Prediction of Heart Disease Using Neural Network

Tülay Karayılan
Department of Computer Engineering Yıldırım Beyazıt University
Ankara, Turkey
155101122@ybu.edu.tr

Özkan Kılıç
Department of Computer Engineering Yıldırım Beyazıt University
Ankara, Turkey
ozkankilic@ybu.edu.tr

***Abstract*—Heart disease is a deadly disease that large population of people around the world suffers from. When considering death rates and large number of people who suffers from heart disease, it is revealed how important early diagnosis of heart disease. Traditional way of diagnosis is not sufficient for such an illness. Developing a medical diagnosis system based on machine learning for prediction of heart disease provides more accurate diagnosis than traditional way. In this paper, a heart disease prediction system which uses artificial neural network backpropagation algorithm is proposed. 13 clinical features were used as input for the neural network and then the neural network was trained with backpropagation algorithm to predict absence or presence of heart disease with accuracy of 95%.**

Keywords—heart disease, artificial neural network, Cleveland database, backpropagation, multilayer perceptron, machine learning

I. INTRODUCTION

Heart disease is the number one killer according to World Health Organization (WHO) statistics.[1] Millions of people die every year because of heart disease and large population of people suffers from heart disease. Prediction of heart disease early plays a crucial role for the treatment. If heart disease could be predicted before, lots of patient deaths would be prevented and also a more accurate and efficient treatment way could be provided.

A need to develop such a medical diagnosis system arises day by day. The important key points of such medical diagnosis systems are reducing cost and obtaining more accurate rate efficiently.

Developing a medical diagnosis system based on machine learning for prediction of heart disease provides more accurate diagnosis than traditional way and reduces cost of treatment.

In this paper, prediction of heart disease by an automated medical diagnosis system based on machine learning is proposed to satisfy this need. Backpropagation Algorithm which is commonly used Artificial Neural Network learning methodology, was used for the prediction system.

The remainder of this paper is arranged as follows: Next section gives insight on previous studies on heart disease based on machine learning. In Section III, proposed methodology of the prediction system is explained in detail. Section IV presents the experimental results which are obtained from the proposed methodology. Finally, conclusion of the paper is shown in Section V.

II. LITERATURE SURVEY

There are a lot of studies on prediction of heart disease as a medical diagnosis system. Firstly, R.W.Jones, M.Clarke, Z.Shen and T. Alberti have proposed a study applying neural network to self-applied questionnaire (SAQ) data to develop a heart disease prediction system. The study not only clarifies common risk factors of the disease but also the other data collected in SAQ. The validation of the work was provided by checking against the result of the neural network with "Dundee Rank Factor Score" which is related to statistically 3 risk factors (blood pressure, smoking and blood cholesterol) together with sex and age to determine risk of having heart disease. In the study, they used multi-layered feedforward neural network which was trained with Backpropagation Algorithm. There were three layers in the neural network they used: input, hidden and output layers. The performance was improved to Relative Operating Characteristic (ROC) area of 98% by increasing input numbers of the neural network.[2]

Ankita Dewan and Meghna Sharma have discussed various kinds of techniques for developing a heart disease prediction system and proposed using Backpropagation Algorithm as best classification technique for the targeted system. They also have proposed using Genetic Algorithm as optimizer against the Backpropagation Algorithm drawback of being stuck in local minima. The proposed methodology was intended for implementing in future with an accuracy of nearly 100% or with minimal error. [3]

Another study on heart disease prediction has been proposed and implemented by SY Huang, AH Chen, CH Cheng, PS Hong and EJ Lin. The classification and prediction was trained via learning Vector Quantization Algorithm which is one of Artificial Neural Network learning technique. There were three steps in their methodology. The first one was to select of 13 clinical features which are important compared to others, i.e., age, cholesterol, chest pain type, exercise induced angina, max heart rate, fasting blood sugar, number of vessels colored, old peak, resting ecg, sex, slope, thal and trestbps. Second one was using Artificial Neural Network algorithm for classification. Lastly, the heart disease prediction system was developed. The accuracy of prediction rate which was obtained from the study is near 80%. [4]

Jayshril S. Sonawane and D. R. Patil have come up with another Artificial Neural Network methodology for heart disease prediction. The used network was trained by Vector

978-1-5386-0930-9/17/$31.00 ©2017 IEEE

Quantization Algorithm using random order incremental training. There were three layers in the used network, including input, hidden and output layers. There were 13 neurons ,which is equal to the number of clinical data of heart disease database, in the input layer. The neurons of hidden layer could be changed to obtain less error and high accuracy. There was only single neuron in the output layer that denotes if heart disease present or not. The system performance was improved by training with varying number of neurons and variable epochs. The result shows that they obtained the highest accuracy of 85.55% when compared to others as stated in the paper. [5]

Majid Ghonji Feshki and Omid Sojoodi Shijani have another study on heart disease prediction by using feature selection and classification approach with a specific dataset. The proposed approach had three steps. The first step was process of dividing dataset into two subsets as sick and healthy people. In the second step, extracting 8192 subsets was done from the total features. In the third step, the best subset which has highest accuracy was found using PSO algorithm with a classifier algorithm, Feed Forward Backpropagation Algorithm. In the approach four classifier algorithms; C4.5, Multilayer Perceptron, Sequential Minimal Optimization and Feed Forward Backpropagation were used. Using feature selection and Backpropagation, neural network with PSO algorithm was pointed out as the most efficient method. The accuracy rate which was obtained from the study is 91.94%. [6]

R. R. Manza , Shaikh Abdul Hannan, R. J. Ramteke and A.V. Mane have a study whose aim is to predict prescription of heart disease by using an artificial neural network as classifier. The proposed methodology had five steps. In step 1, data which is about medicines given by doctor and patient suffering from heart disease was collected. In step 2, symptoms of heart disease and medicines were converted to binary form 0 or 1. 1 indicates medicine or symptom is present. In step 3, Radial Basis Function was used for training. In step 4, testing data was applied to evaluate classifier performance. In step 5, the Radial Basis Function prescribed the medicines for patients. The used network contained of three layers, including input, hidden and output layers. Responsibility of the input layer is not processing information. The task of input layer is just distributing the input vectors to hidden layer. There were a number of Radial Basis Function units in the hidden layer. 97% accuracy was obtained from the study. It was pointed out that the proposed approach could be extended by using Generalized Regression Neural Network.[7]

Syed Umar Amin, Dr. Rizwan Beg and Kavita Agarwal have proposed a hybrid system using Genetic Algorithms and Artificial Neural Networks for prediction of heart disease based on risk factors. The neural network was trained with Backpropagation Algorithm. It was pointed out that Backpropagation Algorithm has two major disadvantages. First problem is that finding out initial weights which are globally optimized is almost impossible. Second problem is slowness of Backpropagation Algorithm in convergence. The problems were solved by using Genetic Algorithm for optimization connection weights of Artificial Neural Network so as to obtain better performance from the network. The neural network used in the study had 12 input, 10 hidden and 2 output nodes. The results show that training accuracy is 96.2% and obtained validation accuracy is 89%.[8]

Jayshril S. Sonawane and D. R. Patil have a study whose aim is prediction of heart disease by Artificial Neural Network methodology. Multilayer perceptron neural network was used in the system. The proposed system had two steps. The first one was process of accepting 13 clinical data as input and as a last step training the network by Backpropagation Algorithm. There were three layers in the network, including input, hidden and output layers. There were 13 neurons which is equal to the number of clinical data of heart disease database in the input layer. The neurons of hidden layer could be varied to obtain less error and high accuracy. In the output layer, there was only single neuron denoting if heart disease present or not. The accuracy rate obtained from the study is 98%. [9]

Saba Bashir, M.Younus Javed and Usman Qamar have another study to predict heart disease. The proposed method uses Decision Tree, Support Vector Machine and Naive Bayes as a hybrid model. Majority voting scheme was obtained by these three classifiers. There were two steps in the proposed approach. First one was producing every three classifiers' decision. Second one was combining the decisions in order to acquire new model based on majority voting scheme. The results show that the accuracy rate obtained from the study is much higher than the others. 74% sensitivity, 82% accuracy and 93% specificity were obtained from the study to predict the heart disease. [10]

III. METHODOLOGY

In the heart disease prediction system, there are input variables, which are disease risk factors which are obtained from dataset, and output variables, which are a category, such as "disease absence" and "disease presence". Prediction of heart disease is called supervised learning problem. Because of having output variables are in category type, the prediction heart disease is "classification type of supervised learning".

Backpropagation Algorithm, which is commonly used Artificial Neural Network learning technique, was used for developing heart disease prediction system. Because Backpropagation Algorithm is the only technique which is used for nonlinear relationships which means it is the best classification algorithm for heart disease prediction.[3]

A. Artificial Neural Network

Artificial Neural Network (ANN) takes its inspiration from human brain which has incredible processing ability because of having webs of interconnected neurons. ANNs are designed by using basic processing unit called perceptron. Perceptron has only one layer and solves linearly separable problems. The problems which are not linearly separable can be solved by Multilayer Perceptron Neural Network (MLP). MLP has multiple layers, including input, hidden and output layers.

The proposed heart disease prediction system was designed as a multilayer perceptron neural network. The designed ANN has three layers: namely an input layer, a hidden layer and an output layer.

- *Input Layer* was designed to contain 13 neurons. Number of neurons was decided to be equal to the number of attributes in the data set.
- *Hidden Layer* was designed to contain 3 neurons. This number was decided as a startup point. The number was changed increasing one by one until it reached to the number of neurons of the input layer by comparing performance of them and then selecting the best one. This approach is based on one of machine learning best practices that the number of neurons of hidden layer should be the mean of the number of the neurons of input and output layers.
- *Output Layer* was designed to contain 2 neurons. The designed NN is a classifier going running in Machine Mode which means returning a class label (e.g., "Disease Presence"/"Disease Absence"). Deciding 2 neurons is based on idea that the output layer has one node per class label in model.

B. Backpropagation Neural Network

Backpropagation Algorithm (BA) is the most commonly used ANN learning technique. The steps of the algorithm are listed below:

- All network weights are initialized to small random numbers.
- Training data is received as input and output is computed for each unit with equation below known as Sigmoid Function:

$$o = \sigma(\vec{w} \cdot \vec{x}) \qquad \sigma(y) = \frac{1}{1+e^{-y}} \tag{1}$$

where $\vec{w}$ is vector of unit weight values and $\vec{x}$ is vector of network input values.

- Then error computation step is started. BP algorithm works as follows: Error signal(δ) which is calculated for each network output is propagated to all neurons in the network as input[11].
- Error term δ_k is calculated for each network output unit k using following equation:

$$\delta_k \leftarrow o_k(1-o_k)(t_k-o_k) \tag{2}$$

where o_k indicates network output for output unit k and indicates desired output for output unit k.

- Error term δ_h is calculated for each hidden unit h as below:

$$\delta_h \leftarrow o_h(1-o_h) \sum_{k \in outputs} w_{kh}\delta_k \tag{3}$$

where w_{kh} denotes network weight from hidden unit h to output unit k.

- Each network weight is updated where

$$w_{ji} \leftarrow w_{ji} + \Delta w_{ji} \text{ where } \Delta w_{ji} = \eta \delta_j x_{ji} \tag{4}$$

where η is learning rate and x_{ji} denotes the input from unit i into unit j. [12]

Backpropagation Algorithm was used for the proposed system as learning algorithm. 13 of the attributes of Cleveland dataset was used as input data for the designed neural network. The dataset was split into three parts: training, testing and validation. Then training was done with Backpropagation Algorithm. After training process, the performance of the proposed system was computed by testing the neural network with test data by different metrics including accuracy, precision and recall, which is explained in detail in the next section.

IV. EXPERIMENTAL RESULTS

The proposed heart disease prediction system which uses multilayer perceptron neural network was developed in MATLAB R2015a.

A. Data Source

Cleveland database was used for heart disease prediction system. Because Cleveland database is the most commonly used database by ML researchers. The dataset contains 303 instances and 76 attributes, but only 14 of them are referred by all published studies. The "goal" field which has varying values from 0(absence) to 4 denotes if heart disease present or not in the patient. Studies on the Cleveland database have focuses on distinguishing absence(value 0) from presence (values range from 1 to 4) [13].

The dataset has some missing values in it. Firstly missing values were filled with interpolation values. Then dataset was split into three parts: one for training (%70), second one for testing (%15) and third one for validation(%15). There are 213 instances and 13 attributes in training data. Test data and validation data contain 45 instances and 13 attributes.

Attribute information (only 14 used) is shown in Table 1:

TABLE I.
CLINICAL FEATURES AND THEIR DESCRIPTIONS.

Clinical Features	Description
Age	Age
Ca	Number of major vessels (0-3) colored by flourosopy
Chol(mg/dl)	Serum cholesterol
Cp	Chest pain type
Exang	Exercise induced angina
Fbs	Fasting blood sugar
Num	Diagnosis of heart disease
Oldpeak	ST depression induced by exercise relative to rest
Restecg	Resting electrocardiographic results
Sex	Gender
Slope	The slope of the peak exercise ST segment
Thal	3=normal ; 6 = fixed defect; 7= reversible defect
Thalach	Maximum heart rate achieved
Trestbps(mmHg)	Resting Blood Pressure

13 of the attributes listed above were used as input data for the network. The remaining attribute, num which is predicting value, was used as output data for the network. The num can get values between 0 and 4. Only 0 means absence of disease,

the others show presence of disease levels. So, output of network was designed as having two output type: 0 indicates that heart disease is absent and 1 indicates that heart disease is present.

B. Performance Evaluation

The performance of the proposed system was computed by different metrics like accuracy, precision and recall.

Accuracy is computed dividing number of predictions which are correct by number of all predictions. The obtained result is multiplied by 100 to get value as percentage.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (5)$$

where *TP, TN, FP* and *FN* demonstrate in order of the number of True Positives, True Negatives, False Positives and False Negatives. *TP* demonstrates the number of instances which are sick and diagnosed accurately. *FP* demonstrates the number of instances which are healthy and diagnosed wrongly as they are sick. *FN* demonstrates the number of instances which are sick but the instances are diagnosed wrongly. *TN* contains a number of instances which are healthy and the instances are diagnosed accurately.

Precision denotes the ratio of the instances that are predicted as having heart disease actually have heart disease.

$$Precision = TP/(TP + FP) \quad (6)$$

Recall denotes the proportion of the instances that are actually have heart disease are predicted as having heart disease.

$$Recall = TP/(TP + FN) \quad (7)$$

Accuracy, recall and precision were decided to express success of predicting heart disease system. Using only accuracy can be sometimes misleading. Sometimes selecting a model which has lower accuracy is desirable, because it provides more robust predictor for the problem. All predictions can be predicted as the value of majority class by model, when problem domain has a large class imbalance. This model is not useful when considered proble domain. This is referred to namely Accuracy Paradox. r such problems, classifiers should be evaluated with additional measures. [14]

To optimize network and to get better performance *pruning* which defines a set of techniques for trimming size of network by nodes was used. Hidden layer size of the network was changed from 3 neurons up to 12 neurons. The results related to hidden layer size are shown in Table 2.

TABLE II.
CLASSIFICATION PERFORMANCE WITHOUT PCA

Hidden Layer Size	Accuracy	Recall	Precision
3	82.222222%	85.714286%	78.260870%
4	75.555556%	77.777778%	66.666667%
5	84.444444%	86.206897%	89.285714%
6	75.555556%	85.185185%	76.666667%
7	82.222222%	84.000000%	84.000000%
8	86.666667%	86.956522%	86.956522%
9	71.111111%	73.913043%	70.833333%
10	86.666667%	89.473684%	80.952381%
11	77.777778%	80.952381%	73.913043%
12	84.444444%	86.956522%	83.333333%

To improve performance, dimensionality reduction with Principal Component Analysis (PCA) was done by reducing number of neurons of the input layer from 13 neurons to 8 neurons. The results which are obtained by changing hidden layer size with reduced dimensionality are shown in Table 3.

TABLE III. CLASSIFICATION PERFORMANCE WITH PCA

Hidden Layer Size	Accuracy	Recall	Precision
3	91.111111%	84.615385%	100.00000%
4	88.888889%	95.454545%	84.000000%
5	88.888889%	88.888889%	92.307692%
6	86.666667%	89.473684%	80.952381%
7	93.333333%	100.00000%	89.285714%
8	95.555556%	95.454545%	95.454545%
9	91.111111%	95.833333%	88.461538%
10	91.111111%	100.00000%	85.185185%
11	95.555556%	100.00000%	91.666667%
12	91.111111%	95.652174%	88.000000%

V. CONCLUSION

The proposed heart disease prediction system has been designed as a Multilayer Perceptron Neural Network. For the system Cleveland dataset was used. The neural network in the system used 13 clinical data which are obtained from Cleveland Dataset as input. It was trained with Backpropagation Algorithm in order to predict whether heart disease present or not in the patient.

There are a lot of studies on prediction of heart disease. Results of these studies vary up to almost accuracy of 100%.The proposed system gives 95% accuracy rate which means a very good rate according to related studies on this field. As a further study, the proposed methodology can be enhanced as a hybrid model with other classification algorithms in order to obtain more accurate diagnosis for heart disease.

REFERENCES

[1] "New initiative launched to tackle cardiovascular disease, the world's number one killer," World Health Organization.[Online].Available:http://www.who.int/cardiovascular_diseases/global-hearts/Global_hearts_initiative/en/. [Accessed: 03-Jul-2017].

[2] Z. Shen, M. Clarke, R. W. Jones and T. Alberti, "Detecting the risk factors of coronary heart disease by use of neural networks," Proceedings of the 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Societ, 1993, pp. 277-278. doi: 10.1109/IEMBS.1993.978541

[3] A. Dewan and M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2015, pp. 704-706.

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

[5] J. S. Sonawane and D. R. Patil, "Prediction of heart disease using learning vector quantization algorithm," 2014 Conference on IT in Business, Industry and Government (CSIBIG), Indore, 2014, pp. 1-5.doi: 10.1109/CSIBIG.2014.7056973

[6] M. G. Feshki and O. S. Shijani, "Improving the heart disease diagnosis by evolutionary algorithm of PSO and Feed Forward Neural Network," 2016 Artificial Intelligence and Robotics (IRANOPEN), Qazvin, 2016, pp. 48-53.doi: 10.1109/RIOS.2016.7529489

[7] S. A. Hannan, A. V. Mane, R. R. Manza and R. J. Ramteke, "Prediction of heart disease medical prescription using radial basis function," 2010 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, 2010, pp. 1-6.doi: 10.1109/ICCIC.2010.5705900

[8] S. U. Amin, K. Agarwal and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," 2013 IEEE Conference on Information & Communication Technologies, JeJu Island, 2013, pp. 1227-1231.doi: 10.1109/CICT.2013.6558288

[9] J. S. Sonawane and D. R. Patil, "Prediction of heart disease using multilayer perceptron neural network," International Conference on Information Communication and Embedded Systems (ICICES2014), Chennai, 2014, pp. 1-6.doi:10.1109/ICICES.2014.7033860

[10] S. Bashir, U. Qamar and M. Younus Javed, "An ensemble based decision support framework for intelligent heart disease diagnosis," International Conference on Information Society (i-Society 2014), London, 2014, pp. 259- 264.doi: 10.1109/i-Society.2014.7009056

[11] "Principles of training multi-layer neural network using backpropagation".[Online].
Available:
http://home.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html

[12] Tom M. Mitchell,"Artificial Neural Networks", in "Machine Learning", McGraw-Hill Science/Engineering/Math, 1997, pp:95-99.

[13] "Index of /ml/machine-learning-databases," Index of /ml/machine-learning-databases. [Online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names. [Accessed: 03-Jul- 2017].

[14] "Classification Accuracy is Not Enough: More Performance Measures You Can Use," Machine Learning Mastery, 06-Jun-2016. [Online]. Available: http://machinelearningmastery.com/classification-accuracy-is-not-enough- more-performance-measures-you-can-use/. [Accessed: 03-Jul-2017].