Neural Network based Heart Disease Prediction

Mrs. K. Uma Maheswari
Assistant Professor,
Department of Information Technology,
Anna University (BIT Campus),
Tiruchirappalli-24.

Ms. J. Jasmine
PG Scholar,
Department of Computer Science Engineering,
Anna University (BIT Campus)
Tiruchirappalli-24

***Abstract*— Machine learning is widely used to make the machine to learn and to predict when exposed to new data. Due to many advancements in machine learning, there are various methods that can be adopted to predict the heart disease of an individual. Heart Disease is one among the major diseases affecting the individual around the world. There are several risk factors which leads to heart disease. The combination of logistic regression analysis and neural network provides a novel approach in predicting the heart disease. Initially logistic regression is applied to select the major risk factors for predicting the disease. It produces the significant risk factors that are useful in predicting the heart disease based on statistical p-value. The risk factors which are not having the significant impact are identified and removed. The resultant significant factors are provided as input to the neural network. And neural network is trained for the risk factors that is obtained from logistic regression and used to test whether the person is having the heart disease or not. Thus, the integration of logistic regression and neural network is applied in predicting the heart disease.**

Keywords*— *Logistic regression model, neural networks, prediction, heart disease

I. INTRODUCTION

Heart Disease is one of the leading disease around the world. The effective functioning of the heart plays a vital role in the body. There are many types of heart diseases such as Myocardial infarction, Myocardial ischaemia, Congenital heart disease, Coronary heart disease, Cardiac arrest, Peripheral heart disease etc., There are various methodologies available in predicting the heart disease.

Machine learning is a type of artificial intelligence that makes the machines to learn from training data and makes predictions on the test data based on the learned data. The basic idea behind the machine learning is to find the patterns among the data and make the predictions. There are numerous applications in machine learning such as in recommender systems, medical diagnosis, bioinformatics etc., Basically, there are three types of learning in machine learning such as supervised learning, unsupervised learning and reinforcement learning.

Predictive analytics include various statistical techniques from predictive modeling, machine learning (ML) and data mining to make predictions based on the current or historical data. The use of predictive analytics are in the customer relationship management, healthcare industry and in many other fields. Deep learning has a significant impact on the predictive analytics. There are many models in the predictive modeling [10] such as Naive bayes, Logistic regression [9], Neural networks [20], Support Vector Machine [18], Classification and Regression trees [8] etc.,

Artificial neural network (ANN) is one of the mathematical or algorithmic approach. It is similar to the human brain neurons. The artificial neural network has connections, propagation direction and discrete layers. Each layer is made up of nodes with the arrows that represents the interconnections between them. In the neural network, there is many nodes in the input layer. These input layer nodes are connected to the hidden layer nodes. Each input is assigned with the weights. The input nodes in the network passes the data to the nodes in the hidden layer which performs some tasks or computations and send the processed data to the output node. The output layer has the node which yields the final result. This is an overview of the process of neural network.

II. RELATED WORK

The risk factors for Coronary heart disease (CHD) or atherosclerosis is identified by V. Sree Hari Rao et al., [1] using the in-built imputation algorithm and particle swarm optimization. It is obtained that physical inactivity also forms the risk factor of CHD. The decision rules are used in predicting the risk factors of heart disease.

In Paolo Melillo et al., [2] the risk assessment in patients suffering from congestive heart failure is done by automatic classifier. Using the measure called Long-term heart rate variability, this classifier classifies the lower risk patients from higher risk patients automatically. Classification and Regression tree (CART) is useful in classifying the higher risk patients and the lower risk patients. The heart rate variability is also major important factor in finding the risk assessment of congestive heart failure.

The risk assessment of heart targeting the reduction of Coronary heart disease (CHD) events are classified into before the event and after the event by Minas A. Karaolis et al., [3]. The event before the CHD and after the event which are non-modifiable and the modifiable are identified. The events are percutaneous coronary intervention (PCI), myocardial infarction (MI) and coronary artery bypass graft surgery (CABG). The C4.5 decision tree algorithm is used for these three events of coronary heart disease.

Carlos Ordonez et al., [4] used association rules for prediction of heart disease. These association rules are applied on the medical dataset and it generates many rules which are irrelevant. In order to identify the rules that are

truly essential for predicting the heart disease are identified by using search constraints which searches the association rules in training dataset and finally validates on the test set.

The hybrid system is used with the global optimization of genetic algorithm in Syed Umar Amin et al., [5] and this system is used to initialize the neural network weights. A multilayered feed-forward network is used. The input nodes, hidden nodes and output nodes are 12, 10 and 2 respectively. The input nodes are basically the risk factors which are used in predicting the heart disease.

The risk factors of heart attack otherwise called myocardial infarction is identified based on the decision trees and apriori algorithm Sikander Singh Khurl et al., [6] Based on these methods, the risk factors which are identified as efficient in the detection of heart attack are chest pain, diabetes, smoking, gender and physical inactivity, age, lipids, cholesterol, triglyceride, blood pressure.

III. PROPOSED SYSTEM

The ultimate goal is to combine the logistic regression model and neural network based approach in the prediction of heart disease. The heart disease dataset has 303 observations of individuals out of which 297 observations are taken for consideration. The proposed system mainly consists of two parts. The first part is to find the important risk factors in predicting the heart disease from the available risk factors in the dataset based on the p-value. This p-value yields the significant codes [19] for each attribute. And the second part is to divide the dataset into training and testing dataset. The neural network is build for the training dataset and the learned neural network is able to predict the testing dataset.

A. Data Collection

The data used in this project is obtained from the Cleveland Heart Disease database. A total of 297 records with 14 medical attributes [7] which is used to predict the heart disease. The description of dataset is tabulated in Table 1.

Table I. Description of Dataset

S.No	Attribute Name	Description
1	Age	Age in years
2	Sex	1= male, 0= female
3	Cp	Chest pain type (1= typical angina, 2= atypical angina, 3= non-anginal pain, 4= asymptomatic
4	Trestbps	Resting blood pressure (in mm Hg on admission to hospital)
5	Chol	Serum Cholesterol in mg/dl
S.No	**Attribute Name**	**Description**
6	Fbs	Fasting blood sugar >120 mg/dl (1= true, 0= false)
7	Restecg	Resting electrographic results (0= normal, 1= having ST-T wave abnormality, 2= left ventricular hypertrophy
8	Thalach	Maximum heart rate
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment (1= upsloping, 2= flat, 3= downsloping)
12	Ca	Number of blood vessels colored by fluoroscopy
13	Thal	3= normal, 6= fixed defect, 7= reversible effect
14	Num	Class (0= healthy, 1= have heart disease)

B. Logistic Regression Model

Logistic regression model is one of the statistical regression model and it has the capacity to measure the relationship between the categorical dependent variable and one or more independent variable. Here the independent variables are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrographic results, maximum heart rate, exercise induced angina, oldpeak-slope of the peak, slope of the peak exercise, blood vessels affected, thal defect. The dependent variable is the class which is to be predicted as healthy or having heart disease. In the binary logistic regression analysis, the dependent variable is coded as "0" or "1" indicates that the individual having the heart disease or not respectively.

The logistic regression model computes the probability of the haert disease as a function of the risk factors. We can compute the conditional probability $p(y=1 \mid X)$, where $X = (x1, x2 \ldots xn)$ represents n risk factors associated with the disease. As a result we could calculate the likelihood of an individual suffering the disease. The cutoff value can be set to 0.5. If the cutoff value is greater than 0.5, we can infer that the individual suffers from the heart disease; otherwise, the individual is free from the heart disease. Apart from this, the logistic regression model has the capacity to select factors that have significant impact on the heart disease based on the statistical significance *p*-value.

The *p*-value is the short form for probability value and it is the probability that is given by the summary of the logistic regression model. Statistical hypothesis testing make use of *p*-values and it is used in many fields of research such as political science, economics etc., The *p*-value is defined as the marginal significance to represent the probability of the occurrence of a given event. The logistic regression model is able to select the important risk factors or attributes that are used in predicting the heart disease from the fourteen attributes of the dataset. By applying the logistic regression model, the risk factors or attributes which are having the *p*-value less than 0.05 ($p<0.05$) as variable inclusion criteria.

And for variable exclusion criteria, the statistical significant p-value is greater than 0.1 ($p>0.1$). The larger *p*-value indicates the changes in the independent variable are not associated with the changes in the dependent variable. The significant risk factors which are obtained from statistical significant *p*-value of logistic regression model are sex, chest pain type, resting blood pressure, fasting blood sugar, exercise induced angina, slope of the peak exercise, blood vessels affected and thal defect.

C. Training and Testing Dataset

The dataset of 297 records [16] are divided into training and testing dataset. The training dataset is used to build a predictive relationship and it is the set of examples that is used for learning and to fit weights of the classifier. The test set is set of examples which is used to evaluate the performance of a fully-specified classifier. The training and testing dataset is divided into 75% and 25% respectively.

D. Building Neural Network

The neural network is a computational model based on biological neural networks. Artificial neural networks (ANN) is based on observation of a human brain. Human brain is very complicated web of neurons [17]. Analogically ANN is an interconnected set of three units such as input, hidden and output unit. In medical diagnosis, the patient's risk factors or attributes is used as an input. The effectiveness of artificial neural network was proven in medicine. ANN are used in predicting coronary heart disease. Here the input layer consisting of 8 neurons corresponds to 8 significant attributes. There is one output class variable which takes the value either 0 or 1. The value 0 represents that the individual is not suffering from heart disease and the value 1 represents that the individual suffers from heart disease. The number of nodes used in the hidden layer are 3. The Sample Artificial Neural Network is shown in Fig 1.

The main advantage of neural networks is high accuracy. The applications of neural network are accounting, medicine [14], fraud detection etc.,. Based on the learned network or training dataset, the neural network is able to predict the presence or absence of heart disease for the testing dataset.

E. Performance Measures

The performance measures of neural network are calculated using various measures such as accuracy, specificity and sensitivity. The accuracy obtained by the neural network is 84%. And the sensitivity and specificity obtained are 91.4% and 77.5% .

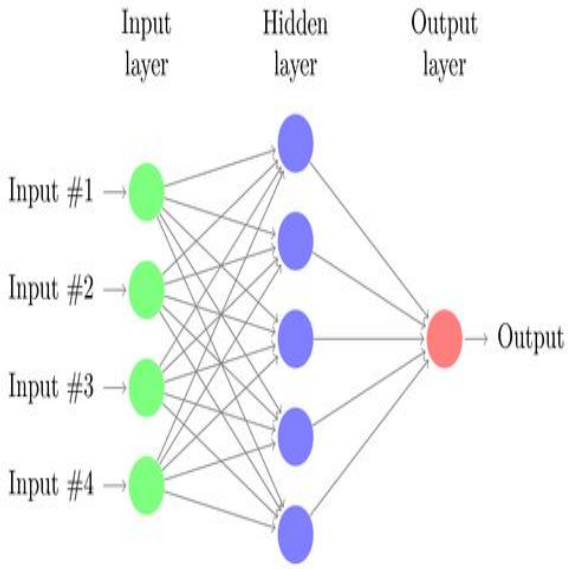

Fig. 1. Sample Artificial Neural Network

Accuracy = (TP+TN) / (TP+TN+FP+FN)
Specificity = TN / (TN+FP)
Sensitivity = TP / (TP+FN)

where,
TP = True Positive; that is the number of samples which are classified as having heart disease while they were actually have heart disease.
TN = True Negative; that is the number of samples which are classified as not having heart disease while they were actually not.
FN = False Negative; that is the number of samples which are classified as not having heart disease while they were actually have heart disease
FP = False Positive; that is the number of samples which are classified as having heart disease while they were actually not.
The accuracy, sensitivity and specificity of Neural Network is tabulated in Table 2.

Table 2. Accuracy, Sensitivity and Specificity of Neural Network

Classifier	Accuracy	Sensitivity	Specificity
Neural Network	84%	91.4%	77.5%

IV. CONCLUSION

Machine learning is active research area in which many researchers are working in healthcare domain for disease risk identification. The advantage of logistic regression is the interpretability of model parameters and ease of use. The advantage of neural network is it requires less formal statistical training to develop and can implictly detect complex non-linear relationships between dependent and independent variables. The integration of logistic regression and neural network gives the novel approach in predicting the heart disease of an individual. The future work can be extended for longitudinal studies of the patients and to improve the accuracy in prediction of heart disease.

REFERENCES

[1] V. Sree Hari Rao, M. Naresh Kumar, “Novel Approaches for Predicting Risk Factors of Atherosclerosis,” IEEE Journal of Biomedical and Health Informatics ., vol. 17, No. 1, Jan 2013.
[2] Paolo Melillo, Nicola De Luca, Marcello Bracale and Leandro Pecchia ,"Classification Tree for Risk Assessment in Patients Suffering From Congestive Heart Failure via Long-Term Heart Rate Variability", IEEE Journal of Biomedical and Health Informatics., Vol. 17, No. 3, May 2013.
[3] Minas A. Karaolis,Joseph A. Moutiris, Demetra Hadjipanayi, Constantinos S. Pattichis, “Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees,” IEEE Transactions on Information Technology in Biomedicine, Vol. 14, No. 3, May 2010
[4] Carlos Ordonez ,“Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction", IEEE Transactions on Information Technology in Biomedicine, Vol. 10, No. 2, April 2006.
[5] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg,"Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors", Proceedings of IEEE Conference on Information & Communication Technologies, 2013.

[6] Sikander Singh Khurl, Gurpreet Singh, "Ranking Early Signs of Coronary Heart Disease Among Indian Patients", IEEE International Conference on Computing for Sustainable Global Development, 2015
[7] The UCI Machine Learning Repository[online] https://archive.ics.uci.edu/ml/datasets/Heart+Disease
[8] Stephenie C. Lemon, Jason Roy, and Melissa A. Clark, Peter D. Friedmann, William Rakowski, "Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison With Logistic Regression", The Society of Behavioral Medicine, Vol. 26, No. 3, 2003, pp. 172-181.
[9] David W. Hosmer, Jr., Stanley Lemeshow, Rodney X. Sturdivant, "Applied Logistic Regression", 2013.
[10] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction ", International Journal of Computer Applications, Vol. 17, No.8, March 2011, pp. 43-48.
[11] Harleen Kaur , Siri Krishan Wasan and Vasudha Bhatnagar, "The Impact of Data Mining Techniques on Medical Diagnostics", Data Science Journal, Vol. 5, October 2006, pp. 119-126.
[12] K.Srinivas, B.Kavihta Rani , A.Govrdhan , "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal on Computer Science and Engineering (IJCSE), Vol. 02, No. 02, 2010, pp. 250-255.
[13] Asha Rajkumar, G.Sophia Reena, "Diagnosis Of Heart Disease Using Datamining Algorithm", Global Journal of Computer Science and Technology, Vol. 10 , September 2010.
[14] Qeethara Kadhim Al-Shayea, "Artificial Neural Networks in Medical Diagnosis ", International Journal of Computer Science Issues (IJCSI), s, Vol. 8, No. 2, March 2011, pp. 150-154.
[15] R. Dybowski and V. Gant, "Clinical Applications of Artificial Neural Networks", Cambridge University Press, 2007.
[16] Tutut Herawan, Rozaida Ghazali, Nazri Mohd Nawi, Mustafa Mat Deris, "Recent Advances on Soft Computing and Data Mining", The Second International Conference on Soft Computing and data mining (SCDM-2016).
[17] Mirza Cilimkovic," Neural Networks and Back Propagation Algorithm".
[18] Abhisek Acharya," Comparative Study of Machine Learning Algorithms for Heart Disease Prediction", Helsinki Metropolia University of Applied Sciences, Thesis, April 2017.
[19] Jamal Jokar Arsanjani, Alexander Zipf, Peter Mooney, Marco Helbich, "OpenStreetMap in GIScience: Experiences, Research, and Applications", 2015, pp.132.
[20] Sonali. B. Maind, Priyanka Wankar," Research Paper on Basic of Artificial Neural Network", International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), Vol. 2, No. 1, January 2014, pp. 96-100.