu State, Nigeria is the secondary source of data collection which provided the new heart disease features of kolanut identified. The sample size of data collected is 700MB of heart beats recorded between the frequency of 0-128Hz from Parklane and 20 gigabyte of recorded heart disease signal from physiobank. These were organized into classes of standard and clinical attributes. The data was integrated to develop the data model for the detection of cardiovascular disease. The dataset were extracted using Best First Search extraction method in (Han and Kamber, 2006) to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes and then trained with feed forward neural network to develop the classification model needed for the detection of cardiovascular heart disease.

IV. SYSTEM DESIGN

The system design used the table 1 to present the problem formulation as shown in the attributes of the cardiovascular data collected. The data model was presented in figure 1 using class diagram which showed how the various attributes of the data collection were classified as standard or clinical symptoms.

Table 1: Data Table

Attributes ID	Attributes	Data description	Data type
1	Age	Number of patient years	Numeric
2	Sex	Gender of patient	Numeric
3	Weight	Body weight of the patient	Numeric
4	Cholesterol	The level of cholesterol in the patient body	Numeric
5	Blood	The level of sugar in the patient’s blood	Binary
6	Blood	Level of pressure in the blood flow	Numeric

7	Smoke	Constituents level of smoke in the blood level	Numeric
8	Alcohol level	Alcohol level in the blood	Numeric
9	Exercise	Exercise induced angina	Numeric
10	Caffeine	Level of caffeine in blood	Numeric
11	Kolanut	Constituents of kolanut level in the blood	Varchar
12	Cough	How often cough occur	Binary
13	Palpitation	How often palpitation occur	Binary
14	Chest pain	If chest pain occur	Binary
15	Breathlessness	If patient struggle to breath	Binary
16	Legs swelling	If the patient leg is swollen	Varchar
17	Orthopnoea	Inability of patient to lie down	Varchar
18	Cough	How often cough occur	Binary

Data model

A data model is a conceptual representation of the data structures that are required by a database. The database was divided into four classes which are based on standard symptoms attributes, clinical symptoms attributes, sickness attributes and knowledge based attributes. This was developed using a class diagram as in figure 1;

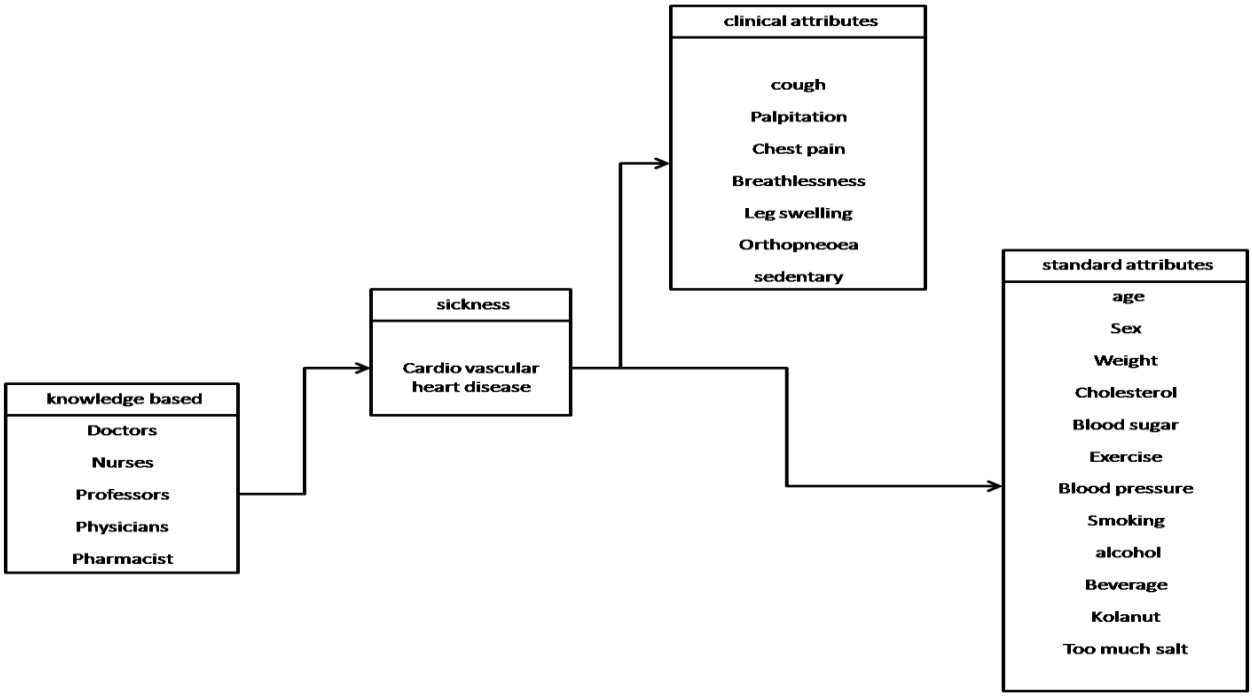

Figure 1: dataset model entity diagram

The data model in figure 1 was used to develop the improved dataset containing new attributes like kolanut and alcoholic features which were omitted in the conventional data (see physiobank) collected and also existing models. The data model classified all the cardiovascular disease attributes collected such as Age, Weight/BMI, Cholesterol, Blood sugar, Blood pressure, Exercise/sedentary, Smoking/no of sticks, Alcohol/units/duration, Beverages/Coffee with caffeine. Clinical symptoms are: Cough, Palpitation, Chest pain, Breathlessness, Legs swelling, inability to lie down because of breathlessness into one dataset of two classes.

Data extraction Model

The data extraction model was developed based on Best First Search Method in Han and Kamber (2006) which extracted features such as Blood Pressure, Cholesterol level, Blood sugar, Heart rate, Weight, Age, Chest pain, Palpitation and Leg swelling from the data model as shown in equation 1;

$$Fx = \frac{\sum_{i=1}^{total\ feature\ vectors} cadiological\ attributes\ (n)^2}{total\ feature\ vectors} \quad 1.0$$

Where i is the total feature vectors, n is any of the selected input symptoms, Fx is the feature vectors.

Model of the data mining based classification model

The data mining algorithm used is the Feed Forward Neural Network (FFNN) adopted from Samuel and Oluwarotimi (2017) and reconfigured to train the data collected. The FFNN was made of interconnected neurons which has weight and bias function. The neurons were activated with tanh activation function and then trained with back propagation algorithm to learn the cardiovascular features and generate the classification reference model. The model of the neural network workflow to generate the cardiovascular classification model was presented in figure 2;

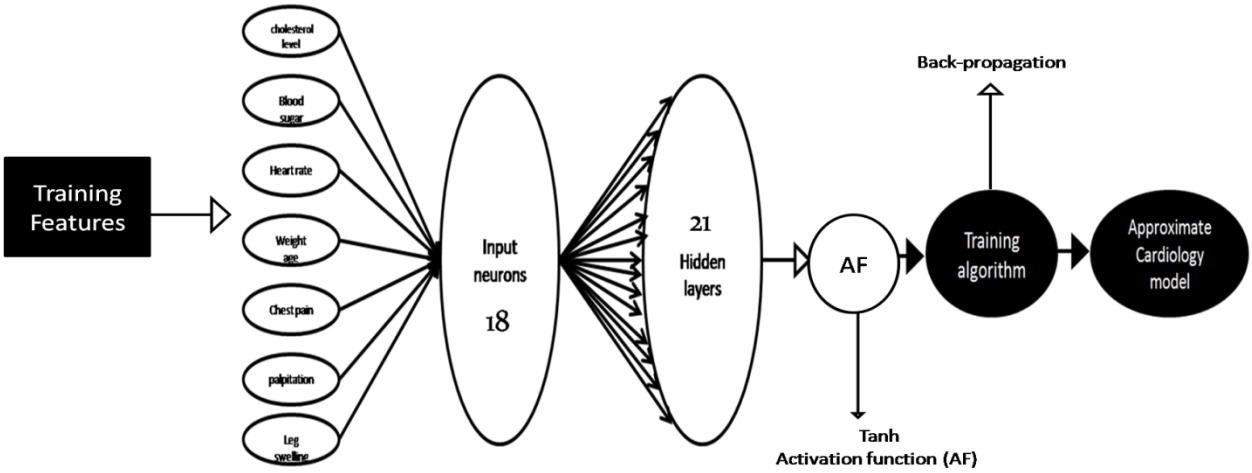

Figure 2: The FFNN modeling diagram

The architectural model showed how the FFNN was developed. The number of input was inspired from the cardiovascular attributes collected which are 18, the hidden layer was specified to be 27 to enable fast computation of the neurons, activation function and training algorithm used were all represented in the model and then trained the cardiovascular features to generate the classification model for the diagnosis of cardiology. Other parameters inclusive used for the development of the FFNN are reported in table 2 while the logical flow chart was presented in figure 3;

Table 2: Neural network configuration parameters

Parameters	Values
Training epochs	19
Size of hidden layers	27
Controller training segments	30
No. delayed reference input	2
Maximum feature output	3.1
Maximum feature input	21
Number of non hidden layers	2
Maximum interval per sec	2
No. delayed output	2
No. delayed feature output	2
Minimum reference value	-0.7
Maximum reference value	0.7
Number of input	18

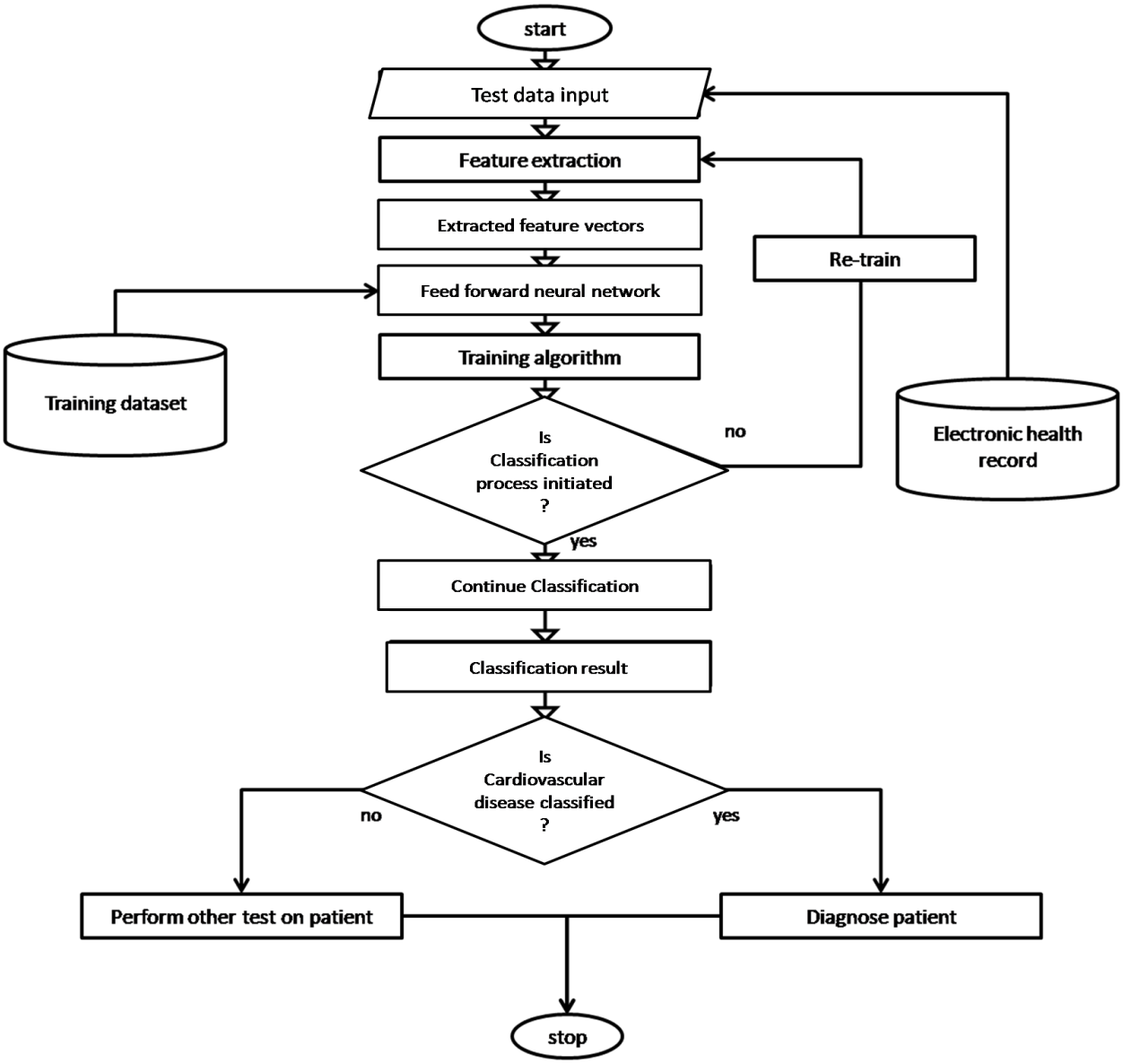

Figure 3: System flow chart

The system flow chart in figure 3 presented the logical data flow of the system, showing FFNN trained with the cardiovascular training data to generate the classification model was used to classify heart disease. When input from the patient electronic health recorded are collected, the features are extracted based on the Best First Search Method to identify key cardiovascular attributes for classification with the FFNN developed classification model. If the classification result is cardiovascular disease with the input feature vectors, then the patient is recommended for diagnosis, else the patients is recommended for other test.

V. IMPLEMENTATION

The system was implemented using neural network toolbox in Matlab. This was achieved using the models to configure the neural network application software in the pattern recognition tool and then upload the cardiovascular features when were automatically extracted by the tool using the Best First Search feature extraction model in equation 1. The number of hidden layers was input and back propagation training algorithm selected for the training process. Before the training the tool automatically divided the features in the ratio of 70:15:15 for training, test and validation set, then training commences with test and validation process respectively at each epoch. This continued until the neurons learn the cardiovascular features and generate model needed for the detection of cardiovascular heart disease.

VI. PERFORMANCE EVALUATION MODELS AND RESULTS

To test the performance of these models, the sensitivity (true positive rate) and specificity (false positive rate) measures was used respectively. Sensitivity is also referred to as the proportion of correctly classified vessel, while specificity is the false positive rate is the proportion of correctly detecting vessel. These models are presented in equation 2 and 3 respectively.

$$True\ \ Positive\ Rate\ (TPR) = \frac{TP}{TP+FN} \quad 2.0$$

$$False\ positive\ Rate\ (FPR) = \frac{TN}{TN+FP} \quad 3.0$$

Where TP is true positive, FN is false negative, TN is true negative and FP is false positive. The overall accuracy of a classifier is estimated by dividing the total correctly classified positives and negatives by the total number of samples in the dataset. The accuracy of the classifiers was measured using the relationship between equation 2 and 3 as shown in the model of equation 4

$$Accuracy\ (ACC) = \frac{TP+TN}{TP+TN+FP+FN} \quad 4.0$$

The validation of the model was presented using the validation model in equation 5;

$$\text{CVA} = \frac{1}{10} \sum_{1}^{10} \text{Ai} \quad 5.0$$

Where CVA stands for Cross Validation Accuracy, A is the accuracy measure for each fold. During the training process, Mean Square Error (MSE) and Receiver Operator Characteristics (ROC) curve were selected as evaluation tool to measure the training error and cardiovascular detection performance. The MSE analyzer was presented in figure 4;

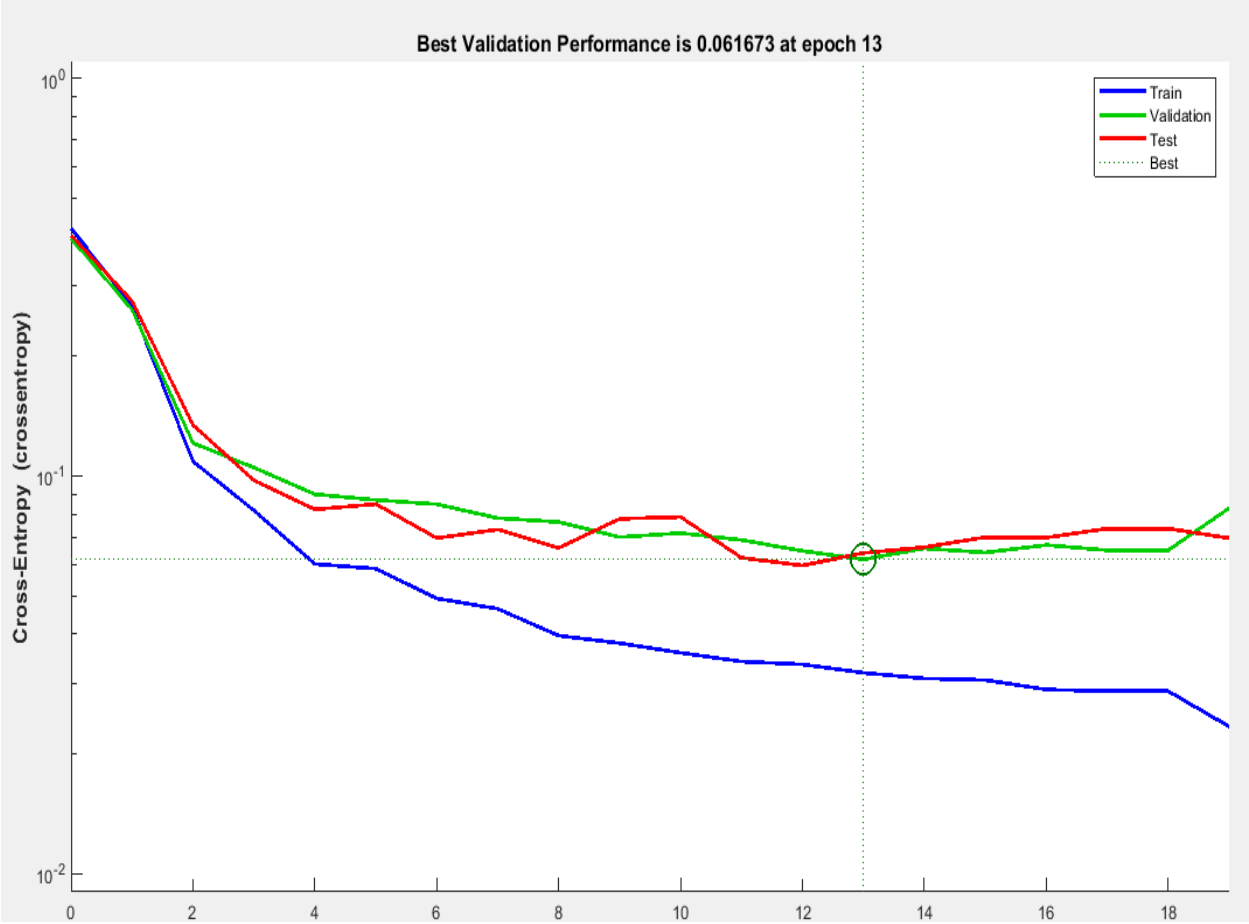

Figure 4: MSE result of the model

The result in figure 4 shows the MSE performance of the classification model generated. The aim of this result was to ensure that minimum error occurred during the training process with a target of zero error. From the result, it was observed that the MSE value is 0.061673 which is good as it is approximately zero at epoch 13. The implication of this result is that the neurons correctly learn the cardiovascular data and produced constant validation output at epoch 13.

The next result showed the cardiovascular detection performance of the neural network classification model using Receiver Operator Characteristics (ROC) curve which used the equation 2 and 3 to measure the true classification performance of the system. The aim of this ROC was to achieved a TPR of 1 or a FPR of zero. These values implied optimal detection of the cardiovascular problem from the test set and the ROC is presented in figure 5;

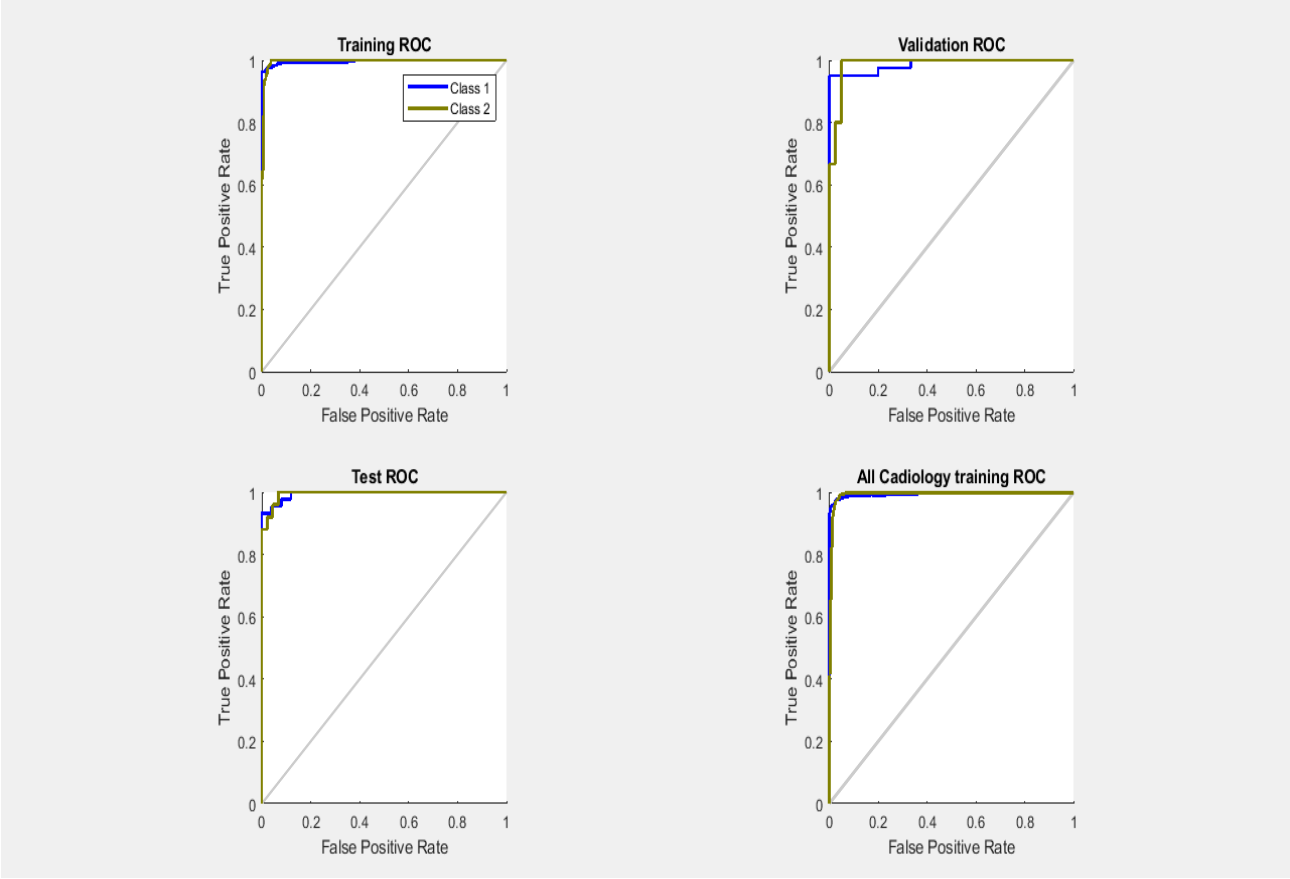

Figure 5: The ROC analyzer

The figure 5 presented the analysis of the classification model developed with neural network. The result was analyzed using the table 3;

Table 3: ROC analysis

Data	False positive rate (FPR)	True positive rate (TPR)
Training	0.23	0.977
Testing	0.57	0.943
Validation	0.29	0.971
All	0.27	0.973
Overall training performance	0.27	0.973

From the table 3, it was observed that both the training, test and validation sets all are very close to the targeted values of 1. The overall performance showed ROC values of 0.973 which is good as it implied correct

classification result for cardiovascular attributes.

To validate the system performance, tenfold validation technique in equation 5 was used. This process iteratively trains the multi dataset in tenfold and the average score recorded as the achieved result. This training process will consider the root mean square error, accuracy and true positive rate, presented in table 4;

Table 4: System Validation performance

Iteration	ACC (%)	TPR	MSE (Mu)
1	98.1	0.991	0.0095612
2	95.1	0.958	0.0773790
3	96.4	0.979	0.0428200
4	95.4	0.924	0.0264560
5	96.6	0.977	0.0156900
6	96.9	0.943	0.0739220
7	95.7	0.989	0.0447392
8	95.9	0.944	0.0629877
9	96.3	0.924	0.0457021
10	98.7	0.970	0.0705542
Average	96.51	0.959	0.0356810

From the validation result the average TPR of the classification model is 0.959 which is very good as it is very close to the ideal value which is 1. The MSE result is also 0.0356810Mu which is approximately zero, indicating very negligible error score which is also good. The average accuracy recorded is 96.51% which is very good too.

VII. COMPARATIVE ANALYSIS

This section compared the performance of the new classification model with some of the conventional sate of the art models and the results presented in figure 6;

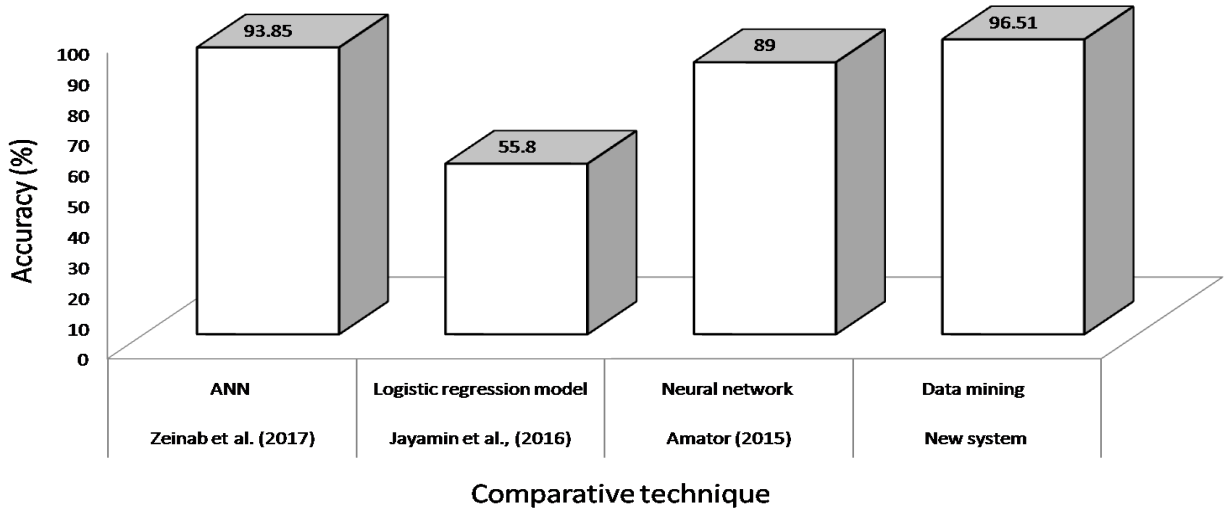

Figure 6: Comparative Analysis

The figure 6 presented a comparative analysis of the new system developed with the existing state of the art models. The result showed that the new system was the best in term of classification accuracy with a percentage improvement of 2.66% compared with the best ANN based approach in Zeinab et al. (2017).

VIII. CONCLUSION

Heart diseases have been one of the major challenging epidermises which the classification is significant in medical field. Overtime, the need to achieve correct classification result has been one of the major aims of researchers, doctors and experts. This is to ensure that a patient with the symptom is not classified negative or vice versa. However to achieve this aim, real time update on the training dataset is required for the learning of the artificial intelligent agent. The challenges of the presents conventional system is that most of the data attributes used for the data model are not up to date. Some of the symptoms and cases of heart diseases are not included in the existing dataset as identified in the literature review. The implication of this process is that there is high chance of poor classification accuracy. This work has successfully proposed and develops a clinical decision system for the real time detection of heart disease using artificial intelligence. The work was validated using the necessary performance evaluation metrics, and the result shows a very précised detection performance

IX. CONTRIBUTION TO KNOWLEDGE

- A heart disease classification model was developed with improved data model using data mining technique.

X. REFERENCE

Agatha M., Brechenrige C., Soyemi E. (1987) "Some preliminary observation in the effect of kola nut in the cardiovascular system" Niger Med. [J]. PMID: 753053; Vol 8 (8):pp 501-505

Amator Filippo (2015) "Artificial neural networks in medical diagnosis", pp. 47-58

Anitha. S., and Sridevi, N. (2019). Heart Disease Prediction Using Data Mining Techniques. Journal of Analysis and Computation. 8(2), 48-55.

Carlos, O. (2004), *Improving Heart Disease Prediction Using Constrained Association Rules*, Seminar Presentation at University of Tokyo

Chaurasia, Vikas, and Saurabh Pal (2017), "Data mining approach to detect heart diseases",

Cindy H. (2008). *5 Common Types of Heart Disease*. EzineArticles.com. Available at http://ezinearticles.com/?5-Common-Types-of-Heart-Disease&id=1073496 (March 03 January 2011)

Gomath, Shanmugapriyaa (2016), "Heart Disease Prediction Using Data Mining Classification", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol.4, Issue 2.

Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques.* Second Edition, Morgan Kaufmann Publishers, San Francisco

Jayamin P., Tejal U., Samir P. (2016) "Heart disease prediction using machine learning and data mining" Vol 1: Pp129-137.

Physio Bank, (2021), Retrieved from https://physionet.org/physiobank/database/#ecg.

Richard N. Fogoros (2009). *Key Symptoms of Heart Disease*.About.com Health's Disease and Condition. http://heartdisease.about.com/b/2010/12/03/the-key-symptoms-of-heart-disease.htm

Samuel T., Oluwarotimi W. (2017) "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction", *Expert Systems with Applications*, Vol. 68, pp. 163-172.

Sarangam Kodati, and Dr. R Vivekanandam (2018)"A Comparative Study on Open Source Data Mining Tool for Heart Disease" , International Journal of Innovations & Advancement in Computer Science, Vol. 7, Issue 3.

Zeinab Arabasadi (2017), " Computer aided decision making for heart disease detection using hybrid neural network Genetic algorithm" , Computer Methods and Programs in Biomedicine-ELSEVIER, Vol. 141, pp.19- 26.