Investigation of Hepatitis Disease Diagnosis using Different Types of Neural Network Algorithms

Barakat Saeed Alshamrani† and Ahmed Hamza Osman††

†, †† King Abdulaziz University, Department of Information System, Faculty of Computing and Information Technology, Jeddah, Kingdom of Saudi Arabia

Summary
The accessibility of large amounts of medicinal data in clinics and hospitals pointers to the focus on reliable information analysis software to exploit useful information. Many tools tried to diagnosis hepatitis disease but still there is a deficiency of analyzing the biological data of Hepatitis illness in the world, where millions of people are killed in the world by this disease. This research aims at investigating the neural network algorithm for hepatitis disease. The data mining processes applied on the UCI dataset. Our investigation model examined different types of neural network algorithms (Quick, Multiple, Dynamic and RBFN) with different factors such as data size, learning cycle, and processing time to achieve the diagnosis accuracy and estimated error. The Multiple neural networks proved the best performance compared with Quick, Dynamic, and RBF neural network algorithms.

Key words:
Hepatitis; Neural Network; Classification; Diagnosis; Accuracy; Cycle

1. Introduction

Hepatitis is a universal health challenge-attacking people every year and lading to the death. Hepatitis has indistinct symptoms. They are different types of hepatitis diseases named A, B, C, D, E, and G [1].
Data mining approaches have been extensively employed for hepatitis diagnosis diseases to help a medicine doctors to make an accurate diagnosis decision. By assistance of analytic schemes, the probable mistakes in doctors' ouccred in the phase of diagnosis can be reduced. In addition, the medicinal data can examined in short period too [2]. Many studies had been touched a hepatitis classification and feature reduction using genetic algorithm [3] and simulated annealing [4] or statistical methods using (LDA) [5]. In the predication stage, hepatitis information that is obtained from hepatitis patients is used as an input to a classification system, for instance neural network (NN) algorithm [6,7, 8], SVM technique [2,4] artificial immune system [9], and fuzzy logic [5]. In hepatitis classification studies, Ster and Dobnikar achieved predication diagnosis achievements of 83.2%, 85.3%, and 86.4%, using FDA and LDA techniques [10]. Polat and Günes¸ obtained a diagnosis accuracy of 94.14% in their research by Artificial Immune System and Personal Component Analysis [10]. Dogantekin et al. achieved 94.16% of accurate diagnosis by using combined framework between ANFIS and LDA [5]. Javad et al. achieved a combined technique using simulated annealing-SVM. Their diagnosis reported accuracy of 96.25% [4]. Chen et al. obtained a diagnosis accuracy of 96.77% based on mixed SVM and LFDA [5]. In this study, a Neural network algorithms (Quick, Multiple, Dynamic and RBFN) is examined for diagnosis the hepatitis diseases based on training and testing data size, learning cycle, and processing time to achieve the diagnosis accuracy and estimated error. Our research investigated the neural network algorithms on UCI Hepatitis dataset to help doctors in diagnosis and decision making. The content of this research was presented as follows. Neural network algorithms will be presented in the next section. Section III reports the methodology and operational framework of this study. Section IV discusses the dataset while section V explain the proposed model and experimental results. The last section discusses the conclusion of this study.

2. Neural Net Algorithms

A Neural Network algorithm (NN) is a model of information processing that is inspired of the brain biotic nervous schemes of the brain. The main component of this model is the original construction of the information-processing scheme. It is composed of a huge number of interrelated processing neurons employed in unison as a solution of certain challenges. NNs, similar to users, learn by the sample. An NN is organized for certain filed, for example dataset classification or pattern recognition, via a learning procedure [11].
The structure of neural networks is a biological structure. The beginning point for the neural networks algorithms is a neuron unit, as in Figure 1. This neuron contains multiple variables as inputs and a single variable as output. The input is adjusted by weight, which multiplies by the input values. The neuron will mix the input weights and, concerning a threshold value and activation function, use these to control

Manuscript received February 5, 2017
Manuscript revised February 20, 2017

its output. This pattern follows carefully our consideration of how real neurons work [12].

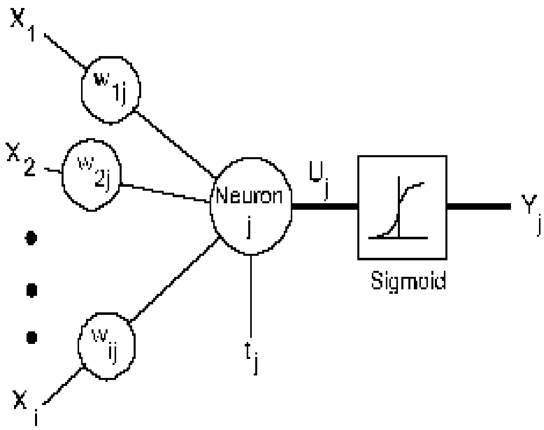

Fig. 1 Neural network structure

Some study tried to explain different types of neural network algorithms [12] such as:

- **RBFN Method** – A radial basis function network (RBFN) is a superior type of neural network. It compoise of an input-layer, a hidden-layer, and an output-layer. The hidden layer contains neurons that represent groups of input features, and these groups are working based on radial basis functions.

- **Quick Method** – This type used a single neural network in the training phase. The quick network has a single hidden layer consisting neurons. The input neurons are cross ponding the output neurons. The back-propagation technique used in the training phase. The quick extract minor hidden layers that are quicker training phase and simplify better.

- **Dynamic Method** – The structure of the Dynamic network modifies during the training phase, with added some neurons to enhance the output till the network obtained the best result. There are two main steps learning of dynamic method; searching for the structure, and learning the ending network.

- **Multiple Method** – This method trained pseudo-parallel style. Every particular network is started, and all networks are learned. This type highest the accuracy and returned as the final model when the criterion stop and met all networks.

3. Methodology and Operational Framework

Our operational framework provides a structured manner that was used to help investigate and examined hepatitis based on the different types of neural network algorithm. The operational framework of this study is an organization with several phases. This phase is demonstrated in Figure2 bellow.

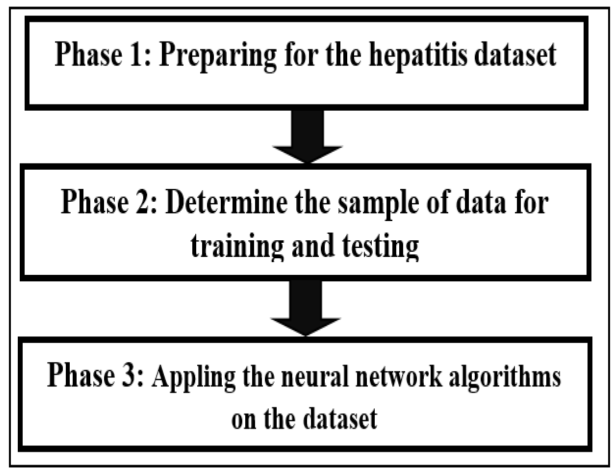

Fig. 2 Operational framework

Phase 1: Preparing the data set
In this phase, we applied pre-processing steps for data mining, which includes data cleansing, outlier values removal, and missing values solving. The pre-processing steps include data cleansing, outlier values removal (by removing the illogical values from the dataset, for example, the Age = 300) and missing values solving (?) by two ways [13]:

1) Remove the row from the dataset (if the number of missing values in less).

2) Calculate the average values of each feature with and the replace it with the missing value (Numeric values).

Counting the number of 0s (1) and 1s (2) in each feature and then replace the high values from 0 and 1 with the missing value in each feature individually.

Phase 2: Divided the data set to training and testing data.
After the data mining pre-processing steps, we then divide the dataset into two sets; the first set is learning dataset, and the data distribution ratio are 50%, 60%, and 70%. The second group is the testing data, and the proportion of the data is 50%, 40%, and 30% respectively.

Phase 3: Applying and investigate the Neural Network algorithms for hepatitis dataset across different factor (Cycle, data size, and processing time). This phase will be discussed in section V.

4. Dataset

In this study, we used the hepatitis diseases dataset collected from UCI machine learning repository [13]. The Dataset is a defined to identify whether patients casualty hepatitis are alive or not. It comprises 19 features and 155 cases. Target feature has been represented as 0 for those die with 32 (20.6%) and 1 for those alive with 123 (79.4%) samples. The missing value approximately around 48.30% of the dataset. The dataset features shown in Table I.

Table 1: Margin specifications

NO.	Attribute	Type	Values
1	Class	Categorical	Used as output: - Die - Live
2	Age	Numeric	Numerical values
3	Sex	Categorical	Male, Female
4	Steroid	Categorical	No, Yes
5	Antivirals	Categorical	No, Yes
6	Fatigue	Categorical	No, Yes
7	Malaise	Categorical	No, Yes
8	Anorexia	Categorical	No, Yes
9	Liver Big	Categorical	No, Yes
10	Liver Firm	Categorical	No, Yes
11	Spleen Palpable	Categorical	No, Yes
12	Spiders	Categorical	No, Yes
13	Ascites	Categorical	No, Yes
14	Varices	Categorical	No, Yes
15	Bilirubin	Numeric	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
16	Alk Phosphate	Numeric	33, 80, 120, 160, 200, 250
17	SGOT	Numeric	13, 100, 200, 300, 400, 500
18	Albumin	Numeric	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19	Protime	Numeric	10, 20, 30, 40, 50, 60, 70, 80, 90
20	Histology	Categorical	No, Yes

5. Experimental Model and Results

This section discusses the building design experiment, results, and discussion about the study. We used the SPSS Clementine tool in our analysis. It is the SPSS enterprise-strength data mining workbench. Clementine helps business companies to enhance citizen and customer associations through an indepth consideration of data [12]. Figure 3 and 4 represent the model of the study and sample of accuracy performance using a different type of neural network algorithms.

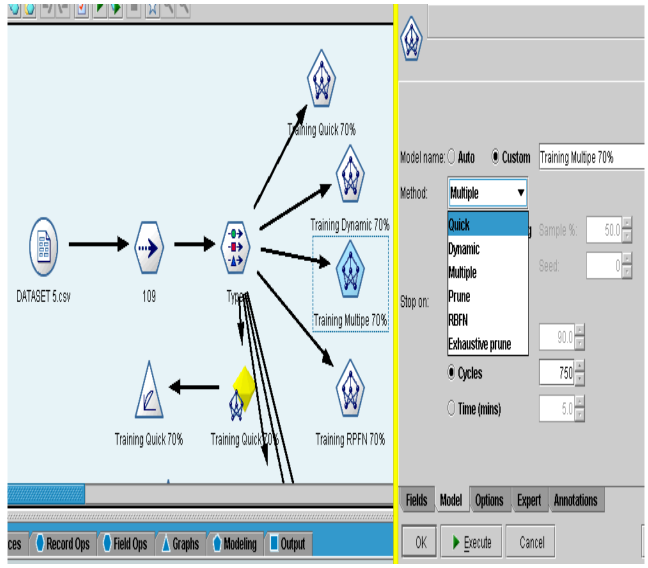

Fig. 3 Model Design

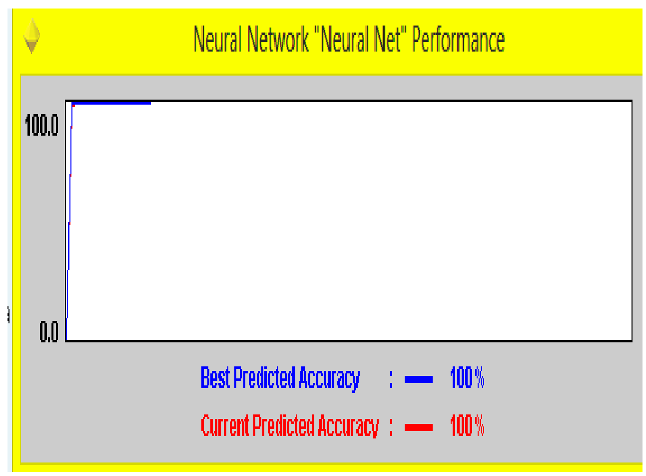

Fig. 4 Sample of accuracy performance

The model was build based on training and testing dataset. The training dataset divided to 70, 60, and 50 where the testing data divided to 30, 40, and 50 respectively. The training results of the neural network algorithms were examined based on learning cycle (1000, 750, 500, and 250) the learning cycle represent the learning time of the neural network in the training phase. Another factor was used such as processing time and size of the data to investigate and determine which type of neural network will obtain the high prediction accuracy and less miss-diagnosis error. The results of the investigation model in training and testing phase is illustrated in table 2 and table 3.

The accuracy of the neural network was calculated as:

$$\text{Accuracy} = \frac{(TN + TP)}{(TN + FP) + (TP + FN)} \times 100 \qquad (1)$$

Table 2: Training Results

Cycles	Algorithm	Size	Estimated Accuracy	Error	Time
1000	Quick	70%	98.165	1.835	1 sec
	Dynamic	70%	99.083	0.917	1 sec
	Multiple	60%	96.774	3.226	1 sec
	RPFN	60%	93.548	6.452	1 sec
750	Quick	70%	98.165	1.835	1 sec
	Dynamic	70%	94.495	5.505	1 sec
	Multiple	60%	91.398	8.602	1 sec
	RPFN	60%	93.548	6.452	1 sec
500	Quick	70%	98.165	1.835	1 sec
	Dynamic	70%	94.495	5.505	1 sec
	Multiple	70%	95.413	4.564	3 sec
	RPFN	50%	94.872	5.128	2 sec
250	Quick	70%	98.165	1.835	1 sec
	Dynamic	70%	94.495	5.505	1 sec
	Multiple	70%	93.578	6.42	1 sec
	RPFN	60%	93.548	6.452	1 sec

The testing results of the neural network algorithms performance demonstrated in table III.

Table 3: Testing Results

Training Cycles	Algorithm	Size	Estimated Accuracy	Error	Time
1000	Quick	30%	75	25	1 sec
	Dynamic	30%	75	25	1 sec
	Multiple	40%	88.043	11.957	1 sec
	RPFN	40%	73.913	26.087	1 sec
750	Quick	30%	75	25	1 sec
	Dynamic	30%	75	25	1 sec
	Multiple	40%	83.696	16.304	1 sec
	RPFN	40%	73.913	26.087	1 sec
500	Quick	30%	75	25	1 sec
	Dynamic	30%	75	25	1 sec
	Multiple	30%	84.259	15.741	1 sec
	RPFN	50%	75.325	24.675	1 sec
250	Quick	30%	75	25	1 sec
	Dynamic	30%	75	25	1 sec
	Multiple	30%	83.333	16.667	1 sec
	RPFN	40%	73.913	26.087	1 sec

The best training and testing results of the neural network algorithms performance that were achieved demonstrated in Table 4.

Table 4: Best Results

Stage	Training Cycles	Algorithm	Size	Estimated Accuracy	Error	Time
Training	1000	Quick	70%	98.165	1.835	1 sec
	1000	Dynamic	70%	99.083	0.917	1 sec
	1000	Multiple	60%	96.774	3.226	1 sec
	500	RPFN	50%	94.872	5.128	2 sec
Testing	1000	Quick	30%	75	25	1 sec
	1000	Dynamic	30%	75	25	1 sec
	1000	Multiple	40%	88.043	11.957	1 sec
	500	RPFN	50%	75.325	24.675	1 sec

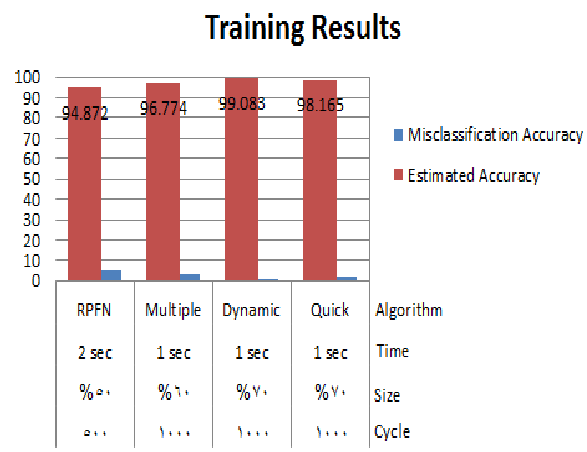

Fig. 4 Result of the best training

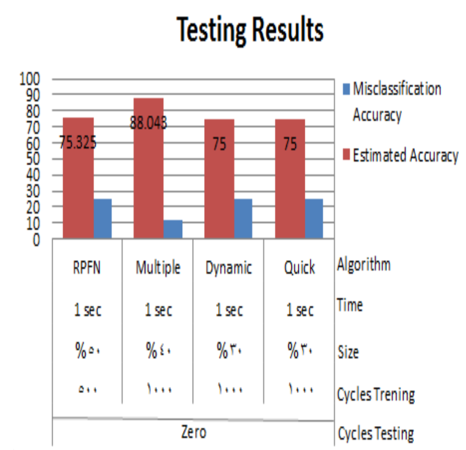

Fig. 5 Result of the best testing

6. Conclusions

This research tried to investigate the hepatitis disease diagnosis to adjust the prediction process based on the neural network type, time process, accuracy, learning cycle and estimated error factor. The quality of hepatitis disease prediction was emphasized using Quick, Multiple Dynamic, and RBFN neural network algorithms. In this study, the experiments conducted based on UCI hepatitis dataset. Our investigation found that the diagnosis results achieved by the Quick, Multiple Dynamic and RBFN neural network algorithms with cycle 250, 500, 750 and 1000 in training and testing process. Also, our deep examinations were highlighted that the Multiple neural network algorithms obtained best result in term of diagnosis accuracy, estimated time, and error factor. In future, we will focus on how to improve the current study using some of the optimization techniques to enhance the diagnosis accuracy of the hepatitis disease.

Acknowledgments

This work was supported by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia. The author, therefore, gratefully acknowledge the technical and financial support from the DSR.

References

[1] J. Cohen, The scientific challenge of hepatitis C, Science 285 (1999) 26.
[2] H. Chen, D. Liu, B. Yang, J. Liu, G. Wang, A new hybrid method based on local Fisher discriminant analysis and support vector machines for hepatitis disease diagnosis, Expert Systems with Applications 38 (9) (2011) 11796–11803.
[3] C.L. Huang, C.J. Wang, A GA-based feature selection and parameters optimization for support vector machines, Expert System with Applications 31 (2) (2006) 231–240.
[4] S.S. Javad, H.Z. Mohammad, M. Kourosh, Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA), Computer Methods and Programs in Biomedicine 108 (2) (2012) 570–579.
[5] E. Dogantekin, A. Dogantekin, D. Avci, Automatic hepatitis diagnosis system based on linear discriminant analysis and adaptive network based on fuzzy inference system, Expert Systems with Applications 36 (8) (2009) 11282–11286.
[6] L. Ozyilmaz, T. Yildirim, Artificial neural networks for diagnosis of hepatitis disease, in: Proceedings of the International Joint Conference on Neural Networks, vol. 1, 2003, pp. 586–589.
[7] M.S. Bascil, F. Temurtas, A study on hepatitis disease diagnosis using multilayer neural network with levenberg marquardt training algorithm, Journal of Medical Systems 35 (1) (2011) 433–436.
[8] M.S. Bascil, H. Oztekin, A study on hepatitis disease diagnosis using the probabilistic neural network, Journal of Medical Systems 36 (3) (2012) 1603–1606.
[9] K. Polat, S. Gunes, Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation, Digital Signal Processing 16 (6) (2006) 889–901.
[10] B. Ster, A. Dobnikar, Neural networks in medical diagnosis: comparison with other methods, in: Proceedings of the International Conference on Engineering Applications of Neural Networks, vol. 1, no. 1, 1996, pp. 427–430.
[11] Azadeh, Ali, et al. "Integration of artificial neural networks and genetic algorithm to predict electrical energy consumption." Applied Mathematics and Computation 186.2 (2007): 1731-1741.Lashley, Karl Spencer. "The problem of serial order in behavior." (1951): 146.
[12] Missing Data Mining - Decision Tree Induction in SAS Enterprise Miner and SPSS Clementine - Comparative Analysis Zulma Ramirez 2901 N Juan St. Edinburg, TX 78541 (956)802-6283.
[13] C.L. Blake, C.J. Merz, UCI Repository of machine learning databases, 1996, Available from HTTP://www.ics.uci.edu./~mlearn/MLReporsitory.html (last accessed 1.12.16).

Barakat *Saeed Alshamrani1* is B.S.c student in King Abdulaziz University, Department of Information System, Faculty of Computing and Information Technology, Jeddah, Kingdom of Saudi Arabia. His research interest in Data Mining and Machine learning techniques

Dr. Ahmed Hamza Osman Ahmed graduated with a Bachelor of Science (Computer Science) from International University of Africa-Sudan. He obtained her MSc Degree in Computer Science from Sudan University of Science and Technology, Sudan. His received PhD degree in Computer Science from University Technology Malaysia (UTM), Malaysia. He is currently work as Assistant Professor in King Abdulaziz University, Department of Information System, Faculty of Computing and Information Technology, Jeddah, Kingdom of Saudi Arabia. His current research interest includes Information Retrieval, Data Mining, Information, Text Processing and Plagiarism Detection, Natural Language Processing and Soft Computing