

Designing a Predictive Model for Heart Disease Detection Using Artificial Neural Network

Antim Dev Mishra¹, Mr. Bijender Mehandia²

¹M. Tech. Student, ECE, Gurgaon Institute of Tech. & Mgmt., Gurgaon, Haryana (India)
antimdevmishra@gmail.com

²Assistant Professor, ECE, Gurgaon Institute of Tech. & Mgmt., Gurgaon, Haryana (India)
bijender82mehandia@gmail.com

Abstract

The diagnosis of heart disease is a complex task requiring professional experience and domain knowledge. Doctor's examine large body of medical records such as ECG, BP, heart MER etc. in traditional approach for diagnosis of heart disease. Unfortunately all doctors are not equally proficient and all hospitals do not have adequate diagnostic facility. Nowadays, data mining technique and neural network tools are widely for predicting medical diseases by using health care industry dataset. In the field of bioelectronics current research focus is on critical disease like Cancer, Diabetes, and Thyroid etc. This paper describes a "Heart Disease Prediction System using two best model of Artificial Neural Network i.e. Multilayer Perceptron and Radial Basis Function. This system helps in predicting the likelihood of a patient developing heart disease based on the current test data. In this heart disease prediction system we are using dataset from "UCI Machine Learning Repository" which consists of ten attributes viz. age, sex, serum cholesterol, episode of chest pain, resting BP, ECG, heart rate, exercise induced angina and 209 instances. It is seen that neural network prediction system predicts heart disease with nearly 92.823% accuracy in Multilayer Perceptron Model, it built with 209 instances where all 10 attributes correctly classified 194 (92.823%) instances while 15 (7.177%) of the instances were classified incorrectly and the accuracy of Radial Basis Function model is 80.382% , it also built with 209 instances where correctly classified instances 168 and incorrectly classified instances 41 attributes and the algorithm took 8 seconds to build the model and more than 5 minutes to produce confusion matrix using WEKA 3.7.9 and WEKA 3.6.6 software.

Keywords: NN, Back propagation, Data mining, Heart Disease, MLP, RBF.

I. INTRODUCTION

Heart is the important part of human body. Our life is dependent on proper working of heart. If working of heart is not proper then it will affect the other parts of the body .Heart disease affect the operation of heart. There are many factors which increases risk of heart disease like cholesterol, high blood pressure, obesity etc.

Nowadays, the Heart disease is the major cause of deaths in whole world. According to WHO report every year about 12

million deaths occur worldwide due to the Heart diseases. An estimated 17.3 million people died from heart disease in 2008. Over 80% of heart disease deaths take place in low and middle income countries. WHO estimated by 2030 more than 23 million people die annually from heart disease. Diagnosis of heart disease is based on symptoms and physical examination of a patient. Doctors are predicting heart disease by deep learning and more experience. The diagnosis of heart disease is not an easy task in medical field because wrong prediction can cause death of patient. Nowadays, Hospitals and healthcare industries collecting the large amount of patient data, these data is the key resources to analyzed and predict the disease efficiently. Only human intelligence alone is not enough for proper diagnosis so this type of disease prediction system can give the efficient result. We have developed an efficient Heart Disease prediction system using neural networks this will improve the efficiency of prediction and reduce the diagnosis time.

II. BACKGROUND

To select a best classifier software tool that best suits the objective of this study some factors were needed to be taken into consideration. The major factors include:

- 1) The specific analysis which needs to be performed, in this case classification. The algorithms supported by the data mining tool.
- 2) The format of the data available.
- 3) The platform where the data mining tool runs on.

After taking the above factors into consideration and comparing different data mining software Weka 3.6.9 machine learning software.

Weka 3.6.9 and Weka 3.6.6 machine learning software was selected because of three principal reasons, First, it is open source, which not only means that it is freeware software, but

it is editable , without depending on the commitment, health, or longevity of any particular institution or company. Second, it provides a wealth of state-of-the-art machine learning algorithms that can be deployed on any given problem. Third, it can be implemented and run on any platform.

III. Algorithms Used For Model Building And Performance Measures

This study aims at applying classification techniques for heart disease detection. An attempt was made to construct prediction model using Neural Network Classifier. After constructing the models, performance of models were evaluated. In this section the algorithms used to build the models and matrices used for performance measure. Neural networks are network structures consisting of a number of nodes connected through hidden node. Every node has its own value, and the links between nodes specify the causal relationship between connected nodes. Neural networks can be used for many purposes, notably descriptive and predictive data mining. They were originally developed in the field of machine learning to try to imitate the neurophysiology of the human brain through the combination of simple computational elements (neurons) in a highly interconnected system .

Neural networks became popular in the 1980s because of its valuable behavior .First; computing power was readily available, especially in the business community where data was available. Second, analysts became more comfortable with neural networks by realizing that they are closely related to known statistical methods. Third, there was relevant data since operational systems in most companies had already been automated. Fourth, useful applications became more important than the holy grails of artificial intelligence. Building tools to help people superseded the goal of building artificial people. Because of their proven utility, neural networks are, and will continue to be, popular tools for data mining.

The architecture (or topology) of a neural network is generated by software and train the data on two special algorithm ,multilayer perceptron (MLP) and radial basis function (RBF) with fixed learning rate 0.3 and 500 epochs.

IV. DESCRIPTION OF THE DATASET

The dataset is taken from UCI Machine Learning Repository (1).The variables in the dataset include Age, Sex, BP, ECG etc. The 10 attributes and their description are shown in following Table IV-1 and IV-2

TABLE IV-1: DATASET MODEL

age	Sex(M=1,F=0)	Serum cholesterol	Chest pain	Rest BP	Blood sugar	Rest ECG	Max Heart Rate	Exr. angina	Disease
45	1	322	asympt	140	0	normal	135	yes	positive
39	0	561	Atyp_angi	160	1	normal	160	no	negative

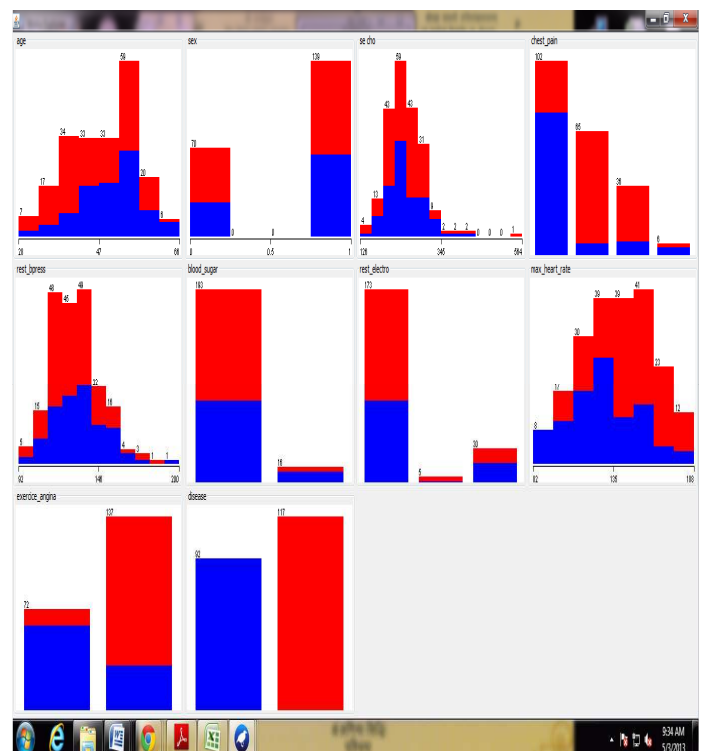


FIG. 1 DESCRIPTIVE DATA VISUALIZATION

Table IV-2: NUMERICAL ATTRIBUTES DATA SUMMARY

Attributes	Min	Max	Mean	Std. Dev
Age	28	66	47.967	8.054
Sex	0	1	.665	.473
Serum Cholesterol	126	564	249.55	52.87
Chest Pain(4type) 1. asympt 2. atyp_angina 3. non_anginal 4.typ_angina	Weight 102 65 36 6			
Resting BP	92	200	133.66	17.43
Blood Sugar >120mg/dl=1 <=120mg/dl=0	0	1	.077	.267
Resting (ECG) Electrocardiogram 1. normal 2. le01ven1_hyper 3. S1_wave_abnor.	Weight 173 5 30			
Max. Heart Rate	82	188	137.5	23.87
Exercise Angina 1. yes 2. no	Weight 72 137			
Disease 1.positive 2.negative	Weight 92 117			

The chest discomfort with activity is known as exercise angina. Angina is generally caused from not getting enough blood through the arteries to supply the walls of the heart with enough blood flow to adequately pump. Angina can be caused by blockage, injury or spasms. It can be particularly irritating when it comes on with exercise. In the dataset used only two stage 'yes' and 'no'. The electrocardiogram is a recording of the electrical activity of the heart as it undergoes excitation (depolarization) and recovery (polarization) to initiate each beat of the heart. Serum cholesterol is the total amount of cholesterol found in our blood. It is a soft, waxy substance found in the blood and cells of the body. It is not possible to determine if the pain you are feeling in your chest is because of a heart problem without a medical professional diagnosing the problem. Experience may play some role but evidence suggests that increasing physician experience decreases threshold for admission without an increase in diagnostic accuracy.

V. EXPERIMENTATION

As the goal of this study is to detect heart disease using data mining techniques a classification technique was adopted to develop a predictive model. The models were built with

supervised machine learning algorithms i.e. Neural Network using Weka 3.7.9 machine learning software.

Model Building Using Neural Network

The experiment was designed to explore the ability of Neural Network in predicting heart disease. From Neural Network Algorithms Multilayer Perception and RBF was selected to conduct the experiment. In this experiment two scenarios were considered, one containing all 10 attributes and the other containing the selected 5 attributes with learning rate 0.3 and 500 epochs.

Experiment 1(Multilayer Perceptron)

On the first scenario the algorithm was run on a full training set containing 209 instances with 10 attributes and the algorithm took 12.9 seconds to build the model and more than 11 minutes to produce confusion matrix. On the second scenario the algorithm was run on a full training set containing 209 instances with selected 5 attributes and took 8.14 seconds to build the model and more than 20 minutes to produce confusion matrix.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{TP rate} = \frac{TP}{TP + FN}$$

$$\text{TN rate} = \frac{TN}{FP + TN}$$

TABLE V-1: CONFUSION MATRIXES FOR EXPERIMENT

Model	Confusion Matrix		
Neural Network With All Attribute	Yes (Predicted)	No (Predicted)	Actual
	86	6	Yes
	9	108	No
Neural Network With Selected Attribute	Yes (Predicted)	No (Predicted)	Actual
	74	18	Yes
	15	102	No

TABLE V-2: DETAILED PERFORMANCE MEASURES FOR EXPERIMENT

Model	Accuracy	TP Rate	TN Rate	Precision	F-Measure	ROC Area
MLP With All Attribute	92.823%	.935	.065	.905	.920	.943
MLP With Selected Attribute	84.2105%	.804	.195	.813	.804	.898

The first Neural Network model built on all 10 attributes correctly classified 194(92.823%) instances while 15 (7.177%) of the instances were classified incorrectly. The overall accuracy rate of the model is successful. The model correctly identified 209 patients out of 194 patients who had heart disease and the remaining 15 were identified incorrectly to be free from the disease while they actually had the disease. This result gave the model a TP Rate of 0.935. The model is identifying negative cases as the TN Rate of the model is 0.065 by correctly identifying 15 patients out of 209 patients who didn't had heart disease and the remaining 194 were identified to have the disease while they actually didn't.

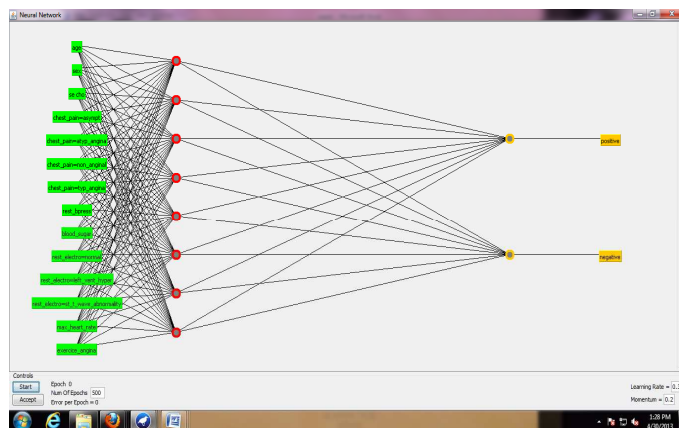


FIG 2: CLASSIFIER THAT USES BACK PROPAGATION TO CLASSIFY 10 ATTRIBUTES

Regarding to precision score of the model with an average precision of 0.905% it is a very successful model in retrieving relevant values for each class. With F-Measure value of 0.920 it can be concluded that the Precision and the Recall of the model

are significantly balanced.

The second Neural Network model built on 6 selected attributes correctly classified 176 (84.2106%) instances while 33 (15.7895%) of the instances were classified incorrectly. This model has a significant decrement regarding to overall classification accuracy rate and increment in execution time compared to a similar experiment conducted on all attributes. The model correctly identified 209 patients out of 176 patients who had heart disease and the remaining 33 were identified incorrectly to be free from the disease while they actually had the disease. This result gave the model a TP Rate of 0.804. The model is identifying negative cases as the TN Rate of the model is 0.184 by correctly identifying 33 patients out of 209 patients who didn't had heart disease and the remaining 176 were identified to have the disease while they actually didn't.

Regarding to Precision score of the model, with an average precision of .813 it is a very successful model in retrieving relevant values for each class. With F-Measure value of 0.804 it can be concluded that the Precision and the Recall of the model are significantly balanced. The results show that Neural Network model performed better on the all attribute in comparison to selected attributes. The classification accuracy increased to 92.823% from 84.2105%. And also, the execution time decreased significantly to 34.14 seconds from 158.94 seconds.

EXPERIMENT 2(RADIAL BASIS FUNCTION)

In the radial basis function algorithm was run on a full training set containing 209 instances with correctly classified instances 168 and incorrectly classified instances 41 attributes and the algorithm took 8 seconds to build the model and more than 5 minutes to produce confusion matrix.

TABLE V-3: CONFUSION MATRIXES FOR EXPERIMENT 2

Model	Confusion Matrix		
RBF With All Attribute	Yes (Predicted)	No (Predicted)	Actual
	68	24	Yes
	17	100	No

TABLE V-4: DETAILED PERFORMANCE MEASURES FOR EXPERIMENT

Model	Accuracy	TP Rate	TN Rate	Precision	F-Measure	ROC Area
RBF With All Attribute	80.382 %	.739	.114	.8	.768	.87

VI. DISCUSSION

Thus, the effort was to develop a predictive model that is capable of diagnosing heart disease being less dependent on the observer, especially when the observer is less experienced. Experiments were conducted and classification performance has been compared in order to determine optimal statistical algorithms for predicting heart disease.

VII. CONCLUSION

In this study, the aim was to design a predictive model for heart disease detection using data mining techniques on UCI .Data collected from UCI Machine Learning Repository 209 instances was selected and preprocessed for this study. The models were built on the preprocessed UCI dataset with supervised machine learning algorithms i.e. Multilayer Perception and Radial Basis Function using Weka 3.7.9 and Weka 3.6.6 machine learning software.

The performances of the model were evaluated using the standard metrics of accuracy, precision, recall and F-measure. This effective model to predict patients with heart disease appears to be a MLP neural network on all attribute with a classification accuracy of 92.823%. Selected attributes with a classification accuracy of 92.823%. from a total of 10 attributes that were available, 6 attributes that are highly relevant in predicting heart disease from UCI dataset were selected. Heart disease is a fatal disease by its nature and misdiagnosis of this disease can be dangerous for life. The best model selected for predicting of heart disease could not exceed a classification accuracy of 92.823% and still much remains to fill the gap of 7.177% misclassified cases.

This study showed that data mining techniques with artificial neural network can be used efficiently to model and predict heart disease cases. The output of this research can be used as an assistant tool by cardiologists to help them to make more consistent diagnosis of heart disease. Furthermore, the resulting model has a high specificity rate which makes it a handy tool for junior cardiologists to screen out patients who have a

high probability of having the disease and transfer those patients to senior cardiologists for further analysis.

VIII. MODEL COMPARISON

After performing the experiments the next step was comparing the models and selecting the best model. The models were compared using different performance measures like accuracy, TN Rate, TP Rate, F-Measure, ROC Area and execution time (time taken to build the model).

TABLE VIII-1: PERFORMANCE SUMMARY OF THE MODELS

Model	Accuracy in %	TP Rate	TN Rate	F-Measure	ROC Area	Time(Sec.)
MLP with all attributes	92.823	.935	.065	.920	.943	12.9
MLP with selected attributes	84.210	.804	.195	.804	.898	8.14
RBF with all attributes	80.382	.739	.114	.768	.87	8

As presented on Table VIII all classification algorithms performed nearly equally well with a remarkable accuracy of up to 92.823% while the lowest accuracy score is 80.382%. A MLP classifier which was implemented on all attributes achieved the highest accuracy (92.823%).

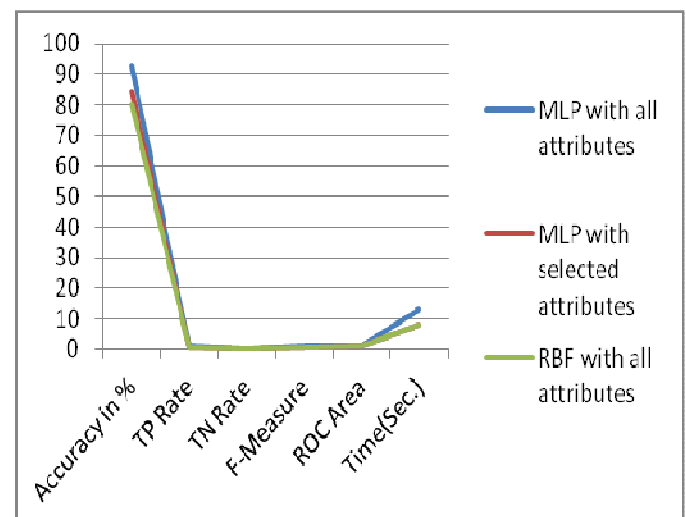


FIGURE 3: PERFORMANCE COMPARISONS OF THE MODELS

IX. RECOMMENDATION

This study has indicated that neural network and data mining techniques can be applied in the prediction of heart disease and the resulting models of this study are worthy of clinical testing. To improve the classification accuracy of the models further researches should be conducted using different algorithms. As a future work, the researcher has planned to perform additional experiments with more dataset and algorithms to improve the classification accuracy and to build a model that can predict specific heart disease types.

REFERENCES

- [1] UCI machine learning repository
<http://archive.ics.uci.edu/ml/datasets.html>.
- [2] Benjamin, M (2006). Introduction to Heart Disease. Available at
http://www.mentalhelp.net/poc/view_doc.php?type=doc&id=4496
- [3] SN Shivnandm Introduction to Neural Networks using Matlab6.0
- [4] Diversity in Neural Network Ensembles. The University of Birmingham. 2004.
- [5] Detection of heart disease technique using data mining by Able Damtey Addis Ababa University.
- [6] Prediction of Heart Disease Medical Preprocessing using Radial Basis Function by Shaikh Abdul Hannan 978-1-4244-5967-4/10/\$26.00 ©2010 IEEE.