PCA-ANN for Classification of Hepatitis-C Patients

Tahseen A. Jilani	Huda Yasin	Madiha Mohammad Yasin
Department of Computer Science, University of Karachi, University Road, Karachi, 75270, Pakistan	Department of Computer Science, University of Karachi, University Road, Karachi, 75270, Pakistan	Department of Anatomy, Liaquat College of Medicine & Dentistry, Karachi, 75290, Pakistan

ABSTRACT

In this paper, an automatic diagnosis system based on Neural Network for hepatitis virus is introduced. This automatic diagnosis system deals with the mixture of feature extraction and classification. The system has two stages, which are feature extraction – reduction and classification stages. In the feature extraction – diminution stage, the hepatitis features were obtained from UCI Repository of Machine Learning Databases. Missing values of the instances are adjusted using local mean method. Then, the number of these features was reduced to 6 from 19 due to relative significance of fields. In the classification stage, these reduced features are given as inputs Neural Network classifier. The classification accuracy of this ANN diagnosis system for the diagnosis of hepatitis virus was obtained, this accuracy was around 99.1% for training data and 100% for testing data.

General Terms

Data Mining

Keywords

Artificial neural network (ANN), Hepatitis-C virus, multilayer perceptron (MLP), principle component analysis (PCA)

1. INTRODUCTION

Nowadays, people who have been affected from hepatitis virus are found all over the world. Liver's swelling and redness without indicating any particular reason is referred to as 'Hepatitis' [1]. Hepatitis virus arises due to one of three viruses. Hepatitis A, hepatitis B, and hepatitis C are the prominent viruses. Furthermore, the Epstein Barr Virus can convert into hepatitis which leads to swelling in liver. Plus, some viruses and bacteria are capable to generate hepatitis D and E, varicella (chickenpox), and cytomegalovirus (CMV) [1]. The following paragraphs elucidate Hepatitis A, B, and C as they are assumed to be the most substantial forms of hepatitis:

Hepatitis A is considered as one of the antediluvian virus found in human beings. In earlier times, it was also referred to as contagious hepatitis, epidemic hepatitis, epidemic jaundice and catarrhal jaundice. This virus affects the functions of the liver whilst imitating in hepatocytes [2].

Hepatitis B virus transmits from one person to another. It is hundred times more infectious than human immunodeficiency virus. Body fluids like spit, semen, blood, and vaginal fluid contains this virus. Both acute and chronic form exists in hepatitis B. Illegitimate use of medicine; tainted needles, infected blood giver, and sexual contact are the major causes of this virus. Moreover, if mother is affected from this virus than it would also be transmitted to her new born baby [3].

Hepatitis C virus is the most severe type. It assaults the liver and caused swelling and redness in liver. Large number of inhabitants who had infected from this virus said that they even didn't feel any indication by which they recognized that they have been infected from such severe virus. Most of them come to know about their state during their usual health check up [4]. Transfusion of blood (particularly that took place before 1992), previously used injections, and sexual contact are the chief reasons of this virus. Furthermore, hepatitis C virus will also be transmitted to the new born baby if the mother is tainted from this virus [5]. In this virus, if the patient is having cirrhosis then the possibility of liver cancer raises [6]. In US, more than four million people have been infected from this virus [7]. According to the professionals, the rate of death from hepatitis will be three times more in the next twenty years [8].

For general practitioners, it is really problematic to detect hepatitis virus [9]. For this cause, an automatic diagnosis system based on Artificial Neural Network (ANN) is given in this paper in order to assist physicians for the diagnosis of hepatitis virus.

Hitherto, many studies have been carried out in the diagnosis of hepatitis virus. Kedziora et al. [10] demonstrated that Phylogenetics trees and Hamming distances best reflect the differences between HCV populations present in the organisms of patients who responded positively and negatively to the applied therapy. Hodgson et al. [11] presented an automated system for the quantification of inflammatory cells in hepatitis-C infected liver biopsies. The required features are extracted from color-corrected biopsy images at positions of interest identified by adaptive threshold and clump decomposition. The experimental results show that this system can rank 15 test images, with varying degrees of inflammation, in strong agreement with five expert pathologists. Kemal Polat and Salih Güne [9] presented a novel method for the diagnosis of hepatitis virus based on a hybrid method that uses feature selection (FS) and artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism. The obtained classification accuracy of their system was 92.59%. Moreover, sensitivity and specificity values for the experimental results were 100% and 85%.

Peng Guan, De-Sheng Huang and Bao-Sen Zhou [12] studied the application of artificial neural network (ANN) in forecasting the incidence of hepatitis A by using ARIMA model and compared the results with an ANN model. They concluded that ANN is superior to conventional methods which have an auto-regression phenomenon. Avendao [13] formulated a model to describe the dynamics of hepatitis C virus (HCV). They discussed the efficacy of the therapy methods for hepatitis C in terms of the threshold parameter. Success of the therapy could possibly be forecasted from the early viral dynamics in the patients. For this purpose, they considered four populations: uninfected liver cells, infected liver cells, HCV and T cells. Moneim and Mosa [14] constructed a mathematical model to study the spread of HCV-subtype 4a amongst the Egyptian

population. The relation between HCV-subtype 4a and the other subtypes has also been studied in the paper.

This paper is organized as follows. In section 2, description regarding data is given. In section 3, we have given a review of principle component analysis. Section 4 briefly describes artificial neural network. Proposed PCA based artificial neural network is presented in section 5. In section 6, results obtained using ANN are given. Finally, discussion and conclusion are presented in section 7.

2. DATA DESCRIPTION

The data available at UCI machine learning data repository contains 19 fields with one output field [15]. The output shows whether patients with hepatitis are alive or dead. The purpose of the dataset is to forecast the presence or absence of hepatitis virus given the results of various medical tests carried out on a patient. This database holds 19 attributes. Hepatitis dataset contains 155 samples belonging to two different target classes. There are 19 features, 13 binary and 6 attributes with 6–8 discrete values. Out of total 155 cases, the class variable contains 32 cases that died due to hepatitis.

Table 1: Description of attributes from UCCI machine learning dataset

S. No	Variable	Values
1	Class	Die, Alive
2	AGE	10, 20, 30, 40, 50, 60, 70, 80
3	SEX	Male, female
4	STEROID	No, Yes
5	ANTIVIRALS	No, Yes
6	FATIQUE	No, Yes
7	MALAISE	No, Yes
8	ANOREXIA	No, Yes
9	LIVER BIG	No, Yes
10	LIVER FIRM	No, Yes
11	PLEEN PALPABLE	No, Yes
12	SPIDERS	No, Yes
13	ASCITES	No, Yes
14	VARICES	No, Yes
15	BILIRUBIN	0.39, 0.80, 1.20, 2.0, 3.0, 4.0
16	ALK PHOSPHATE	33, 80, 120, 160, 200, 250
17	SGOT	13, 100, 200, 300, 400, 500
18	ALBUMIN	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19	PROTIME	10, 20, … , 90
20	HISTOLOGY	No, Yes

3. INTRODUCTION TO DIMENSION REDUCTION TECHNIQUES

Discovering unrevealed or unidentified information and constructive patterns in a database is often referred to as data mining. The terms knowledge discovery, information retrieval, deductive learning and exploratory data analysis can be used in place of data mining [16].

Several different algorithms are incorporated in data mining in order to carry out different tasks. In general, the data mining scopes are segregated into predictive and descriptive areas with application specific changes pertaining to the requisites of the problems. Making prediction regarding data values by means of former identified outcomes from some other data is set by predictive model where as classification of patterns in data is done by descriptive model [16].

3.1 Principal Component Analysis

Dimension of a huge data set can be trimmed down by using principal component analysis which is considered as one of the most prevalent and useful statistical method. This method transforms the original data in to new dimensions. The new variables are formed by taking linear combinations of the original variables of the form:

$$H_1 = b_1'K = b_{11}K_1 + b_{12}K_2 + \cdots + b_{1m}K_m$$
$$H_2 = b_2'K = b_{21}K_1 + b_{22}K_2 + \cdots + b_{2m}K_m$$
$$\cdots$$
$$H_p = b_p'K = b_{p1}K_1 + b_{p2}K_2 + \cdots + b_{pm}K_m$$

In matrix style, we can write $H = B.K$, where $b_{11}, b_{12}, .., b_{pp}$ are known as the loading parameters. The new axes are attuned such that they are orthogonal to one another with utmost expand of information.

$$Var(H_i) = b_i' \sum b_i \quad , \qquad i = 1,2,..,p$$

$$Cov(H_i, H_j) = b_i' \sum b_j \quad , \; i = 1,2,..,p$$

K_1 is the first principal component holding the prime variance. As the direct computation of matrix B is not achievable. So, in feature transformation, the first step is to ascertain the covariance matrix U which can be expressed as [17]:

$$U_{m \times n} = \frac{1}{m-1} \left[\sum_{i=1}^{m} (K_i - \overline{K})' . (K_i - \overline{K}) \right],$$

where $\overline{K} = (\frac{1}{m}) \sum_{i=1}^{m} X_i$

The next step is to determine the eigen values for the covariance matrix 'U'. Eventually, a linear transformation is defined by n eigen vectors match up to n eigen values from a m-dimensional space to n-dimensional space (n<m). Principal axes are also referred to as eigen vectors $E_1, E_2, .., E_m$ correspond to eigen values $\lambda_1 + \lambda_2 + \cdots + \lambda_n$. Generally, the first few principal components hold most of the information. Analysis of variances' proportion represents the total number of principal components that should be retained from the dataset [17].

Table 2: Description of attributes selected after data preparation and data reduction step

Selected Variables	Variable Description
Age	In the dataset, smallest age is '7' and largest age is '78'.
Bilirubin	A pigment fundamentally derived from the go down of hemoglobin from red blood cells damaged in the spleen.
Alk Phosphate	It is an enzyme exists in the blood.

| SGOT	Serum glutamic oxaloacetic transaminase is an enzyme generally exists in serum. Also, it is present in heart.
Albumin	Serum albumin is the chief protein of blood plasma as well as of other serous solutions.
PROTIME	Prothrombin is a predecessor of thrombin which is produced in the liver.
Class	Die/ Live

4. ARTIFICIAL NEURAL NETWORK

Biological neural networks (BNN) are natural machines that can perform a number of different operations that range from vision, hearing to sensing. The brain is a highly intricate, nonlinear and parallel information processing system. It has the capability to organize its structural constituents, known as neurons, so as to perform certain computations like pattern recognition, pattern matching, classification and forecasting. After the development of artificial neuron see [18], that represent under certain limits, (the similar characteristics that exists in biological neuron), this field has grown so rapidly that new applications and theoretical interpretations have given artificial neural networks a major place in solving complex and gigantic real problems [19][20][21][22][23].

A neuron is the fundamental information-processing unit of a NN that consists of, a set of synaptic links, an adder and an activation function. This is presented in Fig. 1. Mathematically, the function of kth neuron in a neural network can be defined as

$$u_k = \sum_{j=0}^{m} w_{kj} x_i \text{ with } x_0 = 1, b_k = w_{k0}$$
$$and \quad y_k = f(u_k)$$

where $x_1, x_2, .., x_m$ are the input signals and $\mathcal{W}_{k1}, \mathcal{W}_{k2}, .., \mathcal{W}_{km}$ are the synaptic weighs of neuron k; u_k is the linear combination output due to the input signals; b_k is the bias parameter; $f(.)$ is the activation function; and y_k is the output signal of the neuron. The bias b_k is an external parameter and shows an affine transformation to the output u_k.
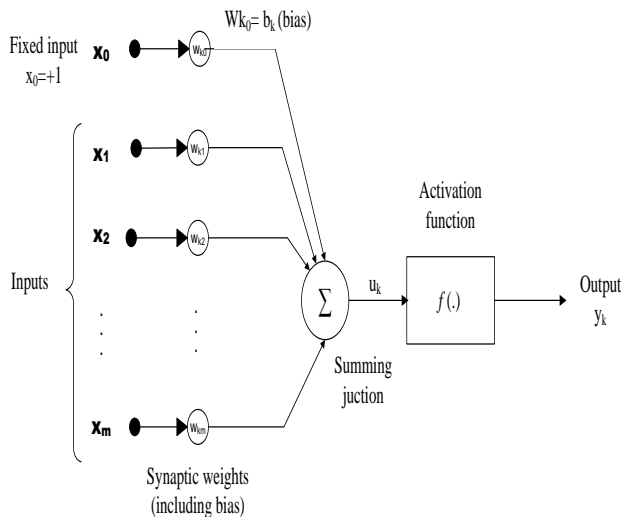
Fig 1: Model of an Artificial Neuron

4.1 Error-Correction Learning

Consider the simple case of a neuron k constituting the only computational node in the output layer of a feed forward neural network. Neuron k is driven by a signal vector $x(n)$ produced by one or more layers of hidden neurons, which are themselves driven by an input vector (stimulus) applied to the input layer (i.e. source nodes) of the neural network. The argument n denotes discrete time, or more precisely, the time step of an iterative process involved in adjusting the synaptic weights of neuron k. The output signal of neuron k is denoted by $y_k(n)$. This output signal, representing the only output of the neural network, is compared to a desired response or target output, denoted by $d_k(n)$. Consequently, an error signal, denoted by $e_k(n)$ is produced, given by

$$e_k(n) = d_k(n) - y_k(n)$$

The error signal $e_k(n)$ actuates a control mechanism; the purpose is to apply a sequence of corrective adjustments to the synaptic weights of neuron k. The corrective adjustments are designed to make the output signal $y_k(n)$ come closer to the desired response $d_k(n)$ in a step-by-step manner. This objective is achieved by minimizing a cost function

$$E(n) = \frac{1}{2} e_k^2(n)$$

The step-by-step adjustments to the synaptic weights of neuron k are continued until the system reaches a steady state. At that point the learning process is ended.

In particular, minimization of the cost function leads to a learning rule commonly referred to as the delta rule or Widrow-Hoff rule, named in honor of its originators [24]. Let $w_{kj}(n)$ denote the value of synaptic weight of output neuron k excited by element $x_j(n)$ of the signal vector $x(n)$ at time step n. According to the delta rule, the adjustment

$$\Delta w_{kj}(n) = \eta e_k(n) x_j(n)$$

where η is the learning rate parameter. Then the update value of synaptic weight is $w_{kj}(n)$ determined by

$$\mathcal{W}_{kj}(n + 1) = \mathcal{W}_{kj}(n) + \Delta \mathcal{W}_{kj}(n)$$

4.2 Perceptron Neural Network

In a perceptron NN, hidden neurons play a critical role in the operations of a multilayer perceptron with back-propagation learning because they act as feature detectors. As the learning process steps forward, the hidden neurons begin to gradually discover the salient features that characterize the training data. They do so by performing a nonlinear transformation on the input data onto a new space called the hidden space, or feature space. In the new space, the classes of interest in a pattern-classification task, for example, may be more easily separated from each other than in the original space [25][26].

4.3 Back Propagation Algorithm

The error signal at the output neuron j at iteration n (i.e., presentation of the nth training example) is defined as

$$e_j(n) = d_j(n) - y_j(n)$$

We define the instantaneous value of the error signal for neuron j as $\frac{1}{2} e_j^2(n)$. Correspondingly, the instantaneous value $E(n)$ of the total energy is obtained by

$$E(n) = \frac{1}{2} \sum_{j \in Q} e_j^2(n)$$

where the set Q contains all the neurons in the output layer of the network. If there are N training input patterns then the average squared error energy is obtained

$$E_{av} = \frac{1}{N} \sum_{n=1}^{N} E(n)$$

where $\frac{1}{N}$ is the normalizing factor. The objective of the study is to minimize E_{av} using LMS algorithm. Network free parameters are adjusted on a pattern-by-pattern basis until one epoch that is complete presentation of the entire set, has been dealt with. The adjustments to the weights are made in accordance with the respective error computed for each pattern presented to the network. Based on LMS method, the back-propagation formula for the local gradient $\delta_k(n)$ is described:

$$\delta_k(n) = f_j' \left(u_j(n) \right) . \sum_k \delta_k(n) . w_{kj}(n) \qquad (1)$$

where neuron j is hidden neuron

The factor $f_j'(u_j(n))$ involved in the computations of induced local field in (1), depends solely on the activation function associated with the hidden neuron j. Thus the general expression of back-propagation algorithm is:

Weight Correction $\Delta W_{ji}(n)$
= (learning rate parameter η). (local gradient)
∗ (input signal of neuron j $y_i(n)$)

5. SIMULATIONS

Table 3: Training and Testing Classification using ANN (6,3,1) Model

			Predicted	
Sample	Observed	1	2	Percentage Correct
		1	22	1	95.70%
Training		2	0	84	100.00%
	Overall Percentage	20.60%	79.40%	**99.10%**
		1	10	0	100.00%
Testing		2	0	38	100.00%
	Overall Percent	20.80%	79.20%	**100.00%**
Dependent Variable: Class				

Table 4: ANN Model Summary

Cross Entropy Error	4.735
Percent Incorrect Predictions	0.90%
Stopping Rule Used	1 consecutive step(s) with no decrease in error [a]
Training Time	00:00.2
Cross Entropy Error	0.264

a. Error computations are based on the testing sample

In classification stage, we have used 69% records for training model and 31% for testing diagnosis. Using ANN, the training and testing classification is represented in Table 3 having 'Class' as a dependent variable. ANN based model summary is

represented in Table 4 giving the value of cross entropy error and percentage of erroneous predictions along with the training time. Table 5 represents classification accuracy in percentage of nineteen different methods which were used for the diagnosis of hepatitis virus and our proposed method [27].

6. RESULTS

Before proceeding for model fitting, we have applied some data reduction techniques in order to trim down the dimensions because the data have the problem of dimensionality's curse. We have applied principal component analysis on the nineteen independent variables, after which we realized that the first six principal components cover more than 98% of the total variability of the continuous data space.

Table 5: Classification accuracies obtained by using hepatitis diagnostic methods

Used method	Article author's	Classification accuracy (%)
RBF	Özyıldırım, Yıldırım, et al.	83.75
15NN, stand. Euclidean	Grudzinski	89
FSM without rotations	Adamczak	88.5
MLP with BP	Stern and Dobnikar	82.1
QDA, quadratic discriminant analysis	Stern and Dobnikar	85.8
ASI	Stern and Dobnikar	82
MLP + BP (Tooldiag)	Adamczak	77.4
LDA	Stern and Dobnikar	86.4
MLP	Özyıldırım, Yıldırım, et al.	74.37
RBF (Tooldiag)	Adamczak	79
1NN	Stern and Dobnikar	85.3
Naïve Bayes and semi-NB	Stern and Dobnikar	86.3
Fisher discriminant analysis	Stern and Dobnikar	84.5
LVQ	Stern and Dobnikar	83.2
GRNN	Özyıldırım, Yıldırım, et al.	80
ASR	Stern and Dobnikar	85
IncNet	Norbert Jankowski	86
CART (decision tree)	Stern and Dobnikar	82.7
LFC	Stern and Dobnikar	81.9
Proposed Method	Used in this study	**89.6**

Therefore, we have acquired six independent variables after applying data reduction approaches, these six autonomous variables are age, bilirubin, Alk phosphate, serum glutamic oxaloacetic transaminase (SGOT), Albumin and prothrombin time (PROTIME). The normalized importance percentage of these independent variables is given in Table 6, which indicates that PROTIME, SGOT, and age hold the utmost percentage. For the analysis of hepatitis virus, the classification precision of this ANN diagnosis system was achieved, this accuracy was about 99.1% for training data and 100% for testing data.

Table 6: Independent variable importance

	Importance	Normalized Importance
Age	0.205	81.10%
Bilirubin	0.053	20.90%
Alk Phosphate	0.143	56.70%
SGOT	0.224	88.60%
Albumin	0.123	48.50%
PROTIME	0.253	100.00%

7. CONCLUSION AND FUTURE STUDIES

In this paper, we have investigated factors which have higher pervasiveness of the risk of hepatitis-C virus. We have observed that out of 19 attributes, 6 appeared to be the most imperative factors. Our automated diagnosis system deals with the grouping of feature withdrawal and classification. The accuracy obtained from our applied process was 100% for testing data and 99.1% for training data. In future, we would like to consider other factors such as hepatic enzymes (ALT etc.), blood picture, urine analysis (bilirubinuria, mild proteinuria, urobilinogen), serology for viral markers etc. Above and beyond, we will apply fuzzy learning models on a large number of patients.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] Watkins, A. (2001). AIRS: A resource limited artificial immune classifier; Master thesis, Mississippi State University. Available at: http://www.kidshealth.org/parent/infections/bacterial_viral/hepatitis.html#.

[2] Documentation Prepared by World Health Organization. Available at: http://www.who.int/csr/disease/hepatitis/HepatitisA_whocdscsredc2000_7.pdf.

[3] Bupa-Health and Life Care; available at: http://www.bupa.co.uk/individuals/.

[4] Information regarding hepatitis C from the staff of Mayo Clinic; available at: http://www.mayoclinic.com/health/hepatitis-c/DS00097.

[5] Causes of hepatitis C according to the staff of Mayo Clinic; available at: http://www.mayoclinic.com/health/hepatitis-c/DS00097/DSECTION=causes.

[6] Blake, C. L., & Merz, C. J. (1996), UCI repository of machine learning databases. Available from: http://www.ics.uci.edu./~mlearn/MLReporsitory.html.

[7] Cara Bruce, Lisa Montanarelli, M.D. Teresa Wright, and Teresa L. Wright, Hepatitis C: An Essential Guide for the Newly Diagnosed; 2nd edition, 2007.

[8] Mark Jenkins, Hepatitis C: practical, medical, and spiritual guidelines for daily living with HCV, 2000.

[9] Polat K., Gune S., "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 (2006), pp. 889–901.

[10] Kedziora P., Figlerowicz M., Formanowicz P., Alejska M., Jackowiak P., Malinowska N., Fratczak A., Blazewicz J., and Figlerowicz M., "Computational Methods in Diagnostics of Chronic Hepatitis C", Bulletin of the Polish Academy of Sciences, Technical Sciences, 53 (3), 2005, pp.273-281

[11] Hodgson S., Harrison R. F., Cross S. S., An automated pattern recognition system for the quantification of inflammatory cells in hepatitis-C-infected liver biopsies, Image and Vision Computing 24, 2006, pp. 1025–1038.

[12] Peng Guan, De-Sheng Huang, Bao-Sen Zhou, "Forecasting model for the incidence of hepatitis A based on artificial neural network", China World Journal of Gastroenterol; 10(24), 2004, pp. 3579-3582.

[13] Avendao R., Esteva L., Flores J. A., Fuentes Allen J. L., Gomez G., Lopez-Estrada J. E., A Mathematical Model for the Dynamics of Hepatitis C, Computational and Mathematical Methods in Medicine, Publisher by Taylor & Francis (Health-Sciences), 4(2), 2002, pp.109 — 118.

[14] Moneim I. A. and Mosa G. A., "Modeling the Hepatitis C with Different Types of Virus Genome", Computational and Mathematical Methods in Medicine, 7(1), 2006, pp. 3-13.

[15] Information concerning hepatitis C; available at: http://www.medicalnewstoday.com/articles/145869.php

[16] Dunham M. H. and Sridhar S., Data Mining: Introductory and Advanced topics, Pearson Education, 2006

[17] Jang, Sun and Mizutani, Neuro-fuzzy logic and Soft Computing; A computational approach to learning and machine intelligence. New York: Practice-Hall, 2003, Chap. 2-4.

[18] Rosenblatt, Principles of Neuro-dynamics-An investigation on Perceptrons, 1962.

[19] Aqil Burney S.M., Jilani A. Tahseen and Cemal Ardil, "A comparative study of first and second order training algorithms for artificial neural networks", Int. Journal of Computational Intelligence, vol. 1, no.3, 2004, pp. 218-224.

[20] Chen and Huang, "Dynamic Analysis of Stochastic Recurrent Neural Networks", 2003

[21] Y. Abu-Mostafa, A. Atiya, M. Magdon-Ismail, and H. White, "Guest Editorial: Special Issue on Neural Networks in Financial Engineering", IEEE Transactions on Neural Networks, Vol. 12, No. 4, July 2001, pp. 653-656.

[22] Kantardzic M., Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, 2003, chapter 5.

[23] Battiti, Training Neural Nets with the Reactive Tabu Search, 1992

[24] Widrow and Hoff, Adaptive Switching Circuits, In *Western Electronic Show and Convention, Volume 4*, pages 96-104, Institute of Radio Engineers (now IEEE), 1960.

[25] Aqil Burney S.M., Jilani A. Tahseen and Saleemi, "Approximate Knowledge Extraction using MRA for TYPE-I Fuzzy Neural Networks", 2006.

[26] Aqil Burney S.M., Jilani A. Tahseen and Cemal Ardil, "Approximate Bounded Knowledge Extraction using Type-I Fuzzy Logic", 2005

[27] Dogantekin E., Dogantekin A.and Avci D., Automatic hepatitis diagnosis system based on Linear Discriminant Analysis and Adaptive Network based on Fuzzy Inference System, Expert Systems with Applications, 36, 2009, pp. 11282-11286.