TuEngr Group

International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies

http://TuEngr.com

PAPER ID: 11A15F

EFFICIENT DIAGNOSTIC CARDIAC SYSTEM USING MACHINE LEARNING APPROACH

Mujtaba Ashraf Qureshi [1*], **Azad Kumar Shrivastava**[2]

[1] *Department of Information Technology, Mewar University, Chittorgarh (Raj), INDIA.*
[2] *Department of Computer Science, Mewar University, Chittorgarh (Raj), INDIA.*

ARTICLEINFO

Article history:
Received 19 May 2020
Received in revised form 10 August 2020
Accepted 24 August 2020
Available online 03 September 2020

Keywords:
Data mining; neural networks; feature selection; WEKA tool; Cardiovascular disease prediction; Cardiac disease prediction.

ABSTRACT

Heart disease is considered one of the ultimate threats to human life. To predict cardiac diseases in the early stages has become a challenge to medical science. Machine learning has acted as a rescuer to assist and develop various cardiac diagnostic systems. Data mining techniques mostly used as a synonym to machine learning plays an important role to mine useful knowledge. However, machine learning (ML) emphasis more on the prediction of diverse diseases. In this research work, three models are devised to predict cardiovascular diseases using artificial neural networks. Models are devised based on the application of a different number of hidden layers. The backpropagation algorithm is used to calculate the desired value by the adjustment of weights of the neurons in the network. In the very last stage of the experimental work performance measures of the three devised models are compared to reach the most efficient model.

Disciplinary: Computer and Information Technology, Cardiology (Cardiovascular Health and Disease).

1 INTRODUCTION

Data mining techniques are employed to extract valuable information from the distributed and voluminous databases by the application of machine learning, mathematical equations, and statistical methods. The suitable and useful diverse datasets remain unused in the absence of data mining technology and methodologies. In the present era, data mining technology accomplishes the principal role to swing medical science to modern approaches from the older traditional approaches. The field of data mining exists as a comprehensive technical field and consequently adopts some of the absolute techniques to analyze and predict lethal diseases. Supervised and unsupervised techniques are the two existing major branches of data mining technology. Supervised techniques are guided by the application of trained datasets and unsupervised datasets do not necessitate any training from the

*Corresponding author (Mujtaba A. Qureshi). Email: Mujtaba170@gmail.com outer world. Classification techniques fall in the category of supervised data mining techniques to classify the input data based on the previously supplied trained datasets. Classification techniques are more effective and realistic to diagnose various diseases predominantly cardiac diseases.

Heart disease is considered one of the ultimate and common threats to human life. As per the reports of the World Health Organization (WHO), more than 12 million people die every year because of cardiovascular diseases and will upsurge at an alarming pace if suitable measures are not taken. To diagnose cardiac diseases with high accuracy and at right time is a challenge to medical science. However, the evolution of data mining techniques transformed this problem sphere into the real domain. Very special attention is needed all over the world to devise the algorithms and models to diagnose cardiac diseases with acceptable results.

An artificial neural network (ANN) is a type of classification technique based on the reflection of the brain of humans [12]. This is designed as a layered structure to transform the supplied input data to the required output result. The very first layer is called the input layer and the final/last layer is called the output layer. The layers existing between the input and output layers are called hidden layers. Hidden layers play an important role to weaken network progress or strengthen the power of network results. In this research paper, three models are devised to predict cardiovascular diseases using artificial neural networks. Models are devised based on the application of a different number of hidden layers. In the very last stage of the experimental work performance measures of the three devised models are compared to reach the most efficient model.

An extensive literature survey is conducted related to the prediction of cardiac diseases using various mining techniques given subsequently. The work [1] used neural networks to develop a predictive system for heart diseases. This system gives the probability of heart diseases. In this experimental work, 14 attributes are used and some of the prominent are blood pressure, cholesterol, age, sugar level, smoking, obesity, etc. Neural networks (NN) shows approximately 100% accuracy for the prediction of heart diseases. The work [2] recommended an algorithm for the prediction of developing cardiovascular disease (CVD). The hybridization method is incorporated in this research work by the author and utilized the services of backpropagation and genetic algorithm. Further, they found that the neural network technique is recommended as an acceptable technique, particularly for the non-linear data. Backpropagation is the most used algorithm along with the ANN for training purposes. Backpropagation is used with ANN until the minimum difference is not attained between expected and obtained values. One of the disadvantages observed by the author is that ANN gets trapped in the local minima problem. The work [3] proposes a system by using neural networks and support vector machines to predict CVDs. SVM shows prediction with 80.41% accuracy whereas the multilayer perceptron (MLP) neural network shows 97.5% accuracy for the prediction of heart diseases. With data of 935 patients to predict and diagnose heart diseases by the application of neural network data mining techniques, the work applied radial basis function (RBF) and MPL of neural networks for prediction. This experimental work shows that RBF neural networks show 83% accuracy with ECG findings and an accuracy of 78% with clinical features. The work [4] performed an experimental approach to predict CVDs using neural networks, decision trees, and naïve Bayes data mining techniques. This system shows 82.5% accuracy. The work [5] developed a heart disease prediction system based on the neural network data mining technique (Feedforward NN). The developed system shows an accuracy of 90%. [7] developed a CVDs diagnosis system utilizing

ANN with a genetic algorithm. Dataset is collected from the online UCI repository for the experimental work. Experimental work is implemented using MATLAB GUI simulation tool for designing the system. The developed system shows an accuracy of 97.83. The work [6] developed a heart disease diagnosis system using SAS software. In this experimental work, the neural network data mining technique is used for the development of the model by conjoining from multiple previous models. An accuracy of 89.01% is achieved by the application of the Cleveland heart disease dataset. Also, specificity and sensitivity values attained are 95.91% and 80.95% respectively. The role of machine learning is explored in various renowned and distinguished fields such as robotics, medicine, academics, business, etc. The researcher has made effective attempts to apply various machine learning algorithms to improve the accuracy of the prediction of academics [10, 11].

1 FRAMEWORK

The design and workflow of the experimental work are depicted in Figure 1.

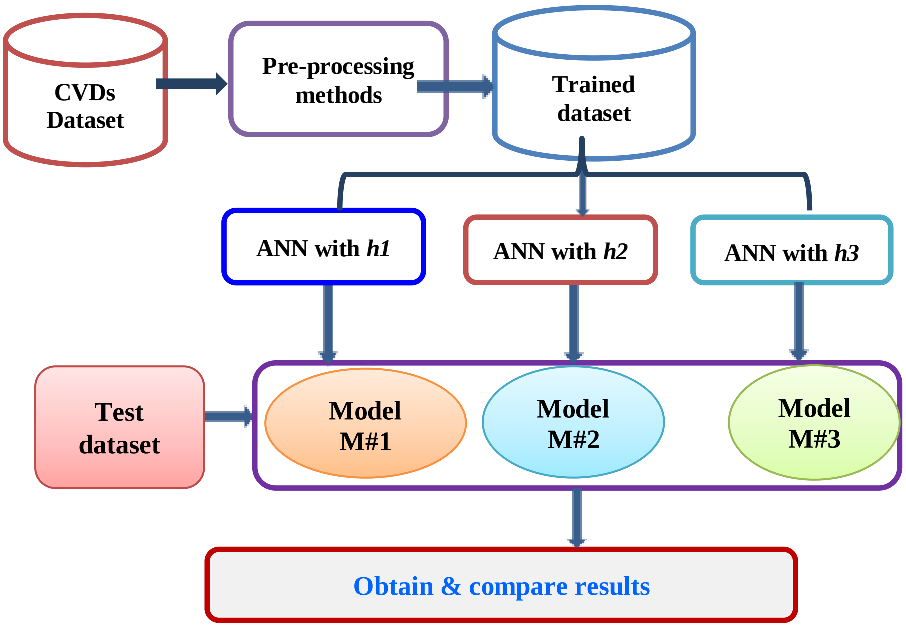

Figure 1: Framework for this study.

The framework revealed in Figure 1 is devised to implement the real-world application of the devised models. Three devised models in the framework are represented by model-M#1, model-M#2, and model-M#3. The framework also depicts *h1*, *h2,* and *h3* as one hidden layer, two hidden layers, and three hidden layers of model-M#1, model-M#2, and model-M#3 respectively.

1 ARTIFICIAL NEURAL NETWORK

The neural network consists of artificial neurons that mimic the neurons of the human brain. ANN is a mathematical representation or computational model that is stimulated by the organization and/or functional aspects of biological neural networks [8, 9]. ANN consists of an input layer, hidden layer/layers, and an output layer. The input layer receives an input of signals and transfers the same to the hidden layer/layers and finally, the processed signal/output is transferred to the output layer as the output of the query. Actually, the input layer receives raw data to forward to the hidden layer/layers. The hidden layer neurons perform their activity based on the input values, weights, and bias value. Finally, the output processed value is obtained via the output layer.

- The input layer receives values and forwards them to a hidden layer of the network.
- The values of the input layer are multiplied with some specified weight values and a constant bias value is
- The values from the hidden layer move to the output layer after some modification is performed using specified weight values.
- Finally, after processing the information, the output layer gives the desired output. During this process, an activation function is used to process the output data.

Typically backpropagation algorithm is used with the artificial neural networks to calculate error values and thus to obtain the desired values. BP allows us to change the weights of the network until the desired output is not obtained. The following diagram defines the basic structure of ANN:

The algorithmic flow of the applied technique is shown below in Figure 2.

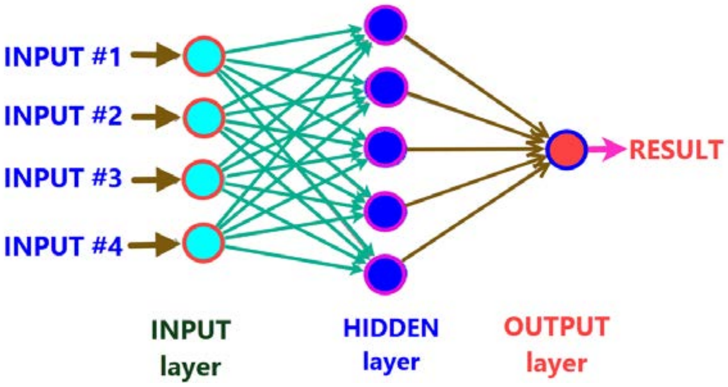

Figure 2: Neural Networks, with one hidden layer.

The final result of the neural network is achieved from the output layer. Equation (1) is used to calculate the output value of the network.

$$y_j = \sum_{f=1}^{k} w_{ij} x_i \qquad (1),$$

where

y_j is an output value of output neuron,

x_i is the value of input neuron,

w_{ij} is connecting weight between x_i and y_j,

$\sum$ is the final sigmoidal function.

2 DATASET

The efficacious output of the network is very much exaggerated by the nature and lucidity of the employed datasets. The dataset employed to perform the experimental approach is acquired from the online available Cleveland database. The scholar/reader of this paper may access to the database related to heart disease data as also https://archive.ics.uci.edu/ml/datasets/Heart+Disease. Cardiac related datasets are accessible in sufficient expanse. Preprocessing methods and feature selection techniques available in the WEKA simulation tool are assimilated to accomplish the high profile attributes only to train and testify the developed models. The selection of only 13 high profile attributes is prepared which consists of 500 instances only. An assessment of the selected attributes is offered in Table 1.

Table 1: Attributes used to predict Heart Diseases.

S.NO	Attributes	Description
1	age	In years
2	trestbp	Resting blood pressure
3	Sex	Male or female
4	Bp	Chest pain type
5	chol	Cholesterol level in mg/dl
6	Fbs	Fasting blood sugar
7	restecg	Electrographic results at rest.
8	Thalach	Maximum Heart Rate Achieved.
9	exang	Exercise-Induced Angina
10	Oldpeak	ST Depression Induced By Exercise.
11	slope	The slope of the Peak Exercise ST Segment.
12	numvessels	No. of vessels colored by fluoroscopy
13	thal	Type of Defect of Heart.
14	diagnosis	Absence or presence of disease.

3 EXPERIMENTAL APPROACH

To stretch the real-world character to the proposed framework, the Waikato Environment knowledge Analysis simulation tool (WEKA) is used. The processed dataset is divided into training and testing datasets in the ratio of 70:30. ANN is used to devise the three models. Models differ from one another by the number of hidden layers incorporated. Models are labeled as model-M#1, model-M#2, and model-M#3. Model-M#1 has one hidden layer (*h1*), model-M#2 has two hidden layers (*h2*) and model-M#3 has three hidden layers (*h3*) integrated. Thus model-M#1 consists of an input layer, an output layer, and one hidden layer. Model-M#2 consists of an input layer, an output layer, and two hidden layers.Model-M#3 consists of an input layer, an output layer, and three hidden layers.

Trained or labeled datasets are supplied as input to the network of all the three models to make them competent enough for the prediction of diseases. All three developed models based on the ANN can predict heart diseases. So we supplied an unlabeled dataset to each of the three developed models to achieve the performance results shown. A comparative study of the predictive results depicted by the models is performed to reach a more efficient and accurate model for cardiovascular prediction. Six measures of results are taken into contemplation to relate and investigate the capability of the developed heart disease prediction models. Description of the performance measures taken into consideration for the comparative study includes

1. Accuracy is defined as how well our developed model/classifier is performing to divide tuples into respective classes.

$$\text{Accuracy} = (TP+TN)/\,(TP+TN+FP+FN) \qquad (2).$$

Machine learning considers the true positive (TP, equivalent with hit), true negative (TN, with correct rejection), false positive (FP equivalent with a false alarm, Type I error), false negative (FN equivalent with miss, Type II error).

2. Precision refers to the measurement of accurateness or percentage of tuples labeled as positive or negative are actually as such by the developed classifiers.

$$\text{Precision} = (TP)/ (TP+FP) \tag{3}$$

3. Sensitivity refers to positive tuples correctly labeled by the classifier or to identify appropriately those individuals actually involving in diseases. Sensitivity is also referred to as a true positive rate.

$$\text{Sensitivity (Recall)} = (TP)/ (TP+FN) \tag{4}$$

4. Specificity is defined as the number of negative tuples correctly labeled/identified.

$$\text{Specificity} = (TN)/ (TN + FP) \tag{5}$$

5. Elapsed Time of Training is the time taken by the classifier to predict CVDs.

4 RESULTS

The outcome of the developed models is presented. An experimental approach is conducted using the WEKA simulation tool. The training and testing phase is conducted for all the three models using the online processed Cleveland datasets having the selected attributes given in Table 1. The results acquired for the prediction of cardiovascular diseases are recorded. Results of model M#1, M#2, M#3 are revealed and compared in Figures 3.

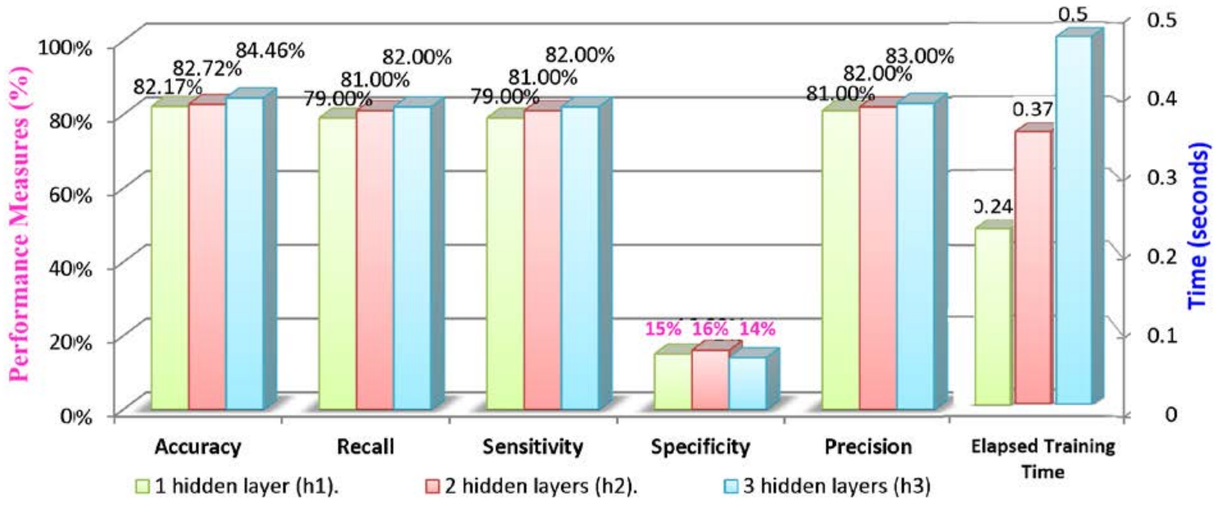

Figure 3: ANN performance measures of models M#1, M#2, and M#3.

The practical approach presents a very much influence upon the performance measures of the developed models using hidden layer number variation. The six most prominent performance measures cited as accuracy, recall, sensitivity, specificity, precision, and elapsed training time of the developed models are taken into attention for the prediction of cardiovascular diseases. Emphasis is specified to perceive that as the number of the hidden layers is augmented from one hidden layer (h1) to three hidden layers (h3), accuracy, recall, sensitivity, and precision increases for the prediction of heart diseases. Also to attain the decline in the percentage of specificity as the number of hidden layers is enhanced from one layer to three layers is considered the exceptional success of the models. However, the undesirable influence is observed, as the number of hidden layers increases, the training time to build the model also upsurge.

5 CONCLUSION

Three cardiovascular predictive models are developed using artificial neural network techniques. Models differ based on the number of hidden layers.

Model M#1 is devised using one hidden layer (*h1*), model M#2 using two hidden layers (h2) and model M#3 is developed using three hidden layers (*h3*). An empirical and comparative study is conducted to verify the efficiency and productivity of the said models. Performance measures taken into consideration illustrates gradual enhancement as the number of hidden layers is varied from one hidden layer (*h1*) to three hidden layers (*h3*). Performance measure outcome presented that the model M#3 is out-performed but specificity, with the highest elapsed time.

This work witnesses that as the number of the hidden layers is augmented from one to two hidden layers and from two to three layers, performance measures revealed by the developed models also increases continuously. One of the fundamental principles of this world is that advantages are always interconnected with disadvantages. Similarly in this research work, time taken by the predictive models also increases as the number of hidden layers is augmented to diagnose heart diseases. In the future, real and live datasets with large sizes can be used to train and testify the predictive models. Moreover, the size of the hidden layers can be intensified to several hidden layers to verify the effectiveness of the models.

6 AVAILABILITY OF DATA AND MATERIAL

The corresponding author will be liable to provide information regarding this paper.

7 REFERENCES

[1] C.S. Dangare, S. Apte. A Data Mining Approach for Prediction of Heart Disease Using Neural Networks. *International Journal of Computer Engineering and Technology, 2012*, 3(3).

[2] A.Dewan, M. Sharma. Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification. *2nd International Conference on Computing for Sustainable Global Development, 2015*, 704-706.

[3] S.A.Pattekari and A.Parveen. Prediction system for heart disease using naive Bayes. *International Journal of Advanced Computer and Mathematical Science, 2012*, 3(3), 290-294.

[4] B.S. Rao, K.N. Rao, S.P. SETTY. An Approach for Heart Disease Detection by Enhancing Training Phase of Neural NetworkUsing Hybrid Algorithm. *Advance Computing Conference, 2014*, 1211-1220.

[5] Vanisree K, J. Singaraju. Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks. *International Journal of Computer Application, 2011*, 19(6).

[6] R. Das, I. Turkoglu, A. Sengur. Effective Diagnosis of Heart Disease through Neural Network Ensemble. *Expert Systems with Applications, 2009*, 36(4), 7675-7680. DOI: 10.1016/j.eswa.2008.09.013.

[7] P.Gupta, B. Kaur. Accuracy Enhancement of Artificial Neural Network using Genetic Algorithm. *International Journal of Computer Applications, 2014*, 103(13).

[8] Rani, K. U. Analysis of heart disease dataset using a neural network approach. 2011.

[9] Tu, J.V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes”. *Journal of clinical epidemiology, 1996*, 49(11), 1225-1231.

[10] Ashraf, Mudasir, Majid Zaman, and Muheet Ahmed. "To Ameliorate Classification Accuracy Using Ensemble Vote Approach and Base Classifiers." *Emerging Technologies in Data Mining and Information Security*, Springer, Singapore, 2019. 321-334.

[11] Ashraf, M., Zaman, M., & Ahmed, M. Performance analysis and different subject combinations: An empirical and analytical discourse of educational data mining. *8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2018.

[12] Nang Y. *The Handbook of data mining*. Lawrence Erlbaum Associates, 2003.

Mujtaba Ashraf Qureshi is a Ph.D – I.T scholar of Department of Information Technology at Mewar University, Chittorgarh, Rajasthan, India. He got his master’s degree in Information Technology from Punjab Technical University (now I.K.Gujral Punjab Technical University), Jalandhar, Punjab, India. He is interested in Data Mining Technology.

Dr.Azad kumar Shrivastava is Professor at the Department of Computer Science, Mewar University, Chittorgarh, Rajasthan, India. He got his doctoral degree from ‘Atal Behari Vajpayee-Indian Institute of Information Technology and Management’, Gwalior, Madhya Pradesh, India. His research includes Artificial Intelligence, Neural Networks, Machine Learning, Deep Learning and Big data on CPU & GPU Cluster for DWH & IOT.

Trademarks Disclaimer: All product names including trademarks™ or registered® trademarks mentioned in this article are the property of their respective owners, using for identification and educational purposes only. The use of them does not imply any endorsement or affiliation.