IDENTIFICATION OF INFLUENTIAL BIOMARKERS FOR HUMAN LEUKEMIA – AN ARTIFICIAL NEURAL NETWORK APPROACH

[1]SOUGATA SHEET, [2]ANUPAM GHOSH, [3]SUDHINDU BIKASH MANDAL

[1,2]A.K.Choudhury School of Information Technology, University of Calcutta, West Bengal, India
[3]Netaji Subhash Engineering College, Kolkata, West Bengal, India
E-mail: [1]sougata.sheet@gmail.com, [2]anupamghosh@rediffmail.com,[3]sudhindumandal@gmail.com

Abstract— Leukemia is a blood cancer. Children and adults under the age of 20 are most affected in leukemia. If it is detected and treated at the early stage cancer can be cured. This paper represents the classification of leukemia blood cells inside the B lymphocytes and plasma cells by using multi layer feed-forward Artificial Neural Network. Here, the sample consists of B lymphocytes, plasma cells, chronic lymphocytic leukemia, and multiple myeloma. The Back Propagation algorithm has been employed to train the network. There are ten input node, five hidden node and one output node are created. However, the network trained by the Back Propagation algorithm and validated using *t*- test and *F*- score.

Keywords— Back propagation; *t*-test; True positive; False positive; False negative; *F*-score.

I. INTRODUCTION

Cancer is a group of 100 diseases. Cancer has two important things. One is that some cells in the body become normal. Another is that the body keeps conducting huge numbers of abnormal cells [1]. In leukemia, abnormal blood cells are creating in the bone marrow. Generally, leukemia includes the creation of abnormal white blood cells. However, the abnormal cells in leukemia do not work in the same way as normal white blood cells [2]. The leukemia cells continue to develop and division eventually crowding out the normal blood cells. The result is that is become hard for the body to fight infection, control bleeding and transport oxygen. To understand leukemia fact, it is most subsidiaries to know about normal blood cells and what happens to them when leukemia develops. When leukemia grows up, the body produces many abnormal blood cells. Sometimes consider a children disease, leukemia generally arrive in adults older than 65 years. The relationship between environmental and leukemia is still not well understand; however research have ascertained that exposure to high levels of ionizing relation has been attached to some types of leukemia in both children and adults. There are above 100 several types of cancer and one of them is leukemia. Leukemia is a cancer of blood and marrow. A large amount of abnormal white blood cells can be produce in leukemia cancer. These types of cells are incomplete and they do not work properly [1]. Leukemia can be a deadly disease without treatment. In this reason more death can be occur among children and adults under the age of 20 [3]. There are several types of leukemia, based upon how speedily the disease develops and the type of abnormal cells generate. The four general types of leukemia are acute lymphocytic leukemia, chronic lymphocytic leukemia, acute myeloid leukemia, and chronic myeloid leukemia. In United States over 50,000 cases of leukemia occur yearly [13]. Each year there are almost 54,000 new cases of leukemia in the United States and about 24,000 deaths due to leukemia. Nearly 27,000 adults and more than 2,000 children in the United States are affected in leukemia each year. Leukemia makes up about 3% of all new cancer cases [3].

Leukemia can be recovery if it is detected and treated in short order. In general, doctor can be detecting leukemia by performing the entire blood computation method [5]. If there is unusualness in this computation, knowledge of morphological bone marrow dirty is made to assure the subsistent of leukemia cells [4]. Specific morphologic character will be accomplished in order to categorize the acute leukemia as either Acute Lymphoblast Leukemia (ALL) or Acute Myelogenous Leukemia (AML) [14 15]. Lymphoblast and myeloblast both are the abnormal white blood cells. The availability of lymphoblast in the blood sample will be respecting of ALL. On the other hand, the availability of myeloblast in the blood sample will be respecting of AML. Recently, new research is completed by hematologists by visual identification under microscope. This alignment is very important for best treatment. The error rate of manual recognition process is between 35% and 45% depending on the hematologists experience and different types of leukemia [4].

Artificial Neural Network (ANN) is externally used for the complicated data analysis. ANN was overcome the narrowness of the influence of computer to finish the certain work. ANN has been used for resolve problem such as speech recognition [6] and diagnosis of different types of cancer such as breast cancer [7 8], cervical cancer [9] etc. Numerous types of effective researches on blood cells using neural network have been done. Ongun *et al*. [10] increased a fully automated classification of blood and bone marrow using different way including neural network and support vector machine (SVM). The best performance of SVM with 91.05% accuracy

as parallelism to Multilayer perceptron (MLP) network using Conjugate Gradient Descent, Linear Vector Quantization (LVQ), and k-Nearest Neighbour (KNN) classifier which generate 89.74%, 83.33% and 80.76% accuracy respectively. Toure and Basu [12] create MLP network using Back Propagation (BP) algorithm to prognosis the several types of leukemia which is either ALL or AML and the result is 58% accuracy on test data sets. Hsieh *et al*. [11] mention a leukemia cancer model which is based on Information Gain SVM technique and classification of two types of leukemia which are ALL and AML. The result shows that the SVM model has create the best result for classification of leukemia cancer with accuracy 98.10%.

In this work, we implement an Artificial Neural Network. We used on leukemia gene expression data sets. In these data sets some samples are normal genes and some samples are diseased (leukemia) genes. We can identify which is normal or diseased. From ANN we can identify important genes.

II. METHODOLOGY

The power of an Artificial Neural Network to become differentiates them from the highest automatic controller. Like humans, neural networks learn by example, and thus necessity to be trained [16]. An Artificial Neural network is generally configured to earmarked applications and have to power to method huge number of data. Complicated trends and pattern can be found out using Artificial Neural Network. An Artificial Neural Networks there is learning methods allow the device to identify a specific pattern and carry out a particular task. These learning method build the network to ordain the weight parameters in order to equalize the given input and output data with activation function. In this frequent method, generate the initial output values by using initial input values. The weight in the network are adjust to match the data based on this input and output values. Next pair of input and output values is correlative and adjust the weight values. This method is carry on until the network gets a good fit for data [17]. Rosenblatt, introduce the single layer feed forward network which are limited to learning almost separate patterns. In non linear the input and output are affix individual data with sufficient training to model any well identify function to unrestricted precision. This Multilayer Feed forward network (MFN) is also known as Multi Layer Perceptron (MLP). The MLP generally trained by using Back propagation algorithm [18 20]. The back propagation algorithm overlook the process input to output by reduce error between the output and calculated output from the input and network. The network selects the weight as a random number and calculates the essential correction. The algorithm can be computed following steps.

Step1: Initialize all weight and bias;
Step2: While terminating condition is not satisfied;
Step3: for each training tuple
Step4: Propagate the input forward
Step5: for each input layer unit "q"
Step6: $O_q = I_q$; Output of an input unit is its actual input value
Step7: for each hidden or output layer unit "q";
Step8: $I_q = \Sigma\, w_{pq}O_p + \alpha_q$; Compute the net input of unit q with respect to the previous layer "p"
Step9: $O_q = \frac{1}{1+e^{-Iq}}$; Compute the output of each unit "q"
Step10: Back propagate the error
Step11: for each unit q in the output layer
Step12: $E_q = O_q\,(1 - O_q)(X_q - O_q)$; Compute the error
Step13: for each unit q in the hidden layer, from the last to first hidden layer
Step14: $E_q = O_q\,(1- O_q)\,\Sigma\, E_r\beta_{qr}$; Compute the error with respect to the next higher layer "r"
Step15: $\Delta\beta_{pq} = (\eta)E_qO_q$; Increment the weight of the layer
Step16: $\beta_{pq} = \beta_{pq} + \Delta\beta_{pq}$; Update the weight of each layer
Step17: $\Delta\alpha_q = (\eta)E_q$; Increment the bias of each layer
Step18: $\alpha_q = \alpha_q + \Delta\alpha_q$; Update the bias of each layer

The activation function of a node identifies the output of that node given an input or set of inputs. A linear access can be generate 1 or 0 output, but for non linear method the activation function can be produce the output in the specific limit. The activation function can be accepted huge forms build on the data sets.

Back propagation learns by frequently producing a set of training samples, comparing the networks prediction for each sample with the actual known class level. For each sample, with weight are changes as to reduce the mean squared error between the actual class and the prediction network. Each step is described below.

2.1. Initialize the weight

Small random number is to initialize the weight of the network. Similarly the each bias of the network is initialized to small random number.

2.2. Propagate the inputs forward

The total input and output of each unit in the output layer and hidden layer are calculated in this step. At first the sample are fed to the input layer of the network. Note that for unit "q" in the input layer, its output is equal to its input, that is $O_q = I_q$ for input "q". Given a unit "q" in a hidden layer or output layer, the net input "I_q" to unit "q" is

$$I_q = \Sigma\, w_{pq}O_p + \alpha_q \qquad (1)$$

Where "w_{pq}" is the weight of the connection from unit "p" in the previous layer to unit "q"; "O_p" is the output of unit "p" from the previous layer; and "α_q" is the bias of the unit. Given the net input "I_p" to the unit "q", then "O_q" is the output of unit "q", is computed as

$$O_q = \frac{1}{1+e^{-I_q}} \qquad (2)$$

2.3. Back propagate the error

In the network the error is created backwards by updating the weight and bias. For a unit "q" in the output layer, the error "E_q" is compute by

$$E_q = O_q (1 - O_q)(X_q - O_q) \qquad (3)$$

Where "O_q" is the actual output unit of "q" and "X_q" is the True output. The error of a hidden layer unit "q" is

$$E_q = O_q (1- O_q) \Sigma E_r \beta_{qr} \qquad (4)$$

Where "β_{qr}" is the weight of the connection from unit "q" to unit "r" in the next higher layer, and E_r is the error of unit "r". The bias and weight are updated to the propagated the error by the following equation, where "$\Delta\beta_{pq}$" is the change in weight "β_{pq}".

$$\Delta\beta_{pq} = (\eta)E_q O_q \qquad (5)$$

$$\beta_{pq} = \beta_{pq} + \Delta\beta_{pq} \qquad (6)$$

The variable "η" is the learning rate, which is a constant typically having a value between 0.0 and 1.0. Biases are updated by the following equation, where "$\Delta\alpha_q$" is the change in bias "α_q"

$$\Delta\alpha_q = (\eta)E_q \qquad (7)$$

$$\alpha_q = \alpha_q + \Delta\alpha_q \qquad (8)$$

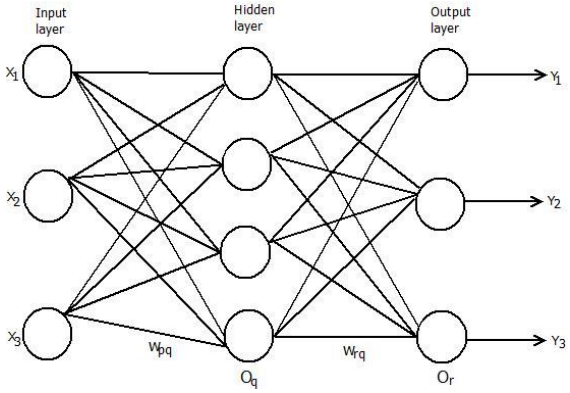

Fig.1. Structure of n-q-m artificial neural network

At least two physical elements are generating a neural network. They are processing components and joining between them. The processing elements which is called neurons and the joining between the neurons which is called the link. Each link has a weight and every neuron accepts the information from the neighboring neurons attached to it and processing the information and generated the output. The neuron which is received the input from the outside is known as the input neurons. The neuron which is received the input from other neurons and whose output is an input for another neuron is called hidden neurons. Neuron whose output is used externally is known as output neurons. The raw information which is fed into the network is representing input layer. This section of the network never changes its values. Hidden layer are received the data from input layer. It uses input value and changing them using some weight values. This new values are dispatch to the output layer and it is also changing some weights from connection between hidden layer and output layer. Then output layer process the information which is received from hidden layer and generate an output. This output is produce by activation function [19].

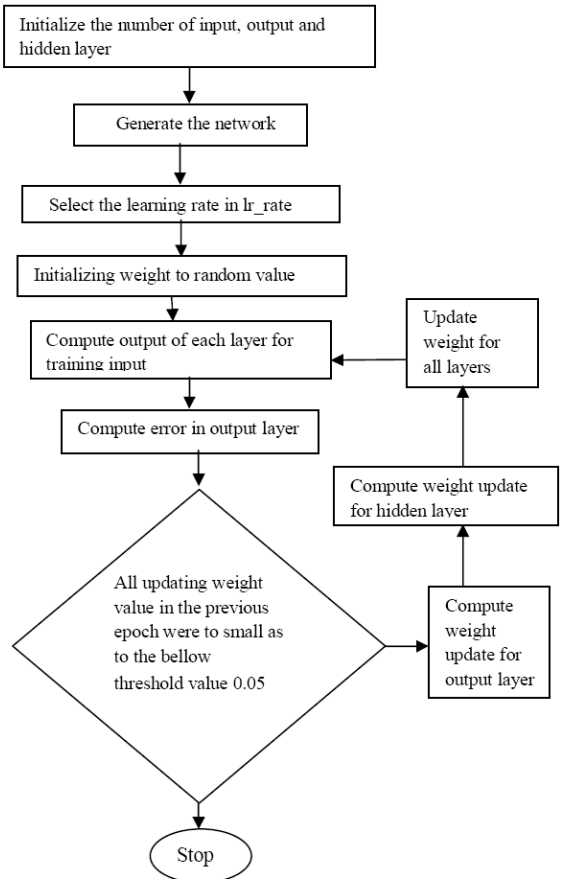

Fig.2. Flow chart of back propagation algorithm

III. DATA SET DESCRIPTION

A genes expression level show the inferential number of copies of that genes RNA generate in a cell and it is correlated with the quantity of the corresponding proteins made. It has been shown that specific patterns of gene expression happen during several biological states such as embryo genesis, cell development and during normal physiological response in tissues and cells. Thus the expression of gene provides the measure of activity of gene under certain biochemical condition. It is known that particular diseases, such as cancer, are reflected in the change of the expression values of particular genes. Normal cell can involve into malignant cancer cell through a series of mutation in genes that control the cell cycle, apoptosis and genome integrity. In this work we select Waldenstrom's macroglobulinemia data sets which consist of B lymphocytes and plasma

cells. The total data set consists of 22283 numbers of genes with 56 samples. Among them, there are 13 normal samples which contents 8 normal for B lymphocytes and 5 normal for plasma cells. On the other hand 43 diseased samples which contents 20 Waldenstrom's macroglobulinemia, 11 chronic lymphocytic leukemia, 12 multiple myeloma samples.

IV. RESULTS AND DISCUSSION

We have selected 22283 numbers of Human lymphocytes and plasma cell expression data sets. In this data sets we are identify which is normal genes and which is diseased. We have found two classifier groups. One is normal class and other is disease class. After some iteration, we found the normalized value of each gene. Here we have considered threshold value which is 0.05. After normalization if the genes value is grater then 0.05, then which type of genes are normal genes and we have to indicate as 1. After normalization if the genes value is less than 0.05, then which type of genes are disease genes and we have to indicate as 0. After normalization, we have found mean of the resulting expression values of genes. In this work we select the learning rate (η) as between 0.0 to 0.9. We calculate TP, FP, FN and *F*-score of each learning values. Show in below. In Table 1 we show that when the learning rate (η) is 0.9 the number of true positive genes is maximum. In this reason we can select the learning rate (η) is 0.9. **Table 1** show the learning rate (η) and corresponding TP, FP, FN and *F*- score value. The algorithm generates the excellent performance with accuracy of 95.70%.

4.1. Validation using t- test

t-test is statistical significance indicates whether
Or not the difference between two group's averages most likely reflects an original difference in the population from which the groups were sampled. We have performed *t*-test on these leukemia data sets. The *t*-value show that most significant genes (99.9%) which p-value <0.001. For leukemia data sets we are applying *t*-test and we gave the *t*-value.

Table1: Learning rate and corresponding TP, FP, FN and *F*- score value

Learning rate (η)	TP	FP	FN	*F*-Score
0.1	1038	818	523	0.61
0.2	1456	912	756	0.63
0.3	1944	1008	928	0.67
0.4	2226	1096	977	0.68
0.5	2618	1233	1137	0.69
0.6	3115	1456	1258	0.69
0.7	3547	1653	1388	0.70
0.8	4212	1856	1523	0.72
0.9	5038	2022	1753	0.73

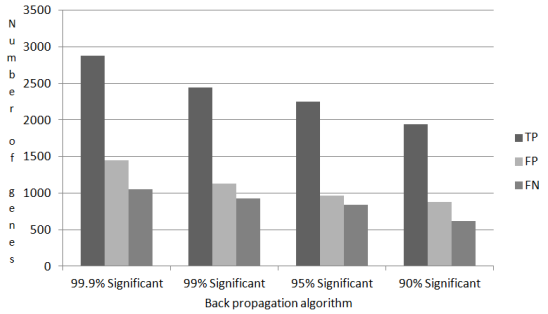

Fig.3. Each Significant level of genes

We have identify some important genes like IARS(5.98),MMP25(4.58),TYMS(3.96), HPS6(5.59), MLX(5.32),CALCA(4.12),HIC2(5.02),ANP32B(4.5), TFPI(5.72),CRYAB(3.98),NCF1C(3.39), HNRNPH1(4.92), etc. The number in the bracket shows *t*-value of the corresponding gene. The *t*-value of this genes exceeds the value for p<0.001. This means that this gene is highly significant (99.9% of significant). Similarly genes like ERCC5(3.12), PRDM2(3.17),PRIM2(2.61),TPT1(3.29),RPS26 (2.83),EFCAB11(3.22),PRPSAP2(3.57), PRKACA(2.84), etc exceeds the value for p<0.01. It indicates that this gene is significant at the level of 99%. Similarly genes like MED17(2.34), MAPK1(2.42),PIK3CB(2.05),NMD3(2.34), ARG2(2.19),EXOC3(2.16),WHSC1(2.18), RFC4(2.26),GLB1L(2.41), HNF1A(2.05) etc exceeds the value for p< 0.05. It indicate that this genes significant at the level of 95%. Similarly genes like FLG(1.97), TXNL1(1.82), RIN3(1.95),CYBB(2.04), ZNF814(1.72), KLF4(1.28) etc exceeds the value for p<0.1. It indicate that this type of genes significant at the level of 90%.

Now we calculate the True positive (TP), False positive (FP) and False negative (FN) of each significant level. Accuracy evaluates the performance of the classifier that can classify both types of normal and disease genes. The higher value accuracy indicates batter performance of classifier. The classifier that can correctly classify leukemia cancer genes will have better result in specificity. Figure 3 show the TP, FP, FN number of genes which significant level is 99.9%, 99%, 95% and 90%. Now we select the top 100 genes. Now we calculate the TP, FP, and FN number of genes whose significant level is 99.9%, 99%, 95% and 90%. Figure 4 show the result.

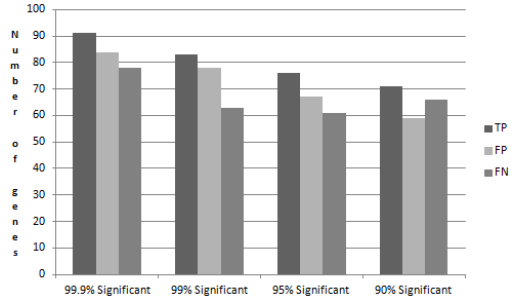

Fig.4. Top 100 significant genes

Now we calculate *F*- score of each significant level of genes. At first we calculate precision and recall value of each significant level of gene then we calculate the *F*- score. Figure 5 show the result. Based on the result in Figure 5 we can show that *F*- score are continuously increase at 99.9%, 99%, 95%, 90%, and NCBI data base respectively.

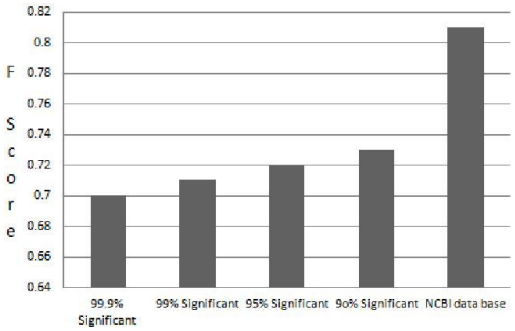

Fig.5. *F*- Score value of each significant level

4.2. Validation using NCBI data base

In this section we assessed the validity and reliability of human leukemia gene expression database which is provide the National Center for Biotechnology Information (NCBI). Now we calculate the TP, FP, FN value of human leukemia genes expression database. NCBI provide 9664 human leukemia genes. Now we are comparing this data sets and leukemia gene expression data sets. As a result we found 6226 true positive genes. The result is show in figure 6.

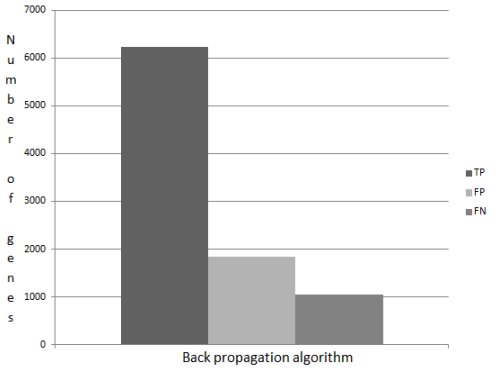

Fig.6. NCBI data base

Now we show the classification performance of epochs and hidden nodes of Back Propagation neural network. Figure7 and Figure 8 show the classification performance of epochs and hidden node for Back Propagation neural network.

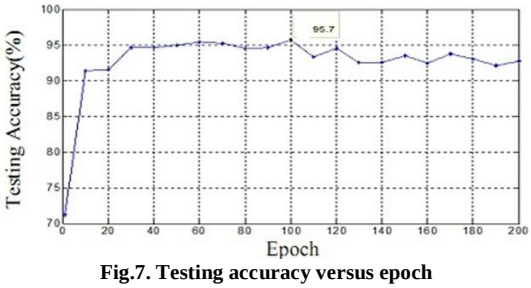

Fig.7. Testing accuracy versus epoch

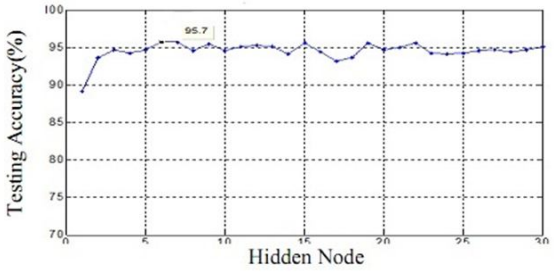

Fig.8 Testing accuracy versus hidden node

Based on the result in figure7 and figure8, the excellent classification performance of epoch and hidden node with the testing accuracy is 95.70%.

CONCLUSIONS

The neural network using BP algorithm has been used to classify leukemia blood cells into two type namely normal samples and diseased sample. In normal sample are consist B lymphocytes and plasma cells. Diseased sample consist Waldenstrom's macroglobulinemia, chronic lymphocytic leukemia, multiple myeloma. The result show that each significant level of genes and corresponding *F*-score. However, the network using BP algorithm has proved excellent classification performance of 95.70%. The result significantly explains the excellence of the proposed characteristic and classification neural network for classification of leukemia blood sample.

APPENDIX

True Positive: A true positive test result is one that detects the condition when the condition is present. True positive rate = *TP/(TP+FN)*.

False Positive: A false positive is an error in some rating method in which a condition tested for is badly found to have been detected. A false positive test result is one that detects the condition when the condition is absent. False positive value = *FP/(FP+TN)*.

False Negative: A result that appears negative when it should not. A false negative test result is one that does not detect the condition when the condition is present. False negative value = *FN/(TP+FN)*.

True Negative: A true negative test result is one that does not detect the condition when the condition is absent. True negative value = *TN/(TN+FP)*.

F-score: F- score are a statistical method for determining accuracy accounting for both precision and recall. The formula for traditional *F* score is, *F = 2* (precision * recall/ precision + recall)*. Where

Precision = *TP/TP+FP*

Recall = *TP/TP+FN*

REFERENCES

[1] D.H. Henry, "Latest News in Blood Cancer Research", Cancer Care, New York, 2010.

[2] G.C.C. Lim, "Overview of Cancer in Malaysia", Japanese Journal of Clinical Oncology, vol. 32, no. 1, pp. 37-42, 2002.

[3] G.C.C. Lim, S. Rampal, H. Yahaya, "Cancer Incidence in Peninsular Malaysia", The Third Report of the National Cancer Registry, Malaysia, 2008.

[4] C. Reta, L. Altamirano, J.A. Gonzalez, R. Diaz and J.S. Guichard, "Segmentation of Bone Marrow Cell Images for Morphological Classification of Acute Leukemia", in Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference, USA, pp. 86-91, 2010.

[5] G.P.M Priyankara, O.W Seneviratne, R.K.O.H Silva, W.V.D Soysa, C.R. De Silva, "An Extensible Computer Vision Application for Blood Cell Recognition and Analysis", pp. 1-13, 2006.

[6] Md. Salam, D. Mohamad, and S. Salleh, "Malay Isolated Speech Recognition Using Neural Network: A Work in Finding Number of Hidden Nodes and Learning Parameters", The International Arab Journal of Information Technology, vol. 8, no. 4, pp. 364-371, 2011.

[7] M..A Mohamed, Abd.EI.F. Hegazy, A.A Badr, "Evolutionary Fuzzy ARTMAP Approach for Breast Cancer Diagnosis", IJCSNS International Journal of Computer Science and Network Security, vol. 11 no. 4, pp. 77-84, 2011

[8] C.L Chi, W.N Street and W.H Wolberg, "Application of Artificial Neural Network-Based Survival Analysis on Two Breast Cancer Datasets", AMIA Annual Symposium Proceedings Archive, pp. 130-134, 2007.

[9] L.J Mango, "Computer-Assisted Cervical Cancer Screening using Neural Networks", Cancer Letters, vol. 77, no. 2-3, pp. 155-162, 1994.

[10] G. Ongun, U. Halici, K. Leblebiciogul, V. Atalay, M. Beksac, S. Beksac, "Feature Extraction and Classification of Blood Cells for an Automated Differential Blood Count System", International Joint Conference IJCNN'01, vol. 4, pp. 2461-2466, 2001.

[11] S.H Hsieh, Z Wang, P.H. Cheng, I-S Lee, S.L Hsieh, F. Lai, "Leukemia Cancer Classification Based on Support Vector Machine", in Proceedings of the 8th IEEE International Conference on Industrial Informatics, Japan, pp. 819-824, 2010.

[12] A Toure, M Basu, "Application of Neural Network to Gene Expression Data for Cancer Classification", International Joint Conference on Neural Network, vol.1, pp. 583-587, 2001.

[13] C Demir, B Yener, "Automated Cancer Diagnosis Based on Histopathological Images: A Systematic Survey", Technical Report, Rensselaer Polytechnic Institute, 2005.

[14] L Seewald, J.W. Taub, K. W Maloney, E. R. B McCabe, "Acute leukemia in Children with Down Syndrome", International Journal of Molecular Genetics and Metabolism, vol. 107, pp. 25-30, 2012, ISSN: 1096-7192.

[15] S Mohapatra, D Patra, "Automated Cell Nucleus Segmentation and Acute Leukemia Detection in Blood Microscopic Images", Proceedings of International Conference on Systems in Medicine and Biology, India, pp. 49-54, 2001.

[16] Y Zhang, D Guo, Z Li," Common nature of learning between BP and Hop field-type neural networks for generalized matrix in version with simplified models", IEEE Transaction on Neural Network and learning system, vol. 24, NO.4, pp. 579-592. 2013.

[17] J. D. J Rubio, D. M. Vazquez, J Pacheco," Back propagation to train an evolving radial basis function neural network", Evolving System, vol.1, no. 3, pp. 173-180, 2010.

[18] F Yu, X Xu, "A short-term load for e-casting model of natural gas based on optimized genetic algorithm and improved BP neural network", Appl.Energy, vol. 134, pp. 102-113. 2014.

[19] J. S Kim, S Jung, "Implementation of the RBF neural chip with the back-propagation algorithm for on-line learning", International journal of Applied Soft Computing, vol. 29, pp. 233-244, 2015, ISSN: 1568-4946.

[20] L Wang, Y Zeng, T Chen, "Back propagation neural network with adaptive differential evolution algorithm for time series forecasting", International journal of Expert System with Application, vol. 42, pp. 855-863, 2015, ISSN: 0957-4174.

★★★