Open Journal of
Oncology & Hematology

Research Article

Selection of Gene Mediating Human Leukemia, using Deep Learning Approach - ə

Sougata Sheet[1]*, Anupam Ghosh[2] and Sudhindu B. Mandal[1]

[1]AK. Choudhury School of Information Technology, University of Calcutta, Kolkata-700098, India
[2]Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata-700152, India

***Address for Correspondence:** Sougata Sheeta, Choudhury School of Information Technology, University of Calcutta, Kolkata-700098, India; E-mail: sougata.sheet@gmail.com

Submitted: 22 December 2016; **Approved:** 16 March 2017; **Published:** 22 March 2017

Citation this article: Sheet S, Ghosh A, Mandal SB. Selection of Gene Mediating Human Leukemia, using Deep Learning Approach. Open J Oncol Hematol. 2017;2(1): 001-009.

ABSTRACT

In is paper, we present a procedure for gene selection, which genes are may be normal or certain cancer. At first, we select an entire set of genes group. Then using Deep Neural Network model (DNN) to determined important group. Finally, individual genes are selected from important group which are valuable in terms of their importance in mediating cancer. The usefulness of the procedure is demonstrated using three microarray human leukaemia gene expression data sets. We have formed a comprehensive comparative analysis with three existing process using biochemical pathway, *p* - value, t - test, F - test, expression profile plots for indentifying biologically and statistically pertinent gene sets. It has been obtained that the present procedure capable to select genes that are most biologically significant in given leukaemia than those gained by another.

Keywords: Deep neural network; t - test; *p* - value; F – score

INTRODUCTION

Leukaemia is a malignancy of blood cells. In leukaemia, abnormal blood cells are generating in the bone marrow [1]. Generally leukaemia entangles the creation of abnormal white blood cells [2]. However, the abnormal cells in leukaemia do not function in the same way as normal white blood cells [3]. There are several types of leukaemia, based upon how speedily the disease is growing up and the type of abnormal cells are generate. The four general types of leukaemia are Acute Lymphocytic Leukaemia (ALL) [4], Chronic Lymphocytic Leukaemia (CLL), Acute Myeloid Leukaemia (AML) and Chronic Myeloid Leukaemia (CML). Acute Lymphoblastic Leukaemia is a cancer of lymphoblast which is white blood cells that fight infection [5]. White blood cells are most usual type of blood cell to become cancer. But red blood cells which carry the oxygen from the lungs to the rest of the body and platelets may also become cancer [6]. Leukaemia occurs mostly in adults older than 55 years, and it is the most general cancer in children younger than 15 years [3]. Leukaemia is either acute or chronic. Acute leukaemia is a fast-growing cancer that generally gets worse speedily. Chronic leukaemia is a slower-growing cancer which gets worse slowly over time. The treatment and prediction for leukaemia depend on the type of blood cell invaded and whether the leukaemia is acute or chronic. In 2015, 54,270 people are prospective to be diagnosed with leukaemia. There are an almost 327,520 people subsistent with, or in relief from, leukaemia in the US. The overall five- year correlative anointing rate for leukaemia has more than fourfold since 1960. From 1960 to 1965, the five-year correlative anointing rate among whites (only data available) with leukaemia was 14%. From 1975 to 1980, the five-year correlative anointing rate for the total population with leukaemia was 34.20%, and from 2004 to 2010, the overall correlative anointing rate was 60.30% [7]. From 2005 - 2010, the five-year relative correlative anointing overall were CML is 59.90%, CLL is 83.50%, AML is 25.40% overall and 66.30% for children and adolescents younger than 15 years and ALL is 70% overall [8], 91.80% for children and adolescents younger than 15 years, and 93% for children younger than 5 years. In 2015, 24,450 people are expected to die from leukaemia with 14,040 males and 10,050 females. In 2007-2011, leukaemia was the fifth most common cause of cancer deaths in men and the sixth most common in women in the US [8].

In the field of microarray data analysis the gene selection is one of the most challenging remittance. Gene expression data usually contains a huge number of variables genes compared to the number of samples [9]. The conventional data mining procedure cannot be directly used to the data due to this identity problem. The analysis of gene expression data used dimension reduction procedure for this reason [10]. The gene selection which deducts the genes extremely related to the pattern of every type of disease in order to escape such problems. The statistical obtainments with parametric and non-parametric tests. For example t - test [11] and Wilcox on rank sum test have been thoroughly used for searching differentially revealed genes since they are instinctive to understand and implement. But they have a restriction to propagate, if more than two classes and require time swallowing coordination to solve the problem of multiple testing. The Kruskal-Wallis test can be used for three or more groups. But it may be generate prejudiced result because of the reliance on the number of samples, when it is used to microarray data whose sample size are generally unbalanced. Microarray and gene chips have formed it feasible to experimental assessing the disclosure of many separate genes at a time [12]. Many diseases, they are reason by the problems such as chromosomal disequilibrium and gene mutations, which give away abnormal gene expression patterns. These patterns give the information about the underlying genetic method and states of several types of disease [13]. If these patterns can be analyzed properly, they can be efficient for recognize the disease sample and detecting the extent to which a forbearing is affliction from the disease and which can be help in the management of disease [14].

The microarray gene expression data have been collected to underlying biological process of a number of diseases. It is very important to narrow down from thousands of genes to a few disease genes and gene ranking; gene selection is most important step in microarray data analysis [15]. For classification data analysis, several types of way have been proposed for gene ranking [16]. These are classified into three several types: filter [17], wrapper and embedded process. Each of these categories has its personal advantages and disadvantages. For example, filter process are computationally useful and simple but minor performance than the other process. On the other hand, wrapper and embedded procedure are comparatively much complicated and computationally costly but it usually gives excellent classification performance as they mainly apply classifier characteristics in gene ranking. Filter procedure include T- score, which is t-statistic standardized interrelation between input and output class labels. On the other hand, wrapper and embedded procedure include Support Vector Machine (SVM) [18] and its variants [19,20], Random Forest (RFE) [21], elastic net [22], guided regularized random forest [23], balanced iterative random forest [24] etc. Main distinction of filter process and wrapper or embedded procedure is how they behave samples when ranking genes. For example, in filter procedure, all the samples are usually used for gene ranking but the quality and relevance data samples are ignore. On the other hand wrapper or embedded procedure, classifier such as boosting algorithm, logistic regression, Support Vector Machine (SVM) etc., is used to gene ranking [25].

In the field of the Machine Learning, like Computer vision, Speech recognition the Deep Learning has earned state of the art result. The developments of the Deep Belief Network are started in the field of the Deep Learning [26]. The unsupervised pre-training and supervised pre-training are used in Deep Belief Network. Different types of unsupervised pre-training algorithm are developed based on same concept such as Stacked Denoising auto encoder [27] and Contractive auto encoder [28]. We get the excellent performance in deep learning without unsupervised pre-training, it has been possible with the development of the Rectified Linear Network [29]. A Deep Neural Network and Multilayer Perception (MLP) are same structure. The hidden layer numbers are defined as the depth of the neural network [30]. At least two hidden layer are present in-depth model [30] and usually needs weights then MLP [31]. If the numbers of hidden layer are increase, then the model complexity are exponentially increase and needs more training samples. The main deep learning methods are Restricted Boltzmann machine model [31] which is energy based training models. Deep learning process is unsupervised, i.e., the model used the input information. In system identification, the deep learning methods can't be applied directly because as classification problem, the input and output are the non binary value [31].

The present paper is an effort in this regard provides a new procedure neural network models for identifying genes mediating normal genes and disease genes. Now we can identify the genes by using this neural network model. We said this model is Deep Neural Network model (DNN). The procedure is to the purpose in a data-rich condition i.e., if the number of sample completely huge comparison to the dimension of each sample. In this problem, the number of microarray samples is fully less compared to the number of genes. To solve this problem, we have suggest a process of generating much data from the given microarray gene expression data sets. The proposed procedure, along with its best performance over other various processes has been demonstrated using three microarray gene expression leukaemia data sets. The existing process with which the results have been compared, Support Vector Machine (SVM), Significance Analysis of Microarray (SAM), and Signal to Noise Ratio (SNR). We can perform comprehensive analysis using biochemical pathway, *p* - value, t - test, F - score. It has been found that the method has been able to the genes are more significant.

SOME EXISTING METHODS

In this section, we have present identification of cancer mediating genes by using neural network model. For comparative analysis, we have created a survey on existing gene selection methods. Among them, we have select three types of gene selection method namely SVM, SAM and SNR.

In machine learning procedure SVM are used to isolate two classes by maximizing the limit between them. For cancer classification, support vector machine are used to identify important genes. For quadratic programming and linear programming methods are used standard SVM and Lasso (L1) SVM. In gene selection methods Recursive Feature Elimination (RFE) SVM is other algorithm which is used weight magnitude for ranking standardization. In SVM-RFE, all genes are hold some scoring function and one or more gene are remove according to their score value. When the highest accuracy value is achieved, the procedure will be stopped.

In SAM, the genes are identified statistically significant changes of expression values using set of gene specified t - test. On the basis of the changing gene expression values, every gene has generated a score value. If score value is grater then threshold value that means which gene is most significant.

According to their discriminative power rank correlated genes are applied SNR method. The method starts single gene evaluation and frequently searches for another genes based on statistical parameter. The important genes are selected based on their high SNR score. The procedure is more efficient for detecting and ranking smaller number of significant genes, if the number of genes can be decreased significantly.

METHODOLOGY

Let us assume a set G = (g_1; g_2; …….; g_n) of n genes are known which is hold the first "m" expression values in normal samples and the subsequent "n" expression values in diseased samples. Now interrelation coefficient of gene-based normal samples is calculated. Therefore, interrelation coefficient R_{pq} within p^{th} and q^{th} genes is given by [17,18]

$$R_{pq} = \frac{\sum_{k=1}^{m}(g_{pk} - y_p)*(g_{qk} - y_q)}{\sqrt{(\sum_{k=1}^{m}(g_{pk} - y_p)^2)} * \sqrt{(\sum_{k=1}^{m}(g_{qk} - y_q)^2)}} \quad (1)$$

Here y_p and y_q are the mean of expression values of p^{th} and q^{th} genes, respectively in normal samples. Similarly, for diseased samples the iteration coefficient R'_{pq} between p^{th} and q^{th} genes is given by

$$R'_{pq} = \frac{\sum_{k'=1}^{m}(g'_{pk'} - y'_p)*(g_{qk'} - y'_q)}{\sqrt{(\sum_{k'=1}^{m}(g'_{pk'} - y'_p)^2)} * \sqrt{(\sum_{k'=1}^{m}(g'_{qk'} - y'_q)^2)}} \quad (2)$$

Each pair of genes is computed by using (1). The genes are located in the similar group if $R_{pq} >= 0.50$. Now we have used interrelation coefficient to narrow downhearted the invention space by searching genes of a comparable behaviour in terms of related expression patterns. The set of responsible genes mediating certain cancers are recognized in this procedure. The choice of 0.50 as a threshold value has been done through extensive experimentation for which the distances among the cluster center have become maximize. The main set of genes is obtained in this pathway.

In Deep learning methodology, the learning rate is considered as a hyper limit and optimized based on their smallest validation error. Magnitude value of the back propagate error derivative is few from lower layers as compared to the upper layers. For usefulness training learning rate are different for different layers. In the possible set of learning rate there is an exponential growth with increase of the depth neural network. For many DNNs with several learning rates in every layer are searching for optimal hyper parameter and the optimization of the learning rate is computationally expensive. For Deep MLP, an attempt has been creating to expand a new hyper parameter free adaptive learning algorithm. The methodology endeavours to avoid extensive searching for getting optimal learning rate hyper parameter in MLP. Learning rate is composed of a function of a parameter β, which is known as learning parameter. The weights of the DNN are updated in the same way learning parameter are updated. Figure 1 shown the structure of the Deep Neural Network.

In this methodology, two functions, namely sigmoid and

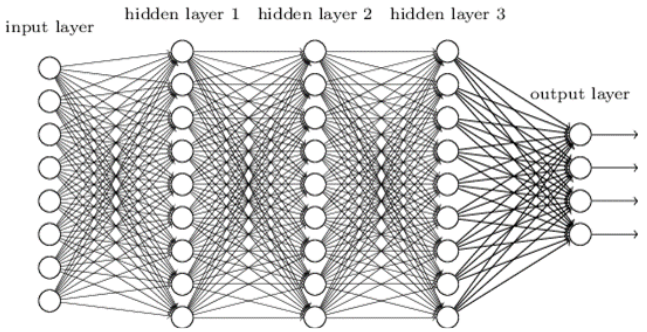

Figure 1: Structure of the Deep Neural Network.

exponential are considered as given bellow.

$$F_1(\beta^{(L)}) = 1/(1+\exp(-\beta^{(L)})) \quad (3)$$

$$F_2(\beta^{(L)}) = \exp(-\beta^{(L)^2}) \quad (4)$$

Where $\beta^{(L)}$ is the learning parameter of the layer L.

In each neuron to every layer, the weight magnitude is changes, expect the input layer. Learning rate updating is used to the Laplacian score concept. The relevance of every neuron of every layer is used to Laplacian score.

For a mini batch, the output from every layer is accepted during the forward propagation. The Laplacian score of the neuron is calculated by using the output of neuron of mini batch for any particular layer. The neuron is depends on Laplacian score of output and output is depends on the incoming weights of the neuron. If the Laplacian score is high, that means incoming weight of the neuron are already trained to generate the output. Now the weight connecting neurons in layer (L-1) to i^{th} neuron in layer L. The effective learning rate is given by

$$\mu_i^{(L)} = (1-l_i^{(L)})F(\beta^{(L)}) \quad (5)$$

Where is the Laplacian score of the output from i^{th} neuron in layer L and $\beta^{(L)}$ learning parameter of the weight values which is connecting in layer of layer L. F is the function which is given by (3) and (4).

Laplacian Score Calculation

Let in a layer contain four neuron and feed-forward step using mini batch of size three. In a particular level of neuron, the activation function is denoted by the A_{ij}. Where i is the index neuron and j is the mini-batch. Activation values can be represented in matrix forms

$$\text{Activation values} = \begin{bmatrix} A_{11}A_{12}A_{13} \\ A_{21}A_{22}A_{23} \\ A_{31}A_{32}A_{33} \\ A_{41}A_{42}A_{43} \end{bmatrix}$$

Calculation of the updated weight

Input for every hidden layer and output layer are calculated

$$\eta_i^L = \sum_{j=1}^{S(L-1)} \omega_{ji}^L (O_j^{(L-1)} D_j^{(L-1)}) + B_j^{(L)} \quad (6)$$

Where η_i^L is the i^{th} neuron in layer L.

S^L is the size of layer L.

ω_{ji}^L is the weight of the j^{th} neuron in $(L-1)$ layer to i^{th} neuron in L layer.

O_j^L is the output of the j^{th} neuron in L layer.

D_j^L is the Dropout mask j^{th} neuron in layer L.

B_j^L is the Binomial

Now we are applied Rectified Linear activation function. Now the hidden activation function which is given by

$$O_i^L = \max(0, \eta_i^L) \forall \text{ i =1 to S} \quad (7)$$

Sigmoid activation function is used only for the output layer. Cross entropy error is used and is given as follows

$$\exists = -\sum_{i=1}^{S(\kappa)} (\Psi_i \log(O_i^\kappa) + (1-\Psi_i)\log(1-O_i^\kappa)) \quad (8)$$

Here is the class level and total number of layer is denoted.

Changing weight for every layer are starting from output to first hidden layer are given by

$$\omega_{nj}^L = \omega_{nj}^L - \mu_j^{(L)} \frac{\partial \exists}{\partial \omega_{nj}^L} \quad (9)$$

For all layer, except output layer is given by

$$\ell_j^{(L)} = \begin{cases} 0 \\ D_j^{(\kappa-L-1)} \sum_{i=1}^{S(\kappa-L)} \omega_{ji}^{(\kappa-L)} \ell_i^{(L-1)} \end{cases} \quad (10)$$

ℓ^κ Is the output layer, which is given by

$$\ell_i^\kappa = (O_i^\kappa - \Psi_i) \quad \forall \text{ j =1 to } S^k \quad (11)$$

The incoming weight for j^{th} neuron in layer L, the learning rate is given by

$$\mu_j^{(L)} = (1-l_j^{(L)})F(\beta^{(L)}) \quad \forall \text{ j =1to } S^L \quad (12)$$

Where $\beta^{(L)}$ is the learning parameter for incoming weight of layer L, and $l_j^{(L)}$ is the Laplacian score of L layer.

Weight update for the bias are given bellow

$$B_i^{(L)} = B_i^{(L)} - \delta_i^{(L)} \frac{\partial \exists}{\partial B_i^{(L)}} \quad (13)$$

Where

$$\frac{\partial \exists}{\partial B_i^{(L)}} = -\ell_i^{(\kappa-L+1)} \quad \forall \text{ i =1 to } S^L$$

Learning rate of the incoming bias of i^{th} neuron in layer L are given by

$$\delta_i^{(L)} = (1-l_i^{(L)})F(\alpha^{(L)}) \quad \forall i = 1 \text{ to } S^L \quad (14)$$

The equation used through the training part are same as the forward pass equation through the testing part but the dropout mask is not used through training, and the output of every hidden layer is multiplied by the terms of $(1-droupout)$. For every layer, the incoming input is given by

$$\eta_i^{(L)} = \sum_{j=1}^{S(L-1)} \omega_{ji}^{(L)} ((1-dropout)O_j^{(L-1)}) + B_i^{(L)} \quad \forall \text{ j=1 to } S^L \quad (15)$$

Rectified linear activation function is used for the hidden layer and sigmoid activation function is used for output layer.

Algorithm

Input: Testing and training date sets. Weight and bias is the initial random values.

In training data set for every mini-batch. //Training
For every layer L from 1 to κ //Forward pass
For every neuron i from 1 to S^L

Calculate input $\eta_i^{(L)}$ and output $O_i^{(L)}$

Calculate Laplacian Score $l_i^{(L)}$

End
End
From every layer L from κ to 1 // Backward pass
For i from 1 to S^L // gradient calculation

For j from 1 to $S^{(L-1)}$

Calculate the error $\frac{\partial \exists}{\partial \omega_{ji}^{(L)}}$
End

Calculate error $\frac{\partial \exists}{\partial B_i^{(L)}}$
End

The learning parameter $\beta^{(L)}$ and $\alpha^{(L)}$ are update
For every neuron i from 1 to S^L //calculate learning rate

Calculate learning rate $\mu_i^{(L)}$ and $\delta_i^{(L)}$
End
For i from 1 to S^L //weight and bias update
For j from 1 to $S^{(L-1)}$

Update the weight $\omega_{ji}^{(L)}$ using learning rate $\mu_i^{(L)}$
End

Update the bias $B_i^{(L)}$ using learning rate $\delta_j^{(L)}$
End
End
End
For every sample in training data sets //Testing
For every layer L, from 1 to κ //forward pass
For every neuron i from 1to S^L

Input $\eta_i^{(L)}$ and output $O_i^{(L)}$ are calculated
End
End
Calculate misclassification error

$$\exists(s) = \begin{cases} 0 \\ 1 \end{cases}$$
End
Misclassification error percentage are calculated

$\exists_{percentage} = 100 * \sum_s \exists(s)$ //In testing data set of total number of samples

Return $\exists_{percentage}$

DESCRIPTION OF THE DATA SETS

In this work we can select three types of data sets. The name of the data set is Waldenstrom's Macroglobulinemia (B lymphocytes and plasma cells). It has been applied for the solution of B Lymphocytes (BL) and Plasma Cells (PC) from patients with Waldenstrom's Macroglobulinemia (WM). The data set ID is GDS-2643. The total data set consists of 22,283 numbers of genes with 56 samples. Among them, there are 13 normal samples which consist of 8 normal for B Lymphocytes and 5 normal plasma cells and 43 diseased samples which consist of 20 Waldenstrom's Macroglobulinemia, 11 chronic lymphocytic leukaemias, 12 multiple myeloma samples. The data base web link is http://ncbi.nlm.gov/projects/geo/ .

Now we can select B-cell chronic lymphocytic leukaemia cell, which consist of 22283 numbers of genes with 16 samples. In this sample analysis of B-cell Chronic Lymphocytic Leukaemia (B-CLL) cells that express or do not express Zeta-Associated Protein (ZAP-70) and CD38. The prognosis of patients with ZAP-70-CD38- B-CLL cells is good, those with ZAP-70+CD38+ B-CLL cells is poor. The data set ID is GDS- 2501.

Now we can select acute myeloid leukaemia gene expression data set. The data set content 26 Acute Myeloid Leukaemia (AML) patients with normal hematopoietic cells at a variety of different stages of maturation from 38 healthy donors. The total data set consist of 22283 numbers of genes with 64 samples. Among them, there are 38 normal samples which contents 10 normal for bone marrow, 10 normal samples for peripheral blood, 8 normal samples for bone marrow CD34plus and 10 normal samples for Primed Peripheral Blood Hematopoietic Stem Cells (PBSC) CD34 plus. On the other hand there are 26 leukaemia samples which contents 7 bone marrow and 19 peripheral bloods. The data set ID is GDS-3057.

ANALYSIS OF THE RESULT

In this section, the usefulness of the methodology has been demonstrated three types of human leukaemia gene expression data sets. A comparative analysis with three existing methods like, SVM, SAM and SNR We have applied this procedure on the gene expression data sets for selecting important genes. We have found two classifier groups. One is normal class and another is disease class. After some iteration, we have found normalized value of every gene. Here we have considered a threshold value which is 0.05. After normalization if the gene value is grater then 0.05, then which types of gene is normal gene. After normalization if the gene value is less than 0.05, then which types of gene is disease gene. We consider several genes that are most significant of our experiment. The gene expression values are significantly changes from normal samples to diseased samples. Applying this process on the first data set (GDS-2643), we have found that genes like ERCC5, TPT1 and CRYAB among the most important genes which are over the expressed the diseased samples. On the other hand RMT minimize the expression value and fully significant in diseased samples. The gene are recognize as an under expressed gene. In order to limited size of manuscript, we have showed only the profile plots of genes of GDS-2643 data set (Figure 2). In the case of GDS-2501 data set, the genes like IGLL3P, CD44, ASCL1, and RCN3 have changed their expression values for normal samples to diseased samples. Similarly, the data set GDS-3057, like TAF4, KCNJ9, TTC12, CTIF have changed their expression values for normal samples to diseased samples. The usefulness of the methodology has been shown three types of leukaemia gene expression data. We have applied the methodology on the aforesaid gene expression data sets for selecting some important gene intercede diseases. In this methodology, at first the genes are placed into groups based on the interrelation values. For first leukaemia gene expression data set (GDS-2643), we have got 8 groups, which contents 2741, 2856, 2691, 2476, 2786, 2813, 2784, and 2673 genes (Table 1). Same test, we have been carried out for the other two data sets.

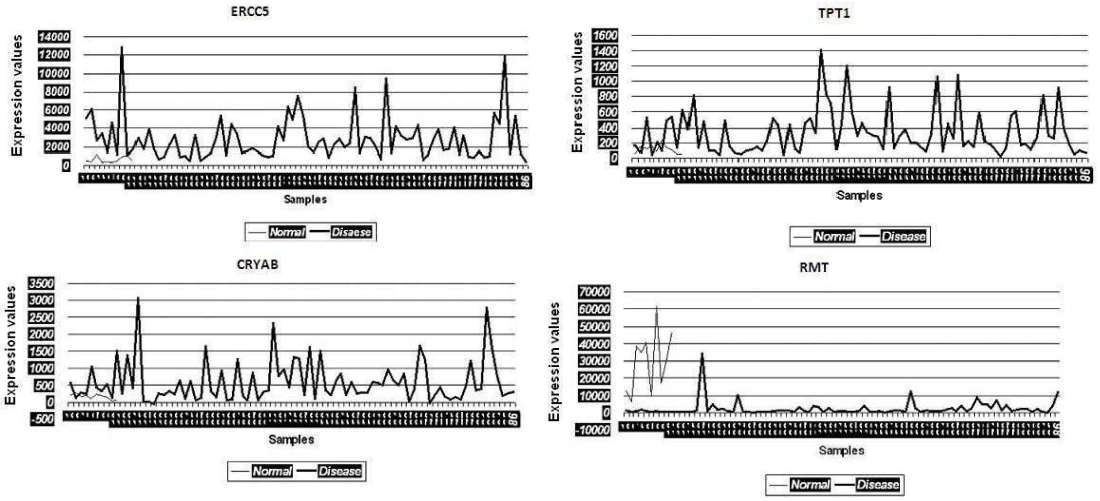

Figure 2: Expression profiles of some over expressed and some under expressed genes in normal samples which is represented by light line and diseased samples which is represented by bold line.

Table 1: Selection of groups and genes for different data set.

Data set ID	Selected group	No of selected groups from selected group	Group	No genes in each group
GDS-2643	4	19	1	2741
			2	2856
			3	2691
			4	2476
			5	2786
			6	2813
			7	2784
			8	2673
GDS-2501	7	18	1	2453
			2	2781
			3	2843
			4	2565
			5	2374
			6	2859
			7	2987
			8	2471
GDS-3057	5	17	1	2814
			2	2957
			3	2607
			4	2576
			5	2696
			6	2878
			7	2598
			8	2770

In the case of second gene expression data set (GDS-2501), we have found eight groups which contenting 2453, 2781, 2843, 2565,2374, 2859, 2987 and 2471 genes. Similarly, eight groups (containing 2814, 2957, 2607, 2576, 2696, 2878, 2598 and 2770) for leukaemia gene expression data (GDS-3057) have been found. An applying DNN, we have found the groups containing 2856 genes for GDS-2643, 2859 genes for GDS-2501 and 2957 genes for GDS-3057 expression data to be the best groups. At last we have found 24, 21 and 22 most important genes corresponding to the data set using SVM procedure. When using SAM procedure, these numbers are 23, 21, and 21 and when using SNR procedure, these numbers are 25, 27 and 22 for three types of data sets. The numbers of genes that are obtained by SVM, SAM and SNR are 19, 18 and 17 corresponding to these data sets.

Validation of the result

In this section, we show the validation of the result in two biochemical pathway namely statistical validation and biological validation which is describe the next sub subsection 5.1.1 and 5.1.2 respectively.

Statistical validation: We compare the results obtained by several types of methods including DNN in this section. In this comparison we can use biochemical pathways, p - value, t - test, F - score and sensitivity. Using some earlier research we have also try validate

some our results. We have considered several types of biochemical pathway that are related in leukaemia gene expression data. From NCBI database (http://ncbi.nlm.nih.gov/projects/geo/). We have found this pathway for leukaemia gene expression data. Thus, we have been determinate to compare the result of this leukaemia gene expression data. In order to validate the results statistically, we have performed t - test on the genes identified by DNN on each data sets. t - Test is the statistical significance which indicates whether or not the difference between two groups average most likely reflects a original difference in the population from which the group wear sampled. The t - value show the most significant genes (99.9%) which p - value < 0.001. For these three types of data set we can apply t - test and we get corresponding t - value. 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 (4.56), 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% level of significance). 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 exceed the t-value for $P = 0.01$. It indicated 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%. Table 2, 3 and 4 shows the list of genes and corresponding t - values of each data set.

For first leukaemia data set (GDS-2643), we have found non-small leukaemia and small leukaemia pathway. In these two pathways a set of 472 genes is involved. This set of genes we have compared obtained by three methods. The result of DNN, we have identified 297 and numbers of genes are common in database information. We have said these genes are True Positive (TP) genes. On the other hand we have found 104 numbers of genes that are in the set of 472 genes respectively which is obtained by DNN but not present in the pathway. These 104 and numbers of gene are said False Positive (FP) and the number are False Negative (FN) gene is 100 for DNN. Similarly, for other methods we have calculated the number of True Positive, False Positive and False Negative genes. From figure 3, compared to all other methods, it is comprehensible that DNN have been efficient to identify more number of True Positive genes but less number of False Positive and False Negative genes.

For second leukaemia gene expression data (GDS-2501), we have found 316 numbers of genes that are existent in leukaemia related pathway. From figure 3, it is comprehensible noticed that DNN perform better along with others 3 methods. Similarly we have found 218 numbers of leukaemia gene expression data sets (GDS-3057). It is clearly show that DNN perform better than others 3 methods.

Biological validation: The disease mediating gene list and corresponding to an earmarked disease can be obtained in NCBI database (http: //ncbi.nlm.nih.gov/projects/geo/). The list is composing in terms of relevancy of the gene. For leukaemia we have got several sets of genes. The database results in349, 381 and 378 numbers of genes for GDS-2643, GDS-2501 and GDS-3057 respectively. For GDS-2643, we have identified 349 numbers of genes each by using DNN procedure. This set of genes we have compared with 349 numbers of genes from NCBI and we can identified 247 numbers of gene for DNN procedure which is common in both set. We said that these genes are True Positive (TP) genes. On the other hand (349 - 247) = 102 numbers of genes for DNN, are not in the list which is obtained from NCBI. We said these genes are False Positive (FP). Similarly (349 - 247) = 102 number of genes that are present in the NCBI list but not in the set of genes which is obtained by DNN. In this reason these genes are call False Negative (FN). Likewise, we have compared our results with other 3 methods, viz., SVM, SAM, and SNR. Figure 4 show the corresponding results. Further our result

Table 2: list of 68 genes (GDS-2643) (the figure within bracket after each gene are the corresponding *t*-values).

Genes(corresponding *t*-value)	IARS(5.98),TYMS(3.96),MMP25(4.58),MLX(5.32),HPS6(5.59),CALCA(4.12),HIC2(5.02), ANP32B(4.56),TFPI(5.72), NCF1C(3.39),CRYAB(3.98),HNRNPH1(4.92),ERCC5(3.12), PRDM2(3.17),TPT1(3.29),PRIM2(2.61),RPS26(2.83),P RPSAP2(3.57),EFCAB11(3.22), GDPD3(2.84),APP(3.82),RPL18AP3(2.82),UTY(5.46),MTMR9(7.71),TCF7L2(4.16), UBXN1(6.96),HAL(2.96),PRH2(7.38),SYT5(2.56),ZBED4(5.36),PPP2R5A(4.48), OPRM1(7.91),TIPIN(3.40),ZBTB33(3.69),POLRMT(3.75),NCLN(4.26),CALD1(5.16), LMF1(3.46),PDK3(3.84),HCFC1(3.32),RCE1(3.58),SKP1(3.95),C4B PA(5.84),PURA(3.50), USP34(2.42),ARFRP1(3.32),DCAF16(3.18),SEPT9(3.88)ETV6(4.10),LIME1(2.16), KMT2A(2. 56),IL5(2.46),GAP43(7.32),MED17(2.34),MAPK1(2.42),PIK3CB(2.05), NMD3(2.34),ARG2(2.19),EXOC3(2.16),WHS C1(2.18),RFC4(2.26),GLB1L(2.41), FLG(1.97),TXNL1(1.82), RIN3(1.95), CYBB(2.04), ZNF814(1.72), KLF4(1.28)

Table 3: list of 55 genes (GDS-2501) (the figure within bracket after each gene are the corresponding *t*-values)

Genes(corresponding *t*-value)	RRAS2(5.93),ZDHC17(3.86),MGAT5(4.38),TNPO1(6.32),TFRC(5.51),H3F3AP4(4.72), RCN3(5.22),ZNF287(3.56),MPL(5.52),KRT 10(3.49),ACTB(3.68),IGLL3P(5.92), MLST8(3.32),GCDH(3.17),CENPN(4.29),KCNE2(3.61),GGA2(2.53),CD44(3.37), GAPDH(3.2 8),KMT2A(2.84),EEF2(3.62),GGTLC1(2.82),CA9(5.86),SFRP4(7.51), ASCL1(4.36),KLK14(6.76),DUSP26(3.96),CD6(7.38),ALCA M(2.26),TACSTD2(4.36), ELK4(4.41),EHD3(6.91),GPL96(3.45),PXN(3.69),GIT1(3.45),CACNB2(4.76), ZNF335(5.36),P7CRA(3.16),VKORC1(3.44),CENPU(3.32),BTN3A2(3.68),FGD6(3.75), CCL1(5.14),SLC7A8(3.52),FRZB(3.42),AKR1B1(3.32),UBE2Z(3.38),O GN(3.48), ESR1(4.18),ZNF835(2.86),TWF2(2.51),MECOM(2.46),TUBB3(7.92),AP1S2(2.84), HOXA11(2.82),

Table 4: list of 63 genes (GDS-3057) (the figure within bracket after each gene are the corresponding *t*-values)

Genes(corresponding *t*-value)	GGCX(5.18),LLS(3.66),TMEM177(5.58),ZNF200(4.62),ETV1(5.19),TAF4(4.52), DSN1(5.81),MUS5B(4.51),CAP2(5.02),ATP2B2(3.3 9),TDG(3.28),KRBOX4(4.22), GNAQ(3.82),CYB561(3.29),FLRT1(3.89),RHOQ(2.61),CHD3(2.23),GOSR1(3.21), USP13(3.92),FRM D8(2.24),VAV13(3.02),KCNJ9(2.18),CALHM2(5.76),CTIF(7.71), CAP2(4.96),SPG7(5.56),PLA2G4B(2.36),AJAK2(7.38),GJB1(2.28) ,HCRP1(4.36), BMP7(4.98),TTC12(7.11),IKZF5(3.49),MARCH6(3.23),GRB10(3.75),BTCC1(4.96), ERAP1(5.28),PRPF23B(3.16),A RHGEF16(3.24),CD24(2.32),ZNF358(3.28),MED24(3.25), NAB2(5.44),BICC1(3.28),MARCH5(2.82),SDF2(3.18),CCNF(3.58),EPB4 1L3(3.18), ALOX12P2(4.10),PAVLB(2.76),PTGDS(2.26),PSMA2(2.81),BTNL2(6.12), SH2D3C(2.84),PRODH2(2.18),SPACA1(2.55), MROH7(2.84),RAP2B(2.72), ATG4B(2.71),ABL1(2.29),DSC2(2.26), CIB2(1.27), UQCRH(1.27)

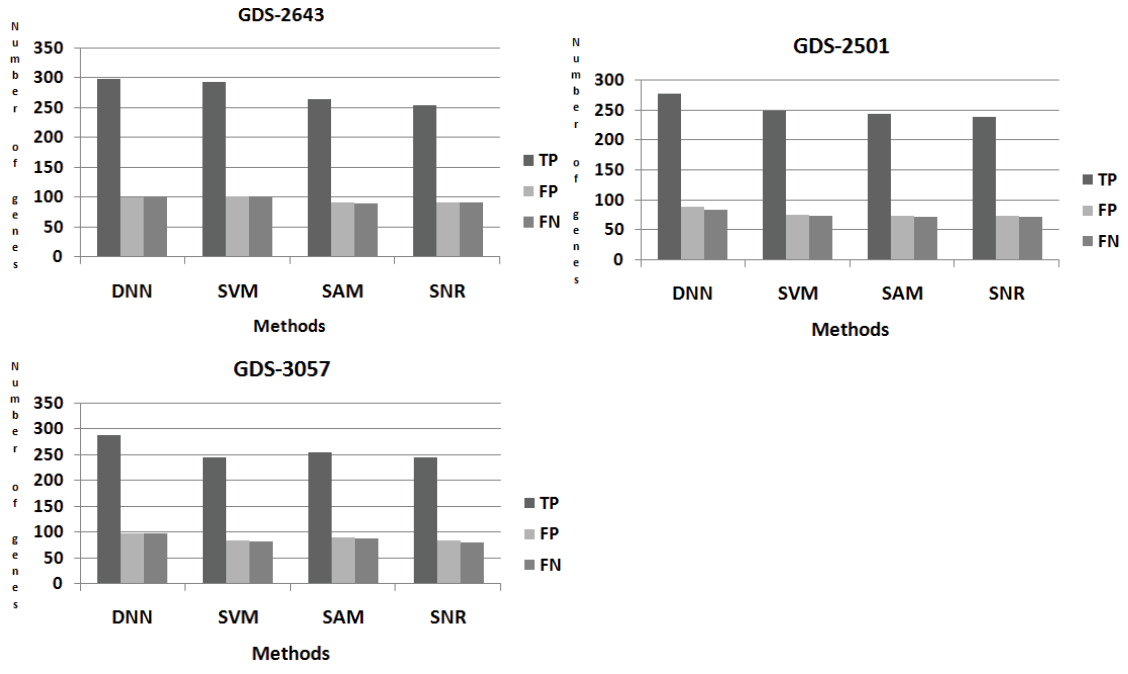

Figure 3: Comparison among the method Here TP, FP, FN indicates the number of True Positive, False Positive, False Negative respectively.

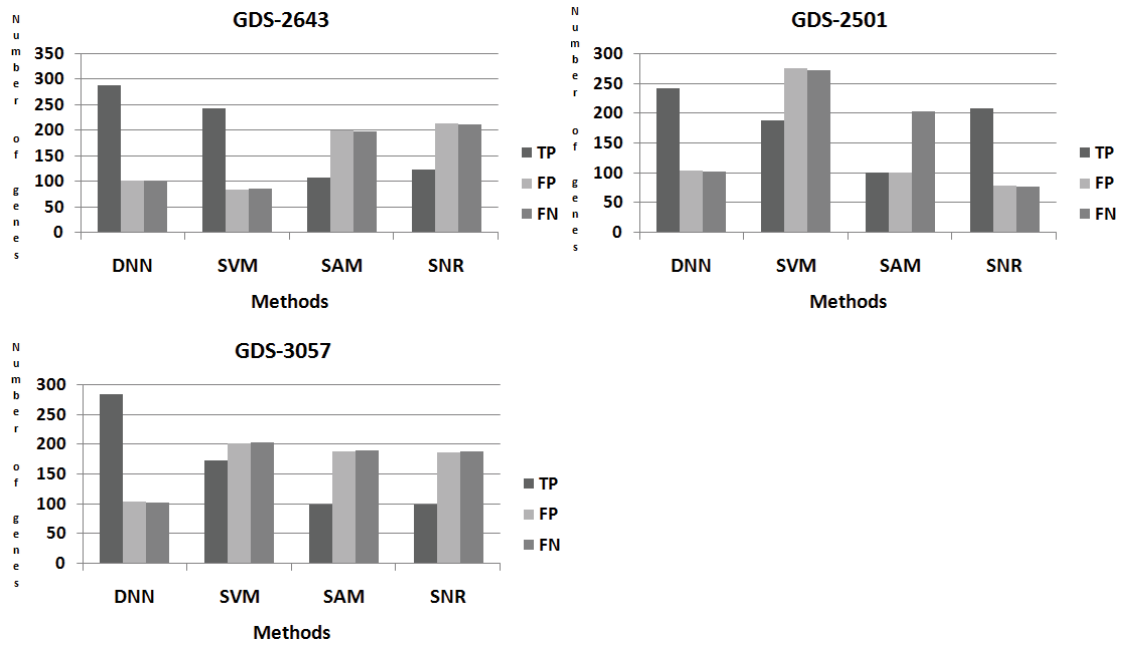

Figure 4: Comparison among the methods using NCBI data base. Here TP, FP, FN indicate True Positive, False Positive, False Negative respectively.

are validate, we have performed Sensitivity on the gene expression data sets. At first we have calculated the numbers of True Positives and numbers of False Negatives corresponding to every method for each data set. Sensitivity is calculated using the following equation

$$Sensitivity = \frac{TP}{TP + FN} \tag{16}$$

Figure 5 show Sensitivity of DNN methodology is much higher than the others existing methods. Figure 5 also comprehensible notice that DNN procedure are performed the best result compared with SVM, SAM and SNR methods.

CONCLUSION

In this article, we have provided a procedure based neural network model for identifying genes that's under or over expression may be normal or may be diseases. The procedure is identifying genes by using Deep Neural Network (DNN) model. The utility of the procedure have been demonstrated three type human leukaemia gene expression data sets. Most important genes are obtained by the procedure have also been corroborated using their p -value. The best performance of the procedure compared to the three existing methods like SVM, SAM and SNR. The result have been corroborated

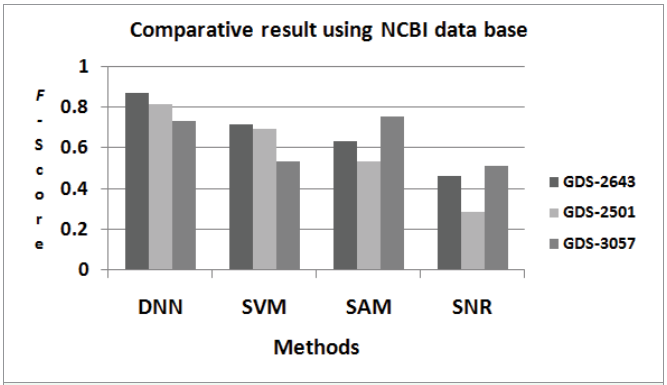

Figure 5: Comparison among the methods using NCBI data base in terms of *F-score*.

using biochemical pathway, *p* - value, t - test, F - score, sensitivity and some existing result expression profile plots. It have been found that the methodology has been able to the genes are more significant.

REFERENCES

1. Baqutayan S, Gogilawan, Mahdzire A, Rahman AH. Cancer Awareness in Malaysia. The Pharma Innovation - Journal. 2013; 2: 81-85. https://goo.gl/DSzlqI
2. Suzuki H, Shigeta A, Fukunaga T. Death resulting from a mesenteric hemorrhage due to acute myeloid leukaemia: an autopsy case. International Journal of Legal Medicine. 2014; 16: 373-375. https://goo.gl/qJ6OrZ
3. Lin PH, Lin CC, Yang HI, Li LY, Bai LY, Chiu CF. Prognostic impact of allogeneic hematopoietic stem cell transplantation for acute myeloid leukaemia patients with internal tandem duplication of FLT3. International Journal of Leukaemia Research. 2013; 37: 287-292. https://goo.gl/gK2KjK
4. Seewald L, Taub JW, Maloney KW, McCabe ER. Acute leukaemia's in children with down syndrome. International Journal of Molecular Genetics and Metabolism. 2012; 107: 25-30. https://goo.gl/n8qzcZ
5. Izraeli S. The acute lymphoblastic leukaemia of Down syndrome - Genetics and pathogenesis. International Journal of European Journal of Medical Genetics. 2016; 59: 158-161. https://goo.gl/VLp0k4
6. Zeidner JF, Karp JE. Clinical activity of alvocidib (flavopiridol) in acute myeloid leukaemia. International Journal of Leukaemia Research. 2015; 39: 1312-1318. https://goo.gl/Bu1eCK
7. Sielken RL Jr, Valdez-Flores C. A comprehensive review of occupational and general population cancer risk: 1,3-Butadiene exposure C response modeling for all leukaemia, acute myelogenous leukaemia, chronic lymphocytic leukaemia, chronic myelogenous leukaemia, myeloid neoplasm and lymphoid neoplasm. International Journal of Chemico-Biological Interactions. 2015; 241: 50-58. https://goo.gl/QH8GvW
8. Ripperger T, Schlegelberger B. Acute lymphoblastic leukaemia and lymphoma in the context of constitutional mismatch repair deficiency syndrome. International Journal of European Journal of Medical Genetics. 2016; 59: 133-142. https://goo.gl/E4Hdg7
9. Mundra PA, Rajapakse JC. Gene and sample selection using T-score with sample selection. International Journal of Journal of Biomedical Informatics. 2016; 59: 31-41. https://goo.gl/XhZ8xB
10. Elyasigomari V, Mirjafari MS, Screen HRC, Shaheed MH. Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization. International Journal of Applied Soft Computing. 2015; 35: 43-51. https://goo.gl/jcRpwB
11. Mishra S, Mishra D. SVM-BT-RFE: An improved gene selection framework using Bayesian T-test embedded in support vector machine (recursive feature elimination) algorithm. International Journal of Karbala International Journal of Modern Science. 2015; 1: 86-96. https://goo.gl/QD0d98
12. Du D, Li K, Li X, Fei M. A novel forward gene selection algorithm for microarray data. International Journal of Neurocomputing. 2014; 133: 446-458. https://goo.gl/OZPFPz
13. Narayanan A, Keedwell CE, Gamalielsson j, Tatineni S. Single-layer artificial neural networks for gene expression analysis. International Journal of Neurocomputing. 2004; 61: 217-240. https://goo.gl/oZMdSm
14. Wang HQ, Wong HS, Zhu H, Yip TT. A neural network-based biomarker association information extraction approach for cancer classification. International Journal of Biomedical Informatics. 2009; 42: 654-666. https://goo.gl/6BSpG3
15. Chakraborty S. Simultaneous cancer classification and gene selection with Bayesian nearest neighbour method: An integrated approach. International Journal of Computational Statistics and Data Analysis. 2009; 53: 1462-1474. https://goo.gl/878apR
16. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Computational Biology and Bioinformatics. 2012; 9: 1106-1119. https://goo.gl/pBcB9i
17. Mundra PA, Rajapakse JC. SVM-RFE with MRMR Filter for Gene Selection. IEEE Trans Nanobiosci. 2010; 9: 31-37. https://goo.gl/1czIDZ
18. Ojeda F, Suykens JA, De Moor B. Low rank updated LS-SVM classifiers for fast variable selection. International Journal of Neural Network. 2008; 21: 437-449. https://goo.gl/hfMQPI
19. Tang Y, Zhang YQ, Huang Z. Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. IEEE/ACM Trans Computational Biology and Bioinformatics. 2007; 4: 365-381. https://goo.gl/Cm1gAV
20. Tang Y, Zhang YQ, Huang Z, Hu X, Zhao Y. Recursive fuzzy granulation for gene subset extraction and cancer classification. IEEE Trans on Information Technology in Biomedicine. 2008; 12: 723-730. https://goo.gl/SjVHRw
21. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006; 7: 1-13. https://goo.gl/B7nW67
22. Zou H, Hastie T. The regularization and variable selection via the elastic net. J R Statist Soc B. 2005; 67: 301-320. https://goo.gl/osDzDq
23. Deng H, Runger G. Gene selection with guided regularized random forest. Pattern recognition. 2013; 46: 3483-3489. https://goo.gl/RJXcG1
24. Anaissi A, Kennedy PJ, Goyal M, Catchpoole DR. A balanced iterative random forest for gene selection from microarray data. BMC Bioinformatics. 2013; 14: 1-10. https://goo.gl/AzJR18
25. Sharma MC, Tuszynski GP, Blackman MR, Sharma M. Long-term efficacy and downstream mechanism of anti-annexina2 monoclonal antibody (anti-ANX A2 mAb) in a pre-clinical model of aggressive human breast cancer. International Journal of Cancer Letters. 2016; 373: 27-35. https://goo.gl/WDoq5x
26. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural computation. 2006; 18: 1527-1554. https://goo.gl/ZvYVmp
27. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. The Journal of Machine Learning Research. 2010; 11: 3371-3408. https://goo.gl/bbDsge
28. Rifai S, Vincent P, Muller X, Glorot X, Bengio Y. Contractive auto encoders: Explicit invariance during feature extraction. In Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011; 833-840. https://goo.gl/ouXLgE
29. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. International Conference on Artificial Intelligence and Statistics. 2011; 315-323. https://goo.gl/Re3Dyv
30. Bengio Y, Delalleau O. Justifying and generalizing contrastive divergence. Neural Comput. 2009; 21: 1601-1621. https://goo.gl/sOJuQn
31. Hinton EG, Sejnowski JT. Learning and relearning in Boltzmann machines. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. 1986; 1: 282-317. https://goo.gl/PpSnMh