DEEP LEARNING BASED HEALTH DATA ANALYSIS TO PREDICT CARDIOVASCULAR DISEASE

A.Grace suji[1], R.P.Anto kumar [2]

[1] *M.E. Student, Department of Computer Science & Engineering, St.Xavier's Catholic College Of Engineering, Nagercoil, Tamil Nadu, India.*

[2] *Professor, Department of Computer Science & Engineering, St.Xavier's Catholic College Of Engineering, Nagercoil, Tamil Nadu, India.*

ABSTRACT

One of the most hazardous disease which causes death to humans all over the world since 15 years is the heart disease. The patient can prevent the heart disease if it is predicted at an earlier stage, this will be also useful for the medical practitioners to understand the cause of heart attack and can avoid before its actual occurrence. The better clarification for the issue is determining the patient's health issue and analyzing the health problems in prospect so that the specialists can begin the treatment to provide better outcome. It is better than stand-in at later when the patient is in danger and prediction of the heart disease is difficult to analyze and is widely researched area. In this heart disease prediction, several research and advance technology are verified. This paper is suggested to provide the detailed description about the advantage of the techniques and prediction model developed for heart disease. To eliminate the problem in heart disease (HD), doctors and many scientists have suggested to utilize intelligent methods i.e., HD prediction problem is solved using artificial intelligence. The fundamental and major factor which causes death is mis-prediction of disease. An intelligent system is designed to prevent the mis-prediction of heart disease. This paper is a simulation which provides better performance than the traditional diagnostic method. In this paper Artificial Neural Network (ANN), which based on the conventional neural networks is suggested for predicting heart disease. The proposed system is a simulation which provides the better performance than the traditional diagnostic method that is used to diagnose the heart disease. The exploitation and exploration of the binary and multi-class heart disease prediction is obtained by Ant Colony Optimization technique. The supervised learning algorithm in neural network provides better performance in classification task.

Keywords: *Heart diseases; ANN; Supervised learning algorithm; Deep neural network; Disease Prediction; Fast Correlation*

1. INTRODUCTION

Heart is the main organ of the cardiovascular system. The heart is about the size of a fist i.e.it weighs about 250-350 grams. The heart beats around 2.5 billion times during a person's lifespan of 66-68 years. The W.H.O states that in 2008, 17.3 million people died because of heart attacks[1]. In India, the leading cause of death cases is caused

by heart attacks, 2.25 million deaths had occurred in 2010 alone. More number of deaths is occurred due to heart disease when comparing to other diseases[2].ANN has been utilized for medical determination; investigation and understanding the pictures and flags to empowering the results have developed. In particular field of medical result forecast, numerous works demonstrate the helpfulness of neural organizations in wellbeing. Fraser and partners, for instance, researched the feasibility of utilizing a particular sort of neural organization (outspread premise work) for the conclusion of myocardial interact Artificial Neural Network (ANN) is broadly applied as a technique for tackling various choices displaying problems. The multilayer insight is forward ANN technique which is utilized broadly for arranging the various issues. An ANN is re-enactment of cerebrum of human. It is the directed strategy utilized for non-straight characterization of coronary heart disease which is a significant pestilence in India and Andhra Pradesh is also in danger of coronary heart disease.

Identifying the heart disease can be very difficult as there are some risk factors such as diabetes, abnormal pulse, high blood pressure and cholesterol [2]. Severity of heart diseases can be found out by some methods. They are K-Nearest Neighbor Algorithm (ANN), Decision Trees (DT), Naïve Bayes(NB), and Genetic Algorithm(GA)[3]. The heart disease should be handled carefully because the nature of the disease is very complex. Metabolic syndromes can be discovered with the help of data mining and neural networks. For the investigation and prediction of heart disease, mining of data plays a major role. Events of accuracy can be predicted by using decision trees. For predicting the heart disease, various methods are used to abstract knowledge by using some methods of Data Mining.

Many pathological conditions of the heart disease can be found out by certain reflected signals. They are electronic and sound signals. When the oxygen and blood supply gets reduced, heart disease occurs. Various fields like Machine Learning, Data Mining, Artificial Intelligence, By allowing knowledge driven decisions, future trends can be predicted by Data Mining[4]. A large set of data which is complex and massive is used for predicting the heart disease. Efficient techniques such as Diverse Data Mining Techniques and Suitable Data Mining techniques are used by the experts for the prediction of heart disease.

The rest of the paper is represented as follows: Related works is explained in Section II, Section III includes problem statement. Section IV describes the proposed methodology. The results and discussion of the proposed method is included in Section V. Finally Section V explains the conclusion of the proposed method.

2. RELATED WORKS

PraanavMotorwar et al.[5]Suggested the machine learning frameworks which has five algorithms such as Support Vector Machine, Naïve Bayes, Random Forest, Logistic Model and Hoeffding Decision .In this paper, the accuracy of heart related diseases are predicted.The final accuracy of 93.44% for Gaussian NB, 16%for Support Vector Machine,95.08% for Random Forest, 81.24% ,for Hoeffding Tree, 80.69% for Logistic Model Tree. But it does not consider more input features for analyzing the outcome.

A model is proposed using Deep Neural Network and χ_2 statistical model for the prediction of risk profile of patients[4].The model is trained using 242 instances and tested using the remaining 61 instances. Optimized configuration search is accomplished using the grid search algorithm. For optimizing the weight, 80-20 holdout validation was used. The model predicts the presence or absence of heart disease. This model showed great results on testing and training data. By including the data of the patient, the model's accuracy can be increased.

In [7],For heart disease prediction, six techniques of machine learning was used. Tenfold cross validation is used in these methods and various performance measures had been evaluated. The accuracy of these techniques are 83% for Naïve Bayes, 77% for Classification Tree, 80% for KNN, 85% for Logistic Regression, 82% for SVM, 84 % ANN. Using logistic regression, the highest classification accuracy of 85%, sensitivity of 89% and specificity of 81% is achieved.

Heart disease can be predicted by the physician with the help of a good algorithm. In[8], a method called Cardio Help is described which predict the probability of the heart disease by using convolutional neural network(CNN). This method has a accuracy of 97%.Temporal data modeling is used in the proposed method by using CNN for the heart disease prediction. The results show that the proposed methods established greater results than the other methods in performance evaluation metrics. This paper contains parameters of cardiac test and human habits

Rubini.P .E et al developed an application which can predict the vulnerability of heart diseases. Here, a modified analysis of techniques of Machine Learning like, Logistic Regression Random Forest (RF), Support Vector Machine (SVM), and Naïve Bayes are performed. Through this analysis, Random Forest is used in the proposed system in the heart disease prediction as it is considered to give more accuracy and reliable algorithm. It is found out to be more reliable in finding the percentage of the heart disease prediction by acknowledging the relation between the heart disease and diabetes[9].

In[10], two phases was proposed for the identification and prediction of heart disease. They are Classification Stage and Feature Selection. Patient's dataset is included in Feature Selection. Classification algorithms are performed by Grey Wolf Optimization, Ant Colony optimization combined with Particle Swarm Optimization. Optimum solutions can be found out by ASO for finding good paths.

3. PROBLEM STATEMENT

In the previous research studies, Machine Learning Techniques are used for classifying and predicting. The studies are based on the specific effects of the machine learning methods and not on the optimization techniques. For the optimized organization of Machine Learning few hybrid optimization methods are used. Particle Swarm Optimization and Ant Colony Optimization are specified with Machine Learning Techniques.

In the proposed work, chi-square Based Feature Selection is used. When all the continuous attributes are discretized, in between the original attributes, attributes are selected that are relevant to mining. Feature Selection provides efficiency in decreasing dimension, removal of unwanted data, increasing result of proposed technique, and to improve the efficiency of learning and feature-selection is used in the pre-processing step of Machine Learning Technique. Artificial Neural Network is used to select the relevant attributes of dataset and is correlated in the second step. Foremost subset characteristics are selected by the characteristic selection and it improves the efficiency of classification. For Heart disease prediction, classification method is applied in the third step. Classification efficiency is used to understand and estimate the performance characteristic of selection methods.

4. PROPOSED METHODOLOGY

The main objective is to use the Artificial Network method for the prediction of heart disease. Out of 300 datasets, 297 datasets were selected from the heart disease dataset for observation. There are two important sections in the proposed system. In the first part, the risk features that determines the heart disease is found because there will be existing risk factors in the dataset. For every feature, the p-value provides the substantial codes. For the testing and training, the dataset is then divided. After the dataset is trained it is then introduced in the neural network. In figure 1, the structure of the proposed system is shown.

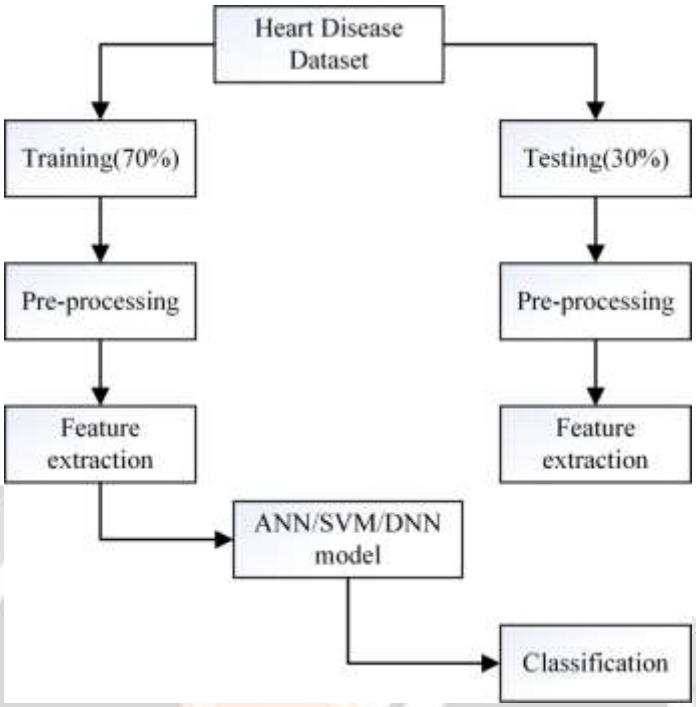

Figure 1:**Proposed Architecture Scheme**

4.1 Data Collection

In data collection, the UCI machine learning is responsible. The UCI machine learning repository contains datasets that are collected from different domains[11].The community of machine learning uses these data.The repository had been directly used by several researchers and academic passes. This repository was developed by David Aha and his fellow students. The Cleveland Clinic Foundation provided a data which is used for the determination of this study.

4.2 Pre-Processing

The data is transformed before they load into the algorithm and this process is called data processing. For processing, the raw data collected from the environment is not suitable. So, the dataset is pre-processed to make it understandable. There are few steps in the proposed system which includes pre-processing steps such as removing duplicate data, empty arrays and null values. The first step in creating a Machine Learning model is the data processing method which initiates the process. Real world data is usually incomplete; also it lacks some attributes, and is in accurate and inconsistent. Data processing helps in cleaning, formatting and organizing the raw data, and is then used for Machine Learning models[12].It is important to identify and correct the missing values in the data-processing, if not it can cause faulty conclusions, inaccuracy and can draw inferences from the data. Two ways are there to find out the missing data.

4.2.1 Discarding the corresponding row

If 75% of the values are missed, then that specific rows and columns will be discarded. There should be no addition of bias after deleting the data.

4.2.2 Calculation of Mean

Mean calculation is used for having the numeric data for calculating the measures of the central tendency of a specific row or column which has a missed value and this method will be replacing the results for the missed value. Variance to the dataset can be added in this method and if there's any loss in the dataset, it can be efficiently negated. This method gives good results than the first method. There is another way for approximating and is done through the deviation of neighboring data and this works well in linear data.

4.3 Feature Extraction

The pre handling stage called Highlight Subset Selection is performed by Artificial Intelligence (AI). It successfully diminishes the dimensionality and removes unimportant information and enlarges the learning precision. Highlight Subset Selection solves the issues of highlights, which predicts the discussion. Highlights are ostensible, consistent, and distinct.

Covering and inserted models are included in the highlight choice[13].There won't be any learning calculation in the channel mode and it completely relies on the investigation of attributes of information and to assess highlights. To recognize the important element, the covering technique uses après decided learning calculation and it uses it to provide highlights in assessment stage. Inserted model includes determination for the model preparation measure. The clinical system's information is large. The achievement of information mining is influenced by numerous components. If the information becomes unessential or excess, trouble occurs in the information revelation during the preparation stage.

Clinical determination should be done precisely and proficiently. Exact outcomes are provided when Highlight Subset Selection is applied with the Clinical Information Mining. Infection from few variables promotes bogus suspicions when connected with whimsical impacts. Highlight subset determination is applied in clinical information.

4.4Classification using ANN

For predicting the results, an algorithm classification called supervised learning method is used[14]. A method is suggested by the proposed technique for heart disease by utilizing the classification algorithms and increasing its efficiency by using the advantages of the classifiers[15]. Heart disease data collection is done on two methods, they are testing and training. The training dataset accomplishes specific classifiers. For testing the accuracy, the test dataset is used.

4.4.1 Classification steps in ANN

Reduction occurs after the filtration of irrelevant data[16]. The classifier then gets the relevant information. In the proposed method, a classification algorithm called Artificial Neural Network is used, which performs the training, classification, and testing operation.

There are three layers in the ANN. They are hidden layer, output layer and the input layer. The attributes of the input layer is shown as x1, x2, and x3. The connection weight of each attributes is represented as w1, w2, w3 respectively. Two functions are performed in this layer named as Summation function and Activation function. Summation function is done by the multiplying each attributes with their weights. Activation function gives correct results for producing the output.

Summation function is represented in equation 1:

$$\xi = \sum W_I \times X_I \quad (1)$$

Where,

χ_i= Input attributes

Y_I = Input attributes of the weights.

Activation function is represented in equation 2:

$$Y=(\xi) \tag{2}$$

The attribute values are produced by the output.Heart disease prediction is done based on these values.

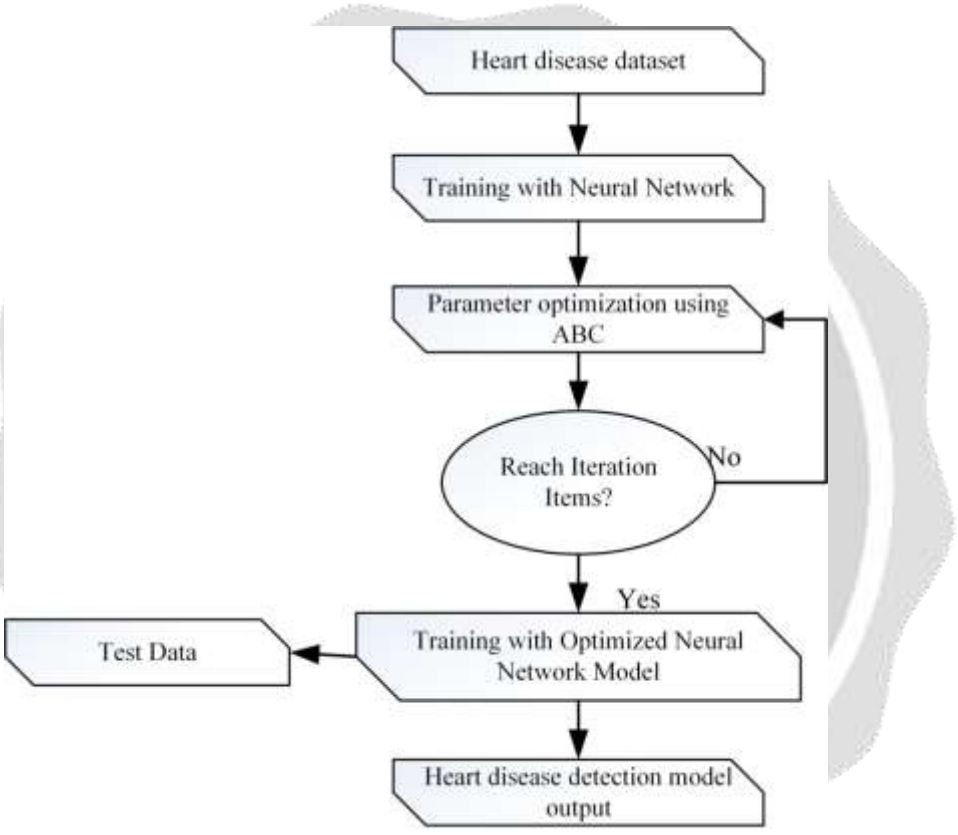

Figure 2:**Data Flow Diagram**

5. RESULTS AND DISCUSSION

The proposed paper contains 85% data in training and 15% data in testing. The performance of the predictable DNN is increased in this technique and from the previous work; the DNN uses the full features by using the search methods of previous work.

Table 1: **Descriptive analysis of different attributes**

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

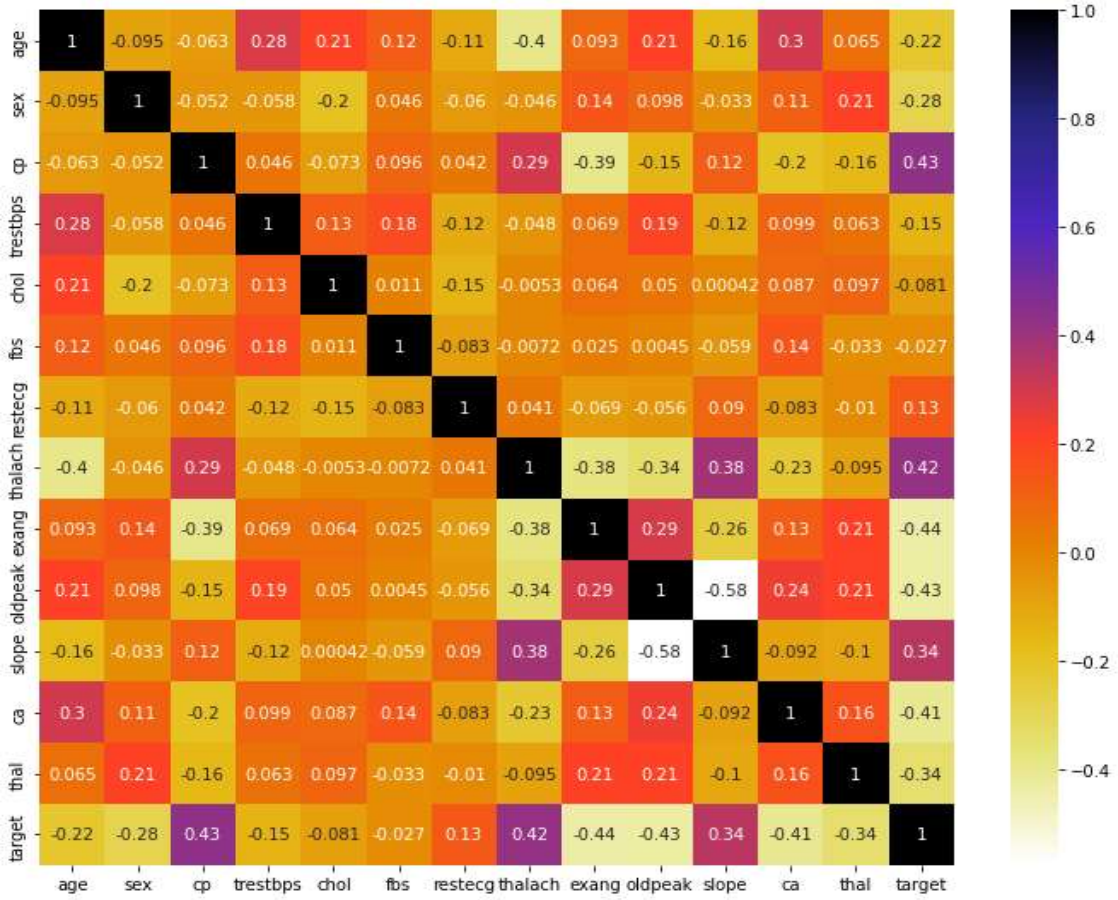

Figure 3: **Comparative analysis of clinical records**.

In figure 3, the dataset has 14 attributes, 8 categorical attributes and 6 numerical attributes are present on the dataset. Feature attributes and their values are mentioned in the table. TRESTB PS is represented as the persons Resting Blood pressure. FBS measures the blood sugar. CHOL is defined as the level of the cholesterol. RESTECG is used for measuring the results of the resting electrocardiographs. THALACH indicates the person's maximum heart rate for measuring the results of the resting electro cardiographs. EXANG means the inducing of Angina by exercise. A ST depression which is induced by the Angina by some rest exercises is known as OLD PEAK. SLOPE

indicates the ST segment's slope. CA indicates the calcium in the heart disease. THAL is the scanning of the Thalassemia disease. Attributes of the class is represented in the Target.

The data is searched by the program and the dataset is divided into 4 types, they are Disease Type 1, Disease Type 2, Disease Type 3 and Disease Type 4. Figure 3 represents the comparative analysis of total records. Figure 4 and 5 shows the plot for loss and the plot for accuracy.

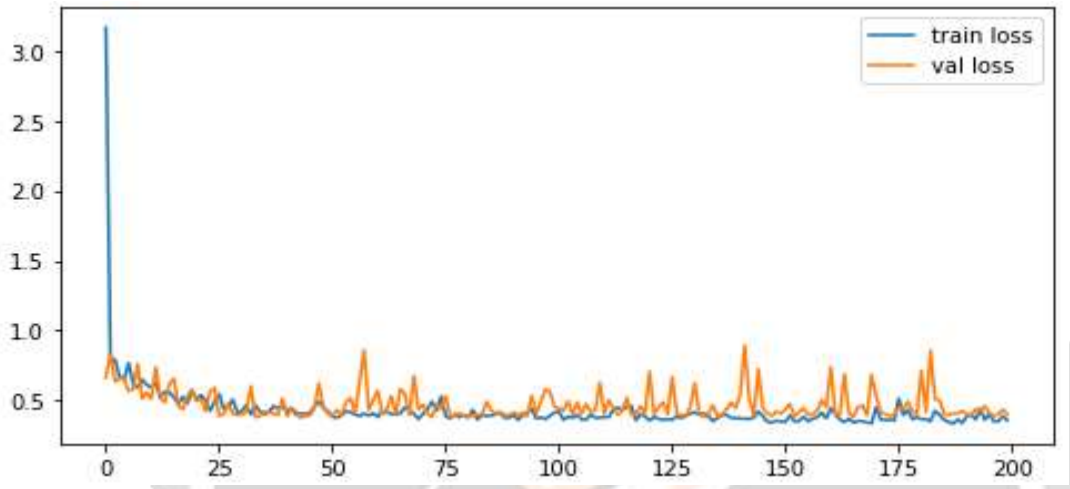

Figure 4: **Plot for loss**

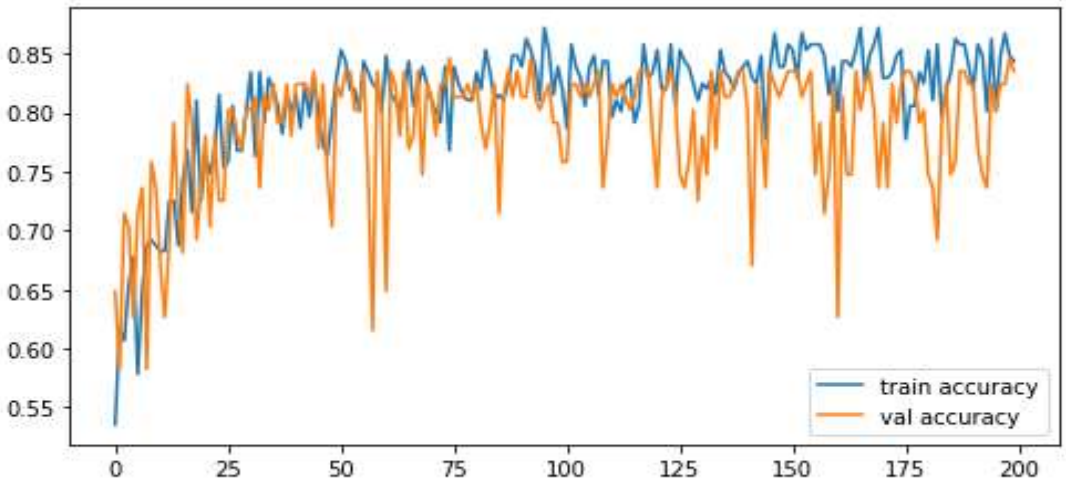

Figure 5: **Plot for Accuracy**

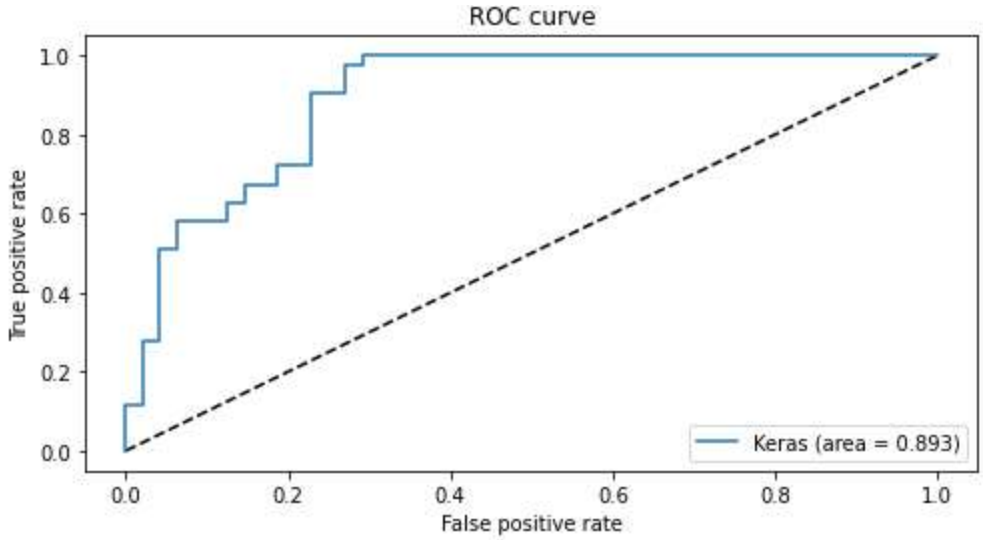

Figure 6**: ROC curve plot**

ROC and AUC metrics are used In the Artificial Neural Network and it increases the efficiency performance[17]. Visual presentation of the model is useful to create a sense of statistics which is removed from the model and to make a learned decision for the completing the limitations without disturbing the Machine Learning Model.

An efficiency of 86.73% is achieved in the proposed method. This suggested method is perfect for feature selection as it increases the presentation of ANN

5.1 Performance of evaluation metrics

Accuracy, Precision, Recall and F1 measures are the various methods used to access the model. Accuracy provides appropriate results from the available instances. Accuracy is represented in the equation 3:

$$\text{Accuracy} = \frac{Tpos^{*}+Tneg^{*}}{Tpos^{*}+Tneg^{*}+Fpos^{*}+Fneg^{*}} \quad (3)$$

Precision evaluates the performance of a classifier in detail. If the model has low positives, then the precision will be high and if the model has high positives then the precision will be low. The formula for precision is represented in the equation 4,

$$\text{Precision} = \frac{Tpos^{*}}{Tpos^{*}+Fneg^{*}} \quad (4)$$

Recall measures the fullness of the classifier. The higher the recall, the more positive samples are detected. The formula for recall is represented in the equation 5,

$$\text{Recall} = \frac{Tneg^{*}}{Tneg^{*}+Fneg^{*}} \quad (5)$$

F1 measure is the mixture of recall and accuracy. The F1 measure is calculated from the precision and recall of the test. F1 measure is represented in the equation 6,

$$\text{F1 measure} = \frac{2\times P\times R}{P+R} \quad (6)$$

Table 2: **Comparison of performance metrics**

Method	Accuracy	Precision	Recall	F1 measure
4- class method	86.6%	86.6%	81.74%	84.14%
Classification tree	77%	79%	73%	84%
Proposed	86.73%	86%	84%	83%

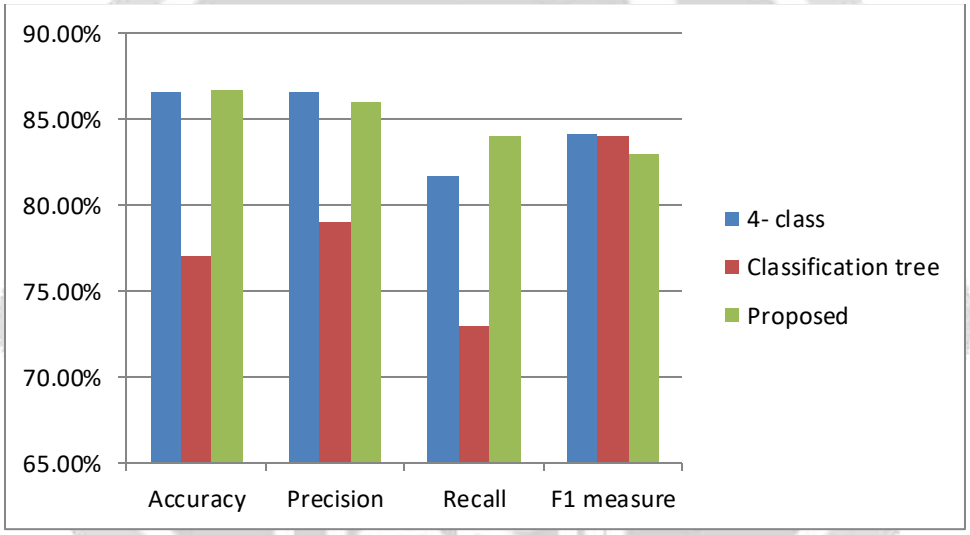

Figure 7**: Performance metrics**

The table 2 shows the proposed method has Accuracy of 86.73%.In the proposed system, the score for recall is 84%, 86%for precision and for F1 measure is 83%.

6. CONCLUSION

Heart disease can be prevented by finding it at an early stage. Heart disease can be predicted with the classification of a good algorithm and can prevent it before occurring. The heart disease detection system is assisted based on the patient's clinical information. In the proposed method, the accuracy, precision, recall and F1 measure is found out. High prediction accuracy of 86.73% is obtained in the proposed method and the proposed model is then used in medical diagnosis. Improved Artificial Neural Network (ANN) Classifier algorithm is used for the heart disease prediction in this paper. Patient's heart disease status can be predicted using this system obtained from UCI

Machine Learning Repository. Effects of the networks different parameter is shown in the prediction accuracy. Principal component analysis is used in this paper to show that the 14 attributes are vital for the diagnosis of the heart disease. In future works, further enhancement of this research paper will be done.

REFERENCES

[1] A. H. Chen, S. Y. Huang, P. S. Hong, C. H. Cheng, and E. J. Lin, "HDPS: Heart disease prediction system," in *2011 Computing in Cardiology*, Sep. 2011, pp. 557–560.

[2] N. Bhatla and K. Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques," *Int. J. Eng. Res.*, vol. 1, no. 8, p. 5, 2012.

[3] "Ensemble Methods for Heart Disease Prediction | SpringerLink." https://link.springer.com/article/10.1007/s00354-021-00124-4 (accessed Jul. 23, 2022).

[4] M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," in *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, Dhaka, Bangladesh, Sep. 2016, pp. 1–5. doi: 10.1109/CEEICT.2016.7873142.

[5] P. Motarwar, A. Duraphe, G. Suganya, and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, Vellore, India, Feb. 2020, pp. 1–5. doi: 10.1109/ic-ETITE47903.2020.242.

[6] P. Ramprakash, R. Sarumathi, R. Mowriya, and S. Nithyavishnupriya, "Heart Disease Prediction Using Deep Neural Network," in *2020 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, Feb. 2020, pp. 666–670. doi: 10.1109/ICICT48043.2020.9112443.

[7] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Comput. Appl.*, vol. 29, no. 10, pp. 685–693, May 2018, doi: 10.1007/s00521-016-2604-1.

[8] A. Mehmood *et al.*, "Prediction of Heart Disease Using Deep Convolutional Neural Networks," *Arab. J. Sci. Eng.*, vol. 46, no. 4, pp. 3409–3422, Apr. 2021, doi: 10.1007/s13369-020-05105-1.

[9] R. P. E, D. C. A. Subasini, D. A. V. Katharine, V. Kumaresan, S. G. Kumar, and T. M. Nithya, "A Cardiovascular Disease Prediction using Machine Learning Algorithms," *Ann. Romanian Soc. Cell Biol.*, pp. 904–912, Mar. 2021.

[10] M. A. Shaik and D. Verma, "Prediction of heart disease using swarm intelligence based machine learning algorithms," *AIP Conf. Proc.*, vol. 2418, no. 1, p. 020025, May 2022, doi: 10.1063/5.0081719.

[11] A. Kondababu, V. Siddhartha, BHK. B. Kumar, and B. Penumutchi, "A comparative study on machine learning based heart disease prediction," *Mater. Today Proc.*, Feb. 2021, doi: 10.1016/j.matpr.2021.01.475.

[12] K. Burse, V. P. S. Kirar, A. Burse, and R. Burse, "Various Preprocessing Methods for Neural Network Based Heart Disease Prediction," in *Smart Innovations in Communication and Computational Sciences*, Singapore, 2019, pp. 55–65. doi: 10.1007/978-981-13-2414-7_6.

[13] R. Kavitha and E. Kannan, "An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining," in *2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*, Feb. 2016, pp. 1–5. doi: 10.1109/ICETETS.2016.7603000.

[14] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," *Expert Syst. Appl.*, vol. 68, pp. 163–172, Feb. 2017, doi: 10.1016/j.eswa.2016.10.020.

[15] N. M. H.S, "ANN Model to Predict Coronary Heart Disease Based on Risk Factors," *Bonfring Int. J. Man Mach. Interface*, vol. 3, no. 2, pp. 13–18, Mar. 2013, doi: 10.9756/BIJMMI.4473.

[16] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection," *Glob. J. Comput. Sci. Technol.*, May 2013, Accessed: Jul. 23, 2022. [Online]. Available: https://computerresearch.org/index.php/computer/article/view/367

[17] R. Kannan and V. Vasanthi, "Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart Disease," in *Soft Computing and Medical Bioinformatics*, N. B. Muppalaneni, M. Ma, and S. Gurumoorthy, Eds. Singapore: Springer, 2019, pp. 63–72. doi: 10.1007/978-981-13-0059-2_8.