Artificial Neural Network Based Method for Classification of Gene Expression Data of Human Diseases along with Privacy Preserving

S.Sathish Kumar[1], Dr N Duraipandian[2]

[1]Research Scholar, Dr MGR Educational and Research Institute University, Chennai, Tamil Nadu, India

Email: Sathish_tri@yahoo.com

[2]Vice Principal, Velammal Institute of Technology, Chennai, Tamil Nadu, India

E-mail: emailpandiandurai@gmail.com

ABSTRACT

In this paper, the author introduces a classification approach using Artificial Neural Network(ANN) with Back-Propagation learning technique for human diseases like Cancer and heart problems from clinical diagnosis data. Clinical diagnosis is done mostly by experienced doctors with expertise in this field. In many cases, the test results are not effective towards the diagnosis of the disease. The author is particular about the wrong diagnosis which leads to a wrong treatment. The author is using Artificial Neural Network technique to classify the disease with reduced number of DNA sequence. The accuracy is differing based on the training data set and validation data set. The other major issue is the privacy preserving of the patients. As we are sharing the critical data from clinical diagnostic centers, there is good chance of patient's anonymity is revealed. To avoid this, the author is using a simple Privacy Preserving in Data Mining (PPDM) technique to crypt the identity of the patients as well as the critical data and discloses only the required data like DNA sequence to the research team, as they are not much interested in the identity or the owner of the diagnosis report.

***Keywords**: Data mining, classification, Cancer disease, Artificial neural networks, PPDM, Back propagation*

1. INTRODUCTION

As we are entering into the post genomic era by successful completion of the Human Genome Project, we are facing to handle a mass amount of data produced either from the research laboratories as well as the clinical research centers. First we will talk about the DNA sequence. The DNA is the material present in every living organism that forms the genes. Genes are carrying the instructions for the body growth as well as the regular functions. Throughout the life, our cells are exposed to various stimulations. These stimulation agents induce some changes in the DNA sequences. For example, Chemicals, Radiations as well as Viruses are the few which will act as stimulation agent for the changes in DNA sequences. The permanent changes are called as mutations. Sometimes, temporary mutations may happen in normal cell replication process. But the normal healthy cell will repair these changes by itself, virtually. Some changes that are not repaired by the normal cells continue to exist as mutations. When the single cell acquires enough mutations in the DNA sequence, it begins to behave in an abnormal way. This is the characteristic of cancer. This will lead to uncontrolled cell growth and ultimately spread throughout the body.

Two types of genes are particularly involved in the development of cancer. They are referred as proto-oncogenes and tumor-suppressor genes. In this, **proto-oncogene** is a normal gene that can become an oncogene due to mutations or increased expression. Once this gene is active, a proto-oncogene becomes a tumor-inducing agent, that is, an oncogene. So, Proto-oncogenes code for proteins that help cells grow and divide, at the same time the tumor-suppressor genes help to stop cell growth whenever necessary. In general, the mutations that cause cancer are mutations in regular cell growth related genes.

Many mutations in these genes have been identified and have been discovered to be associated with one or more types of cancer. Today, the collaboration between the researchers is required to identify the catalogue of cancer mutations in cells, for different cancer types. These details will help the researchers to understand all forms of cancer and to find new as well as effective way of controlling those diseases. Pharmaceutical companies have already succeeded in developing therapies that will attack only the cancer cells with highest efficiency, for some cancer types. Clinical trials are already started for several of these drugs, the details like, the type of tumor and changes happened in those patients after the treatment are extensively recorded. As a result, a new era of personalized care is emerging in which information about the genomic changes specific to the person's cancer will help them and their doctor determine what treatments are most likely to work for them. Now the details from the clinical research centers are collected and they are used as the test data for training our neural networks. After training the ANN, the actual details of the patients are entered. Even after a slight noise in the patient's records, we need an efficient method to identify the intensity of the disease as well as the category of that so as to find out the type of treatment the doctor can give to them. This is coming under the classification technique of our ANN in data mining.

We will explain how the classification works. Data classification is a two-step process. In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or "learning from" a training set made up of database tuples and their associated class labels. In the second step, the model is used for classification. First, the predictive accuracy of the classifier is estimated. If we were to use the training set to measure the accuracy of the classifier, this estimate would likely be optimistic, because the classifier tends to over fit the data. Therefore, a test set is used, made up of test tuples and their associated class labels. These tuples are randomly selected from the general data set. They are independent of the training tuples, meaning that they are not used to construct the classifier.

Data mining is a crucial step in discovery of knowledge from large data sets. In recent years, Data mining has found its significant hold in every field including health care. Medical history data comprises of a number of tests essential to diagnose a particular disease [1]. Clinical databases are elements of the domain where the procedure of data mining has develop into an inevitable aspect due to the gradual incline of medical and clinical research data. Because of this, it is possible for the healthcare industries to gain advantage of Data mining by employing the same as an intelligent diagnostic tool. The industries can acquire knowledge and information concerning a disease from the patient specific stored measurements as far as medical data is concerned. Therefore, data mining is the vital domain in healthcare [2]. It is possible to predict the efficiency of medical treatments by building the data mining applications. The data mining techniques have been utilized by a wide variety of works in the literature to diagnose various diseases including: Diabetes, Hepatitis, Cancer, Heart diseases and the like. Information associated with the disease, prevailing in the form of electronic clinical records, treatment information, gene expressions, images and more; were employed in all these works

2. LITERATURE SURVEY

Few works have been published by XindongWu , Vipin Kumar · J, Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan. They have proposed 10 top best algorithms for datamining like C4.5, *k*-Means, SVM, Apriori, EM, PageRank, AdaBoost, *k*NN, Naive Bayes, and CART. These top 10 algorithms are among the most influential data mining algorithms in the research community. [1]

Anchana Khemphila, Veera introduces a classification approach using Multi-Layer Perceptron (MLP)with Back-Propagation learning algorithm and a feature selection algorithm along with biomedical test values to diagnose heart disease. They are giving elaborative details about the clinical diagnosis. In their paper they classify the presence of heart disease with reduced number of attributes. Original, 13 attributes are involved in classify the heart disease. The authors use Information Gain to determine the attributes which reduces the number of attributes which is need to be taken from patients. The Artificial neural network is used to classify the diagnosis of patients. [2]

In the paper published by Bing Liu, Qinghua Cui, Tianzi Jiang, and Songde Ma, they are using micro array experiments for clinical diagnosis for discovering gene expression patterns that are specific for a particular disease. To date, this problem has received most attention in the context of cancer research, especially in tumor classification. Various feature selection methods and classifier design strategies also have been generally used and compared in the paper. In this paper, the authors are using a combinational feature selection method in conjunction with ensemble neural networks to improve the accuracy and robustness of sample classification. [3]

The survey paper presented by Alex A. Freitas , discusses the use of evolutionary algorithms, particularly genetic algorithms and genetic programming, in data mining and knowledge discovery. The author is focusing on the data mining task of classification. In addition, the author discusses some preprocessing and post processing steps of the knowledge discovery process. The author is showing the influence of datamining in knowledge discovery and design of evolutionary algorithms by some experimentation. [4]

Thair Nu Phyu is elobarting the basic classification technique in datamining in his survey paper. He is comparing several kinds of classification methods like decision tree induction, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques. The outcome of this paper is the comprehensive review on these techniques in datamining. [5]

In the paper on training the neural networks for medical databases by Maciej A. Mazurowskia, Piotr A. Habasa, Jacek M. Zuradaa, they investigates the effect of class imbalance in training data when developing neural network classifiers. They did investigation on medical data, namely small training sample size, large number of features, and correlations between features. They have explored back propagation algorithm and particle swarm optimization with clinically relevant criteria. An experimental study is performed using simulated data and the conclusions are further validated on real clinical data for breast cancer diagnosis. [6]

As the complete human genome sequence is available, an elaborative study on human genome sequence by the influence of mutations and cancer growth is done by Christopher Greenman, Philip Stephens, Raffaella Smith and others. They report more than 1,000 somatic mutations found in 274 megabases (Mb) of DNA corresponding to the coding exons of 518 protein kinase genes in 210 diverse human cancers. There was substantial variation in the number and pattern of mutations in individual cancers reflecting different exposures, DNA repair defects and cellular origins. Most somatic mutations are likely to be ‘passengers’ that do not contribute to oncogenesis. However, there was evidence for ‘driver’ mutations contributing to the development of the cancers studied in approximately 120 genes. Systematic sequencing of cancer genomes therefore reveals the evolutionary diversity of cancers and implicates a larger repertoire of cancer genes than previously anticipated. [7]

3. THE INTEGRATED TECHNIQUE FOR DISEASE CLASSIFICATION

The proposed disease classification technique classifies disease based on the given DNA sequence. The DNA sequence is comprised of four basic nucleotides, Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). Every species has a long DNA sequence, which is formed by the four nucleotides. The DNA sequence defines the attributes, nature and type of the species. The proposed technique is an integration of data mining and artificial intelligence. In the proposed technique, firstly, nucleotide patterns are mined from the sequence. The mined patterns form a nucleotide pattern database with higher dimension. So, secondly, the dimension of the pattern database is reduced by MPCA. Finally, the dimensionality reduced pattern database is used to train the neural network. The technique is described in the further sub sections. [8]

3.1 Mining Nucleotide Patterns from DNA Sequence

The first and initial stage of the proposed technique mines the nucleotide pattern from the DNA sequence. At this stage, patterns formed by different combinations of nucleotides are mined using a novel mining algorithm. Let g be the DNA sequence, which is a combination of four nucleotides *A*, *G*, *C* and *T*. For instance, a sample DNA sequence is given as *CGTCGTGGAA*. From the sequence, the mining algorithm extracts different nucleotide patterns and their support. The algorithm is comprised of two stages, namely, pattern generation and support finding. In pattern generation, patterns with different length are generated whereas in support finding, support values for every generated pattern are determined from the DNA sequence. The basic structure of the algorithm is given as a block diagram in Figure 1. [8]

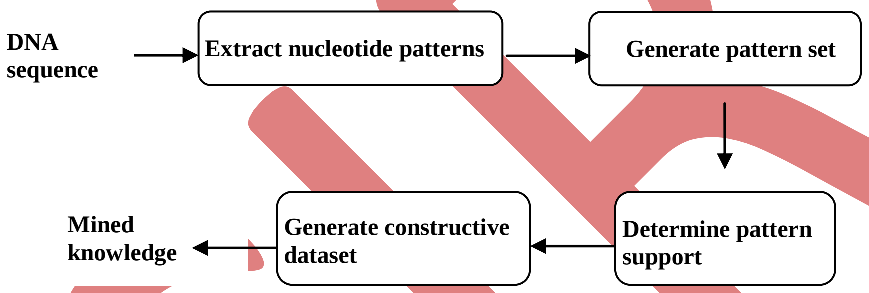
DNA sequence
Extract nucleotide patterns
Generate pattern set
Determine pattern support
Generate constructive dataset
Mined knowledge

Figure 1: Block diagram of the pattern mining algorithm

3.1.1 Data flow chart

To understand the above said concept, we are using the following dataflow diagram to understand our concept very clear. The blocks of this data flow diagram is explained one by one in the following sections.

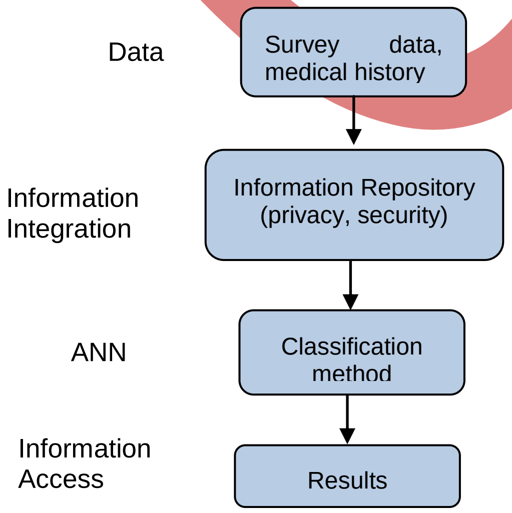

Figure 2: Block diagram for analyzing biomedical Data taken from Research laboratory, etc to get the finding about the disease

3.1.1.1 Data Acquisition

As the genetic datasets are in large-scale, it is significantly a challenge for systems biology to extract all the biological information which is relevant to the bio-informatician. One of these challenges is how to retrieve the relevant information in a timely manner. The main problem is that of classification of the relationships among phenotypes of mutant strains into biological information of gene interaction. Geneticists have determined such classifications based on insights from biological examples. Agent technology and various filtering systems have been employed in order to enhance the relevancy. One of the methods is that depends on maximizing a previously described context-dependent information measure to obtain maximally informative biological networks. The context-dependent information measure is a function only of phenotype data and a set of interaction rules, involving no prior biological knowledge. Personalization will provide extra relevancy to the retrieved information. Analysis of the resulting networks reveals that the most biologically informative networks are those with the greatest context-dependent information scores. The high-complexity networks reveal genetic architecture at a modular level, in contrast to classical genetic interaction rules that order genes in pathways. In our study we are taking data from biomedical research centers and data collected from surveys. [9]

3.1.1.2 Information Integration and Information Repository

Before applying the data mining rules to find out the patterns in our input, we have to clean and filter the necessary data to avoid the creation of deceptive or inappropriate rules or patterns. The preprocessing step contains few stages as normalize the data to match to the internal representation of our database or warehouse and remove the unnecessary records or data fields. After this step, the datamining will become simple as all records resemble each other in their format and only required fields or minimized fields are available for processing. Sometimes, the raw data is changed into data sets with appropriate characteristics and in some cases, the data sets are combined or reduced to fewer sets for minimizing the memory requirement as well as best suited for our algorithm. In our case, we are taking cancer disease data warehouse to refine by removing the duplicate records and supply the missing values. The records are transformed to a form as suitable for classification method. As we are taking records from the research centers and laboratories, there is a good chance that the records contains the personal details of the patients like their name, age, address etc. This leads to disclose of the personal details about the patients. To avoid this we are applying privacy preserving (PPDM) or any anonymity method to remove the sensitive data from the input and hide those details from accessed by the others. [9]

3.1.1.3 Privacy Preserving in Data Mining (PPDM)

Privacy Preserving Data Mining (PPDM) is a novel field in data mining. PPDM is concerned with extraction of information from data warehouse without revealing sensitive information of individuals and company privacy details. Present industry consists of database which is distributed across multiple source locations. Most of them do not want their data to be revealed so that the confidentiality is lost which is of a great concern. The confidentiality of some attribute within the data base must be ensured. There are two major methods in PPDM first by using cryptographic representation and the other by using heuristic algorithms which ensures that sensitive data is not revealed. Most of the current industry require that there data be secured during transmission and also when the data is present in the data warehouse. In this paper we are using a cryptographic representation called Elliptic Curve Cryptography to ensure confidentiality which can be indicated to some attributes during design and data distortion in heuristic algorithm. Extraction Transformation Load (ETL) is an important stage in Data Warehouse and the processing activities can be indicated by means of diagrams as a part of design. Designing ETL and embedding PPDM representation is the main task of this research. Extraction deals with the retrieval of data from various sources. These sources of data can be of different format and different location in space [1]. Each of these locations must be analyzed before it is loaded on to the data warehouse. The most important and complex part of ETL is the Transformation phase. Transformation can be visualized as the change that must be made on to the source data before it is loaded on to the data warehouse. There are various methods that are used for transformation. Usual transformation include deletion of a column, conversion of values from one representation to the other, aggregation of values for better retrieval, removal of null values, converting multiple representations in to a single format and joining of multiple tables. The last phase is the loading phase which copies the information from the transformed data to the warehouse. [10][11][12]

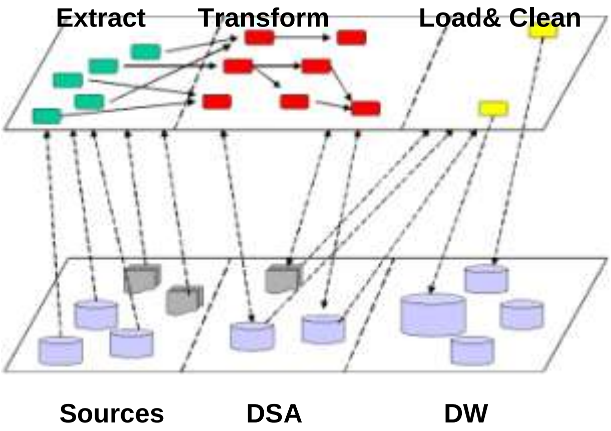

Figure 3: The Environment of ETL processes

The overall process consists of three major steps first identification of the source targets second is the actual data representation also called conversion and the last phase is moving resultant to the target location. These tasks can be done in intermediate stages or as a single logical execution. Each of these stages indicated separately will help improve the overall representation of Data Warehouse. The initial design of conceptual schema plays an important role and also propagates to other stages. The complexity of extraction transformation and load lies in the mapping of data from source to destination. As indicated by Alkis Simitsis[12] the environment of ETL process can be shown in figure. The left side we have source representation where in information is distributed. In the middle we have Data Staging Area (DSA) which encapsulates the major transformation and the last is the Data Warehouse representation.

The sample table which contains the details about the cancer DNA as well as the patient details are available in the following table. As we can see from the table that there are more attributes about the patients and their personal details are present in that. [13]

Table 1: The attributes of the database used for clinical analysis. courtesy: The actual table used by Petitjean A, and team in their original paper " *Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database"*[13]

Mutation_ID		**MUT_ID**	**hg18_Chr17_coordinates**	**ExonIntron**	
Genomic_nt		**Codon_number**		**Description**	
c_description		**g_description**		**g_description_hg19**	**WT_nucleotide**
Mutant_nucleotide	**Splice_site**	**CpG_site**	**Type**	**Mut_rate**	**WT_codon**
Mutant_codon		**WT_AA**		**Mutant_AA**	
ProtDescription		**Mut_rateAA**		**Effect**	
Structural_motif		**Putative_stop**		**Sample_Name**	
Sample_ID		**Sample_source**		**Tumor_origin**	
Topography		**Short_topo**		**Topo_code**	
Sub_topography		**Morphology**		**Morpho_code**	
Grade		**Stage**		**TNM**	
p53_IHC	**Add_Info Individual_ID**	**Sex**		**Age**	
Ethnicity		**Geo_area**	**Country**	**Development**	
Population		**Region**		**TP53polymorphism**	
Germline_mutation	**Family_history**	**Tobacco Alcohol Exposure**		**Infectious_agent**	
Ref_ID		**PubMed**		**Exclude_analysis**	

3.1.1.4 Classification Method

Because of the huge and explosive amount of biological data has been obtained and deposited in various open access databases in the recent years, to analyze those information, many new algorithms as well as programs were developed. The major problem with them is their lack of common standards as they were developed by academic Institutions and commercial companies. They resulted in exhausted program switching and data transformations. In this paper, the author is proposing a method for biomedical data analysis, namely Neural Mining Based Disease Classification (NMBDC). This is nothing but using An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks, in other words, is a replica of biological neural system to do the classification of our medical data. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. The ANN is an adaptive system which changes its structure based on external or internal information.

3.1.1.5 Artificial Neural Networks (ANNs)

Artificial neural networks are inspired by attempts to simulate biological neural systems. The human brain consists primarily of nerve cells called neurons, linked together with other neurons via stand of fiber called axons. Axons are used to transmit nerve impulses from one neuron to another whenever the neurons are stimulated. A neuron is connected to the axons of other neurons via dendrites, which are extensions from the cell body of the neurons. The contact point between a dendrite and an axon is called a synapse. Multilayer is feed-forward neural networks trained with the standard back-propagation algorithm. It is supervised networks so they require a desired response to be trained. It learns how to transform input data in to a desired response, so they are widely used for pattern classification. With one or two hidden layers, they can approximate virtually any input-output map. It has been shown to approximate the performance of optimal statistical classifiers in difficult problems. The most popular static network is the multilayer. The multilayer is trained with error correction learning, which is appropriate here because the desired multilayer response is the arteriographic result and as such known. Error correction learning works in the following way from the system response at neuron j at iteration t, yj (t),and the desired response dj $(t)$for given input pattern an instantaneous error ej $(t)$is defined by $(t) = (t) - (t)$ (1)

Using the theory of gradient descent learning, each weight in the network can be adapted by correcting the present value of the weight with a term that is proportional to the present input and error at the weight, i.e. $(t + 1) = (t) + (t)xk(t)$ (2)

The (t) is the learning-rate patameter. The (t) is the weight connecting the output of neuron k to the input neuron j at iteration t. The local error (t) can be computed as a weighted sum of errors at the internal neurons. [8][9]

3.1.1.6 Classification using ANN

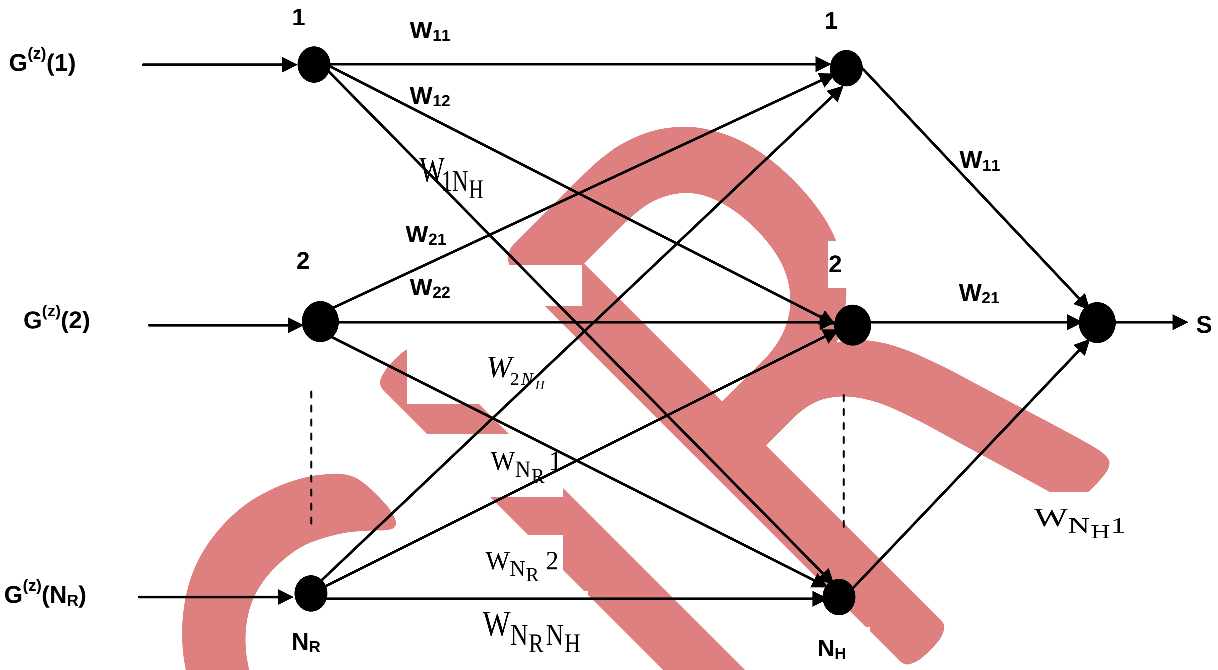

Figure 4: The multilayer feed forward neural network used in the proposed technique

Before performing any task, the ANN must be trained. Once trained, the ANN capably identifies the species by finding the class of the gene sequence. The training phase and classification phase of the ANN are done by back propagation algorithm. [8][9]

3.1.1.7 The Back Propagation Algorithm

The algorithm is represented roughly in the following steps:

1. First apply the inputs to the network and work out the output – note that the initial output could be anything and the initial weights were random numbers.

2. Next work out the error for neuron X. The error is What was the actual planned value – What we get from the present calculation. In other words: ErrorX = OutputX (1-OutputX)(TargetX – OutputX)

The "*Output(1-Output)*" term is necessary in the equation because of the Sigmoid Function – if we were only using a threshold neuron, it would just be *(Target –Output)*.

3. Change the weight. Let W+AB be the new (trained) weight and WAB be the initial weight.

W+AB = WAB + (ErrorB x OutputA)

Notice that it is the output of the connecting neuron (neuron A) we use (not B). We update all the weights in the output layer in this way.

4. Calculate the Errors for the hidden layer neurons. Unlike the output layer we can't calculate these directly, so we *Back Propagate* them from the output layer. To do this, take the Errors from the output neurons and running them back through the weights to get the hidden layer errors. Example: if neuron A is connected as shown to B and C then we take the errors from B and C to generate an error for A.

ErrorA = Output A (1 - Output A)(ErrorB WAB + ErrorC WAC)

Again, the factor "*Output (1 - Output)*" is present because of the sigmoid squashing function.

5. Having obtained the Error for the hidden layer neurons now precede as in stage 3to change the hidden layer weights. This method is repeated for as many layers as required. [14]

3.1.1.8 Information Access

In this layer, the job is very simple as it integrates the results collected from the ANN tool and put them in proper format. The output will be either a table which contains the results accumulated from the previous step or it may be an excel sheet which contains just the results collected from the previous step. The excel sheet information is further used to analyze for coming to a conclusion. In our work, we are producing graphs as well as table to come to a conclusion.

3.2 EXPERIMENTAL RESULTS

Our first example for analysis is PIK3CA and the second one is TP53 gene. Recent evidence has shown that the PIK3CA gene is mutated in a range of human cancers and the **p53** (also known as **protein 53** or **tumor protein 53**), is a tumor suppressor protein that in humans is encoded by the *TP53*gene. The evaluation process is performed using 10-fold cross validation test. Here, nucleotide patterns are mined with $L_{\max} = 5$. The nucleotide patterns for L = 2 and 3, and their corresponding support are given in Table. In Figure, different length patterns and their support are depicted and the constructive dataset that is generated from the pattern set is given in Table.

Table 2: Mined nucleotide patterns from the DNA sequence of Brucella Suis and C.elegans (a) $l = 2$ and (b) l=3

S. No	Pattern	Genes	
		Support	
		PIK3CA	TP53
1	aa	169042	168149
2	ag	56284	59645
3	ac	53509	54824
4	at	100894	101778
5	ga	72354	72651
6	gg	45341	45368
7	gc	46001	43023
8	gt	53423	56002
9	ca	67662	69882
10	cg	47630	40344
11	cc	47205	44205
12	ct	57377	58316
13	ta	70670	73713
14	tg	67864	71687
15	tc	73159	70695
16	tt	171584	169717

S. No	Pattern	Genes	
		Support	
		PIK3CA	TP53
1	aaa	86090	83349
2	aag	17627	18623
3	aac	18374	19422

4	aat	46951	46755
5	aga	19089	19871
6	agg	11100	10442
7	agc	11318	12264
8	agt	14777	17068
9	aca	17239	17901
10	acg	10577	9453
11	acc	10491	10676
12	act	15202	16794
13	ata	20142	21368
14	atg	15783	16752
15	atc	17406	16451
16	att	47563	47207

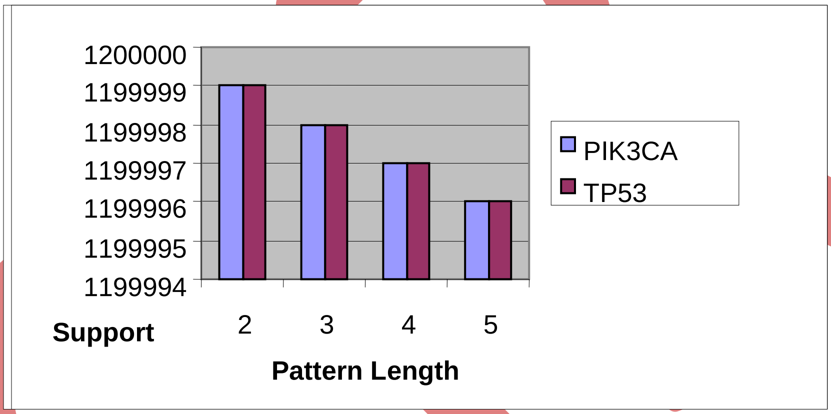

Figure 5: Support obtained for different length patterns

Once the training process has been completed, the technique is validated using the test sequence. The results obtained from 10-fold cross validation are given in Table.

Table 3: Performance evaluation using 10-fold cross validation results

Rounds in cross validation	Genes			
	PIK3CA		TP53	
	ANN Output	Classification Result	ANN Output	Classification Result
1	0.2421	TP	0. 5171	TP
2	0.0769	TP	0. 6272	TP
3	0.0828	TP	0. 6361	TP
4	0.2634	TP	0. 8974	TP
5	0.2493	TP	0. 6063	TP
6	0.2613	TP	0. 0141	TN
7	0.5277	TN	0. 9163	TP
8	0.3616	TP	0. 6714	TP

9	0.5849	TN	0. 5103	TP
10	0.2143	TP	0. 5142	TP
Mean Classification Accuracy	**80%**		**90%**	

From the results, it can be seen that when a gene sequence is given to the proposed technique it identifies the corresponding genes. Here, the technique is evaluated with the DNA sequence of only two genes. The technique is developed in such a way that it can be applied to any kind of DNA sequence. The test results claim that the performance of the technique reaches a satisfactory level.

3.3 CONCLUSION

In this paper, we have proposed a disease identification technique by integrating data mining technique with artificial intelligence. Initially, the nucleotide patterns have been mined effectively. The resultant has been subjected to dimensionality reduction and eventually classified using a well-trained neural network. The implementation results have shown that the proposed technique effectively identifies the gene from its gene sequence and so the disease. Moreover, results obtained from 10-fold cross validation have proved that the disease can be identified even from a part of the DNA sequence. Though the technique has been tested with the DNA sequence of only two genes, the 10-fold cross validation results have reached a remarkable performance level. From the results, it can be hypothetically analyzed that a technique, which identifies the disease only with a part of gene sequence, has the ability to classify any kind of disease.

REFERENCES

[1] survey paper on “top 10 algorithms in datamining” by XindongWu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, Knowledge Information System (2008) 14:1–37, DOI 10.1007/s10115-007-0114-2

[2] Anchana Khemphila, Veera Boonjing , “Heart disease Classification using Neural Network and Feature Selection”, 21st International Conference on Systems Engineering, 2011

[3] Bing Liu, Qinghua Cui, Tianzi Jiang* and Songde Ma, “A combinational feature selection and ensemble neural network method for classification of gene expression data”, BMC Bioinformatics 2004, 5:136 doi: 10.1186/1471-2105-5-136

[4] Alex A. Freitas “A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery”, uran.donetsk.ua

[5] Thair Nu Phyu “Survey of Classification Techniques in Data Mining”, Proceedings of the International Multi Conference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong.

[6] Maciej A. Mazurowskia, Piotr A. Habasa, Jacek M. Zuradaa, Joseph Y. Lob, Jay A. Bakerb, and Georgia D. Tourassib, “Training Neural Network Classifiers for Medical Decision Making: The Effects of Imbalanced Datasets on Classification Performance”, Neural Network. 2008; 21(2-3): 427–436.

[7] Christopher Greenman, Philip Stephens, Raffaella Smith and others, “Patterns of somatic mutation in human cancer genomes”, Nature. 2007 March 8; 446(7132): 153–158. DOI: 10.1038/nature05610.

[8] Sathish Kumar S, Dr. N. Duraipandian, “ Artificial Neural Network based String Matching Algorithms for Species Classification – A Preliminary Study and Experimental Results”, International Journal of Computer Applications (0975 – 8887) ISSN: 0975 – 8887, Volume 52– No.14, August 2012.

[9] Sathish Kumar S, Dr. N. Duraipandian, “ Architecture for NM Based Species Classification”, Wulfenia Journal, ISSN: 1561 – 882X, Volume 20, No. 3;Mar 2013

[10] Kiran P, S Sathish Kumar and Dr Kavya N P, “A Novel Framework using Elliptic Curve Cryptography for Extremely Secure Transmission in Distributed Privacy Preserving Data Mining”, Advanced Computing: An International Journal (ACIJ), ISSN : 2229 - 6727 [Online] ; 2229 - 726X [Print], Vol.3, No.2, March 2012

[11] Kiran P, S Sathish Kumar and Dr Kavya N P, “An Extended Conceptual Modeling for ETL Processes in Privacy Preserving Data Mining”, International Journal of Future Computer and Communication, Volume 1, No: 3, Oct 2012, ISSN**:** 2010-3751

[12] P.Vassiliadis, A.Simitsis, S.Skiadopoulos, Conceptual Modeling for ETL Processes, In Proc. of the 5th ACM Int. Workshop on Data Warehouising and OLAP, McLean, USA, pp. 14-21, Nov 2002.

[13] Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, Hainaut P, Olivier M. *Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database.* Hum Mutat. 2007 Jun;28(6):622-9.

[14] The back propagation algorithm, “ www4.rgu.ac.uk/files/chapter3”