IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol-6, Special Issue-4, May 2018 E-ISSN: 2347-2693

Heart Disease Detection using Autoencoder

D. Rajeswari[1]*, K. Thangavel[2]

[1,2]Deptartment of Computer Science, Periyar University, Salem, India

*Corresponding Author: raji.danapal@gmail.com, Tel.: +91-9894811157

Available online at: www.ijcseonline.org

***Abstract*—** Early detection of heart disease can be achieved by high disease prediction and diagnosis efficiency. Machine learning techniques can help the medical expert in decision making for providing the best treatment. In this paper, an autoencoder neural network classifier is developed for the classification of heart disease medical data sets. The autoencoder were trained to properly classify the clinical data. The proposed classifier is tested on heart disease data sets namely Cleveland and Statlog obtained from the UCI repository and also compared with conventional classification techniques namely Support Vector Machine, Random Forest, K-Nearest Neighbour, Naïve Bayes to concerning its outperformance. Experimental results show that the autoencoder neural network classifier offers much better classification accuracy, precision, recall and f-measure rates when compared with other conventional methods. The proposed method presents itself as an easily accessible and cost-effective alternative to traditional machine learning methods which are used for the diagnosis. In this study, the implementation of the developed model can potentially support in reducing heart disease among patients.

Keywords— Artificial Neural Networks, Autoencoder, Heart disease, Classification

I. INTRODUCTION

Heart disease (HD) is one of the fatality diseases and kills more than 370,000 people every year all over the World [1]. It is the first leading cause of death around the World and 1 out of 3 deaths globally are as result of premature heart disease. By 2030 this is estimated to rise by 22% [2,3]. The effectiveness of early diagnosis is to reduce mortality among patients with heart disease. It is estimated that for every 43 seconds, one person affects from a heart disease. The indicators of diagnosing heart disease are based on some biochemical risk factors such as smoking, cholesterol, physical obesity, blood pressure, blood sugar level, etc.

Detection of HD is an important research problem in health care domains and medical communities. In this study, the demographic characteristics of individuals are used to find the risk of HD and the model is created using Autoencoder Neural Networks (AE-NN) which have been applied to classification [19]. This approach is used extensively in real life applications, especially for decision support system in the medical diagnostic system for earlier detection of some serious disease.

Various machine learning techniques such as Random Forest (RF), Naïve Bayes (NB), etc., each with its unique merits and demerits, have been presented to predict the heart disease outcomes and these traditional methods can easily applied to data sets. However, these methods cannot produce significant classification performance [4-7].

Artificial Neural Network (ANN) is one of the soft computing techniques, inspired by the working mechanism of the biological nervous system such as brain. It is the novel structure of the information processing system. In general, ANN consist of input, hidden and output layers that are interconnected each other and each layer is made up of a number of processing elements called neurons (nodes or units).

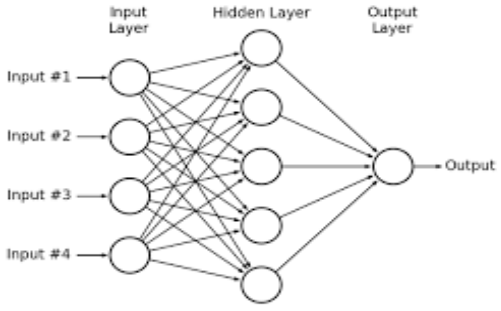

Figure .1. Structure of Artificial Neural Network

The input layer receives the data from the external world or nearby surrounding neurons and each input is multiplied by its corresponding weights. The processing elements are inputs, weights, activation function and one output. The activation function has a set of transfer function which is used to get the output. The performance of neural network depends on the transfer function, architecture itself and learning rule.

The overall organization of this paper is as follows. Background study of neural network models for disease classification and diagnosis are outlined in Section II. Section III describes the methodology and dataset used in this study. Experimental results analysis and the performance of proposed AE-NN classifier and different traditional machine learning methods are compared with AE-NN Classifier in Section IV. This paper concludes in section V, with direction for further research.

II. RELATED WORK

Baldassare et al. [8] used artificial neural network technology to recognize patients at high risk of cardiovascular diseases and obtained an accuracy of 92%. Pelazeo et al. [9] presented a multilayer perceptron to classify ischemic heart disease using ECG analysis and the model provides an accuracy of 88.4%. Adetiba et al. [10] in their paper, automated heart defect detection model for athletes by different ANN architectures were designed and trained with ECG data with an accuracy of 90%. Resul Das et al. [11] proposed a neural network which was based on ensemble method for diagnosing heart disease and obtained 89% accuracy. Arabasadi et al. [12] proposed a highly accurate hybrid method of NN and Genetic algorithm for the diagnosis of coronary artery disease. Guillermo et al. [13] proposed a Radial Wavelet Neural Network (RWNN) classifier for heart murmur with high performance accuracy. Lisboa et al. [14] studied the benefits of neural networks using in the health care sector for decision support system. Yan et al. [15] used a multilayer perceptron-based diagnosis system and achieved the rate of 90% accuracy. Acharya et al. [16] designed ANN and fuzzy equivalence relations used for determine the heart rate variability which yields the classification over 90%. Shi et al. [17] conducted study on backpropagation neural network to diagnose cardiac disease and myocardial infraction. Chatur et al. [18] used ANN for the prediction of thyroid disease with accuracy greater than 90%.

From these works, it is observed that that the variants of neural network have been used to diagnosis the different disease. It is observed that no researcher has used AE-NN to process the heart disease dataset to diagnosis. In this paper proposes and analyse a novel method which is based on AE models to classify the disease status from the observation of clinical data. The method is then compared with conventional algorithms that have been applied in HD detection problem

III. METHODOLOGY

Autoencoder is the feed forward neural network, where the network is trained to recall the inputs [19]. The AE comprises two parts namely encoder and decoder. It sets the dependent values to be equal to the inputs, it uses $y^{(i)} = x^{(i)}$. The autoencoder tries to learn the identity function through a function $h_{W,b}(y)=x$, so as to output y that is similar to x.

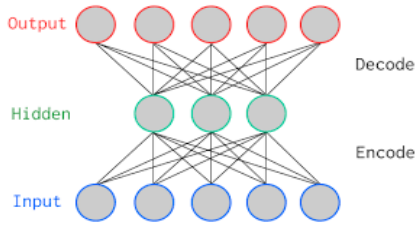

Figure.2. The structure of the autoencoder

The features of an individual, represented by a vector {m}, to one of the diagnosis classes, H_1 or H_2 (where H_1, H_2 represents the status of an individual, which may be Healthy or Unhealthy). The outcome is represented as variable n. The classification model maps the inputs m_1, m_2, . . ., m_n to the output n. A mathematical function describes this mapping, can be written as follows:

$$\{n\} = f(\{m\}, \{w\}). \quad (1)$$

Here {w} is the weights and {m} represents the demographic input parameters and {n} represents the diagnosis status.

Heart Disease Classification using Autoencoder Neural Networks

AE-NN are successfully employed as pre-trained layers of neural networks for classification tasks [20,21]. The network comprises of an input layer, representing input features, mapped to an output layer representing the same features as the input layer via the hidden layer. The network was trained to recall itself (predict the feature inputs shown in Fig.3. One of the input nodes m_1, which was ultimately represented by one of the output nodes, n_1, as well. The neural network equation can be written as in eq. (1). Since the network is trained to recall the input features, the output vector {n} (predicted feature properties) determined will be approximately same as the input vector {m} (actual feature properties). However, an error, exists between the input vector {m} and the output vector {n}, which can be expressed as the difference between the input and output vector. This error is formulated as:

$$e = \{m\} - \{n\}. \quad (2)$$

Substituting for {n} from eq. (1) into eq. (2) we get

$$e = \{m\} - f(\{m\}, \{w\}). \quad (3)$$

An error, exists among the predicted disease status (output vector) {n} and the individual's actual disease status (target vector) {t} during training, which can be expressed as the variance among the target and output vector. For the neural

network heart disease classification, the mean square error function among target output vector {t} and the output vector {n}. In this study, a non-negative and minimum error is required. This can be attained by squaring the error function in eq. (3).

$$e=(\{m\}-f(\{m\},\{w\}))^2 \quad (4)$$

The goal is to learn and obtain a feature expression $h(m,W,b)=\sigma((Wx+b)$,m=m_1,m_2,m_3,…..m_n at the hidden layer so that the output σ (W^Th(m,W,b)+c) is close to or refactoring the input, where W denotes the weights between the input layer and the hidden layer and b is the biases. The hidden layer controls the number of active neurons. In practice, if the output of a neuron is close to 1, the neuron is considered to be "active", otherwise it is "inactive".

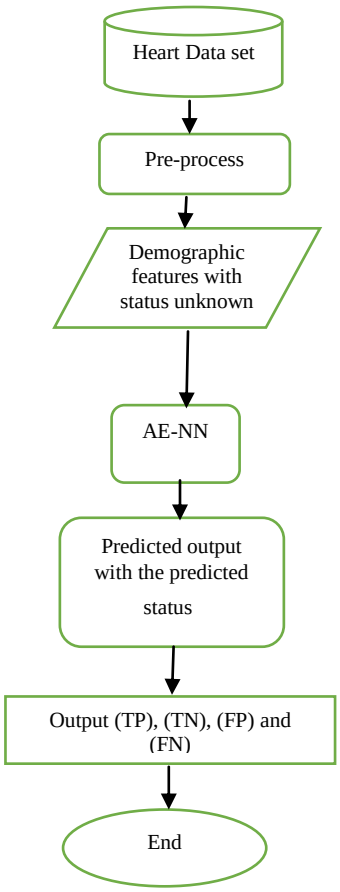

Figure 3: Flow chart of Proposed AE-NN

Datasets

The Cleveland and Statlog heart disease dataset are obtained from the website of the UCI (University of California Irvine). There is a total of 303 and 270 training instances, 13 predictors and 1 class label included in these datasets.

Table 1. Dataset Description

Name	Type	Description
Age	Numeric	Age in years
Sex	Nominal	1 = Male 0 = Female
Fbs	Nominal	Fasting blood sugar > 120 mg/dl: 1 = true,0 = false
Restecg	Nominal	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2=showing probable or define left ventricular hypertrophy by Estes 'criteria
Thalach	Numeric	Maximum heart rate achieved
Exang	Nominal	Exercise induced angina: 1 = yes 0 = no
Slope	Nominal	The slope of the peak exercise segment: 1 = up sloping, 2 = flat, 3= down sloping
Diagnosis	Nominal	Diagnosis classes: 0=Healthy 1=Possible heart disease.

The excellence of the classifier is measured based on the accuracy value as in Eq. (5).

$$\text{Accuracy} = \frac{tn+tp}{tn+fn+tp+fp} * 100 \quad (5)$$

In addition to accuracy, we consider three other performance measures: Precision, Recall, F-Measure (Eq.(6), Eq.(7), Eq(8)) are shown below:

$$\text{Precision} = \frac{tp}{tp+fp} * 100 \quad (6)$$

$$Recall = \frac{tp}{tp+fn} * 100 \quad (7)$$

$$Fmeasure = \frac{precision * recall}{precision + recall} * 2 \quad (8)$$

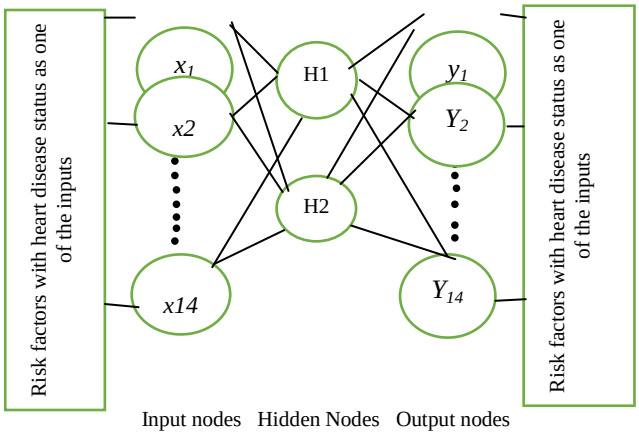

Figure 4: Autoencoder neural network architecture

Algorithm: Autoencoder
Input
The labelled dataset $D_{L,}$ with input X
Output
The predicted labels of samples in D, at the output layer to obtain an output X'.
Step 1: Pretrain the autoencoder with samples D.
Step 2: Initialize the weights of the autoencoder.
Step 3: Train a prediction model with the hidden representation of D_L and compute activations at the hidden layer.
Step 4: Apply it on the hidden representation of D.

Figure 5: Algorithm for Autoencoder

IV. RESULTS AND DISCUSSION

Table 2 presents the result of comparison of the four conventional machine learning algorithms and autoencoder neural network when applied to Cleveland and Statlog datasets. It is observed from Table 2 and Figure 6 that the proposed approach AE-NN produced the highest accuracy of 97.3% and 98.6% for Cleveland and Statlog dataset. AE-NN approaches show better performance with respect to classification accuracy.

V. CONCLUSION

In this paper, an autoencoder classifier was proposed for the detection of heart disease. The results showed that the proposed Autoencoder Neural Network classifier outperformed than the other benchmark methods such as Support Vector Machine, Random Forest, K-Nearest Neighbour, Naïve Bayes for both Cleveland and Statlog datasets. The experimental results showed that AE-NN system predicts heart disease efficiently. Further, the influencing parameters/factors are to be identified for the early detection of heart disease and this is the direction for the further research.

Table 2: Classification accuracy, Precision, Recall and F-Measure for SVM, RF, KNN and NB with AE-NN

Classifier models	Performance Assessment	Dataset	
		Cleveland	Statlog
AE-NN	Accuracy	97.3	98.6
	Precision	96.5	98
	Recall	96.2	97.8
	F- Measure	96.3	97.9
SVM	Accuracy	90.3	94.5
	Precision	89.2	93
	Recall	94	92.9
	F- Measure	91.5	93
RF	Accuracy	90.3	95.4
	Precision	89.2	94.6
	Recall	94	93.8
	F- Measure	91.5	94.1
KNN	Accuracy	75.9	78.8
	Precision	77.2	80
	Recall	78.6	82.6
	F- Measure	77.8	81.2
NB	Accuracy	85.9	85.4
	Precision	86.3	84.4
	Recall	88.6	89.6
	F- Measure	87.4	86.9

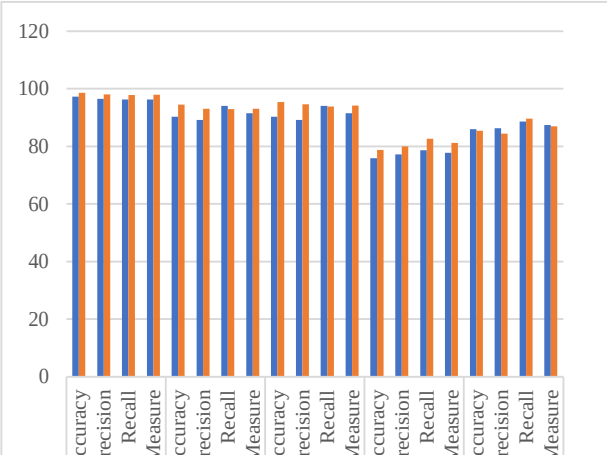

Figure 6: Performance analysis chart

Table 3: Characteristics of the dataset

Dataset	No of attributes	No of instance	Training set	Testing set
Cleveland	13	303	223	80
Statlog	13	270	192	78

ACKNOWLEDGMENT

The research work of the first author is supported by the University research fellowship of Periyar University, Salem.

REFERENCES

[1] Centres for Disease Control and Prevention (CDC), Heart disease in the United States, Available http://www.cdc.gov/heartdisease/facts.htm.

[2] Fact sheet: Cardiovascular disease in India, World health federation.

[3] World Health Organization. The World Health Report 2002. Geneva, Switzerland: WHO, 2002.

[4] M.Ciecholewski, "*Ischemic heart disease detection using selected machine learning methods*". International Journal of Computer Mathematics, Vol. 90, No.8, pp. 1734–1759, 2013.

[5] E.A. Zanaty, A. Afifi, "*Support Vector Machines (SVMs) with universal kernels".* Applied Artificial Intelligence, Vol. 25, pp. 575–589, 2011.

[6] K. Kalaiselvi, K. Sangeetha, S. Mogana, "*Efficient disease classifier using data mining techniques: Refinement of random forest termination criteria",* IOSR-JCE, Vol. 14, pp. 104-111, 2013.

[7] A.K. Ghosh, "*On optimum choice of k in nearest neighbour classification*", Computational Statistics & Data Analysis, Vol. 50, pp. 3113 – 3123, 2006.

[8] D.Baldassare, E.Grossi, M. Buscema, M. Intraligi, M.Amato, E.Tremoli, L. Pustina, S. Castelnuovo, S. Sanvito, L.Gerosa, S.Sanvito,L.C.Egros, "*Recognition of patients with cardiovascular disease by artificial neural networks*", Annals of Medicine Vol. 36 (8), pp. 630-640, 2004.

[9] J.I.Peláez, J.M.Doña, J.F.Fornari, G.Serra, *"Ischemia classification via ECG using MLP neural network*", International Journal of Computational Intelligence Systems, Vol. 7(2), pp.344-352, 2014.

[10] Adetiba, E. Iweanya, V.C. Popoola, S. I.Adetiba, J. N.Menon, "*Automated detection of heart defects in athletes based on electrocardiography and artificial neural network*", Cogent Engineering, Vol. 4, Issue.1, pp.1-21, 2017.

[11] R. Das, I. Turkoglu, A. Sengur, "*An Effective diagnosis of heart disease through neural network ensembles*", Expert Systems with Applications, Vol. 36, No. 4, pp. 7675-7680, 2009.

[12] Z. Arabas Adi, R. Alizadeh Sani, M. Rosha Zamir, H. Moosaei, A.A. Yarifard, "*Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm",* Computer Methods and Programs in Biomedicine, Vol. 141, pp. 19-26, 2017.

[13] J.E. Guillermo, J. Luis, Castellanos, E. Sanchez, A.Y. Alanis, "*Detection of Heart Murmurs Based on Radial Wavelet Neural Network with Kalman Learning*", Neuro Computing, Vol.164, pp. 307-317, 2015.

[14] P.J. Lisboa, "*A review of evidence of health benefit from artificial neural networks in medical intervention*". Neural Networks, Vol. 15, No.1, pp. 11–39, 2002.

[15] H.Yan, Y.Jiang, J.Zheng, C.Peng, Q.Li, "*A multilayer perceptron-based medical decision support system for heart disease diagnosis"*, Expert Systems with Applications, Vol. 30, No. 2, pp. 272-281, 2006.

[16] R. Acharya, P. Subbanna Bhat, S.S. Iyengar, A. Rao, S.Dua, "*Classification of heart rate data using artificial neural network and fuzzy equivalence relation",* Pattern Recognition, Vol. 36, No. 1, pp. 61-68, 2003.

[17] L.Shi, Z. Zhang, L. Wang, J. Zhang, "*The Aide Diagnosis of Cardiac Heart Disease Using a Deoxyribonucleic Acid Based Backpropagation Neural Network"* International Journal of Distributed Sensor Networks, Vol. 5, Issue.1, pp. 38, 2009.

[18] P.N.Chatur, A.Ghatol, FIETE, "*Thyroid Disease Recognition Using Artificial Neural Network* " IETE Technical Review, Vol. 17, No. 3, Issue.3, pp.143-145, 2000.

[19] B.L. Betechuoh, T. Marwala, T. Tettey, "*Autoencoder networks for HIV classification*", Current Science, Vol. 91, No. 11, pp. 1467-1473, 2006.

[20] R. Hata, M.A.H. Akhand, K. Murase, *"Multi-valued autoencoders and classification of large-scale multi-class problem"*, International Journal of Advanced Computer Science and Applications, Vol. 8, No. 11, pp. 19-26, 2017.

[21] F.M.Alakwaa,Kumar D.Chaudhary, L.X.Garmire, "*Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data*" J Proteome Res, Vol. 17(1), pp.337-347, 2017.

Authors Profile

Mrs.D. Rajeswari received B.Sc. degree in Computer Science from Bharathiar University, India in 2007. She completed her M.Sc. in Information Technology from Anna University, India in 2009 and M.Phil. degree in Computer Science from Periyar University, India in 2013. She is currently a research student working for her Ph.D. under the guidance of Dr. K. Thangavel. Her research interests focus on the Machine Learning Techniques for Health care problems.

Dr. K. Thangavel is a Professor at the Department of Computer Science, Periyar University, Salem, India. He received M.Sc. and M.Phil. degrees in Mathematics from Bharathidasan University, India during during in 1986 and 1987 respectively. He received Ph.D. degree in Mathematics in 1999 from the Gandhigram Rural University, India. He received the M.C.A degree in 2001 from Madurai Kamaraj University, Madurai. He started his teaching career as a Lecturer in Mathematics at Gandhigram Rural University in 1989 and promoted as Reader in Mathematics in 1999. He joined the Periyar University as Professor of Computer Science in 2006. He has published more than 300 research articles in journals of international repute and successfully completed 3 major research projects funded by UGC, New Delhi. He was awarded the Tamil Nadu State Scientist Award in 2009 by the Govt. of Tamil Nadu and Sir C.V. Raman Award in 2013 by Periyar University for his outstanding contribution in research. His major research work focuses on Data Mining, Image Processing, Soft Computing, Bio-Informatics Optimization Algorithms and Pattern Recognition.

