Human Liver Cancer Classification using Microarray Gene Expression Data

P. Rajeswari
Research Scholar
Department of Computer Science,
PSGR Krishnammal College for Women,
Coimbatore, India

G. Sophia Reena
Head of the Department
Department of Computer Applications,
PSGR Krishnammal College for Women,
Coimbatore, India

ABSTRACT
Cancer is one of the dreadful diseases, which causes considerable death rate in humans. Cancer is featured by an irregular, unmanageable growth that may demolish and attack neighboring healthy body tissues or somewhere else in the body. There are dissimilar techniques lives for the naming of cancer but none of those techniques afford considerable accuracy of detection. Therefore a new method is highly essential for the cancer classification with improved accuracy. Gene expression profiling by microarray method has been emerged as an efficient technique for classification and diagnostic prediction of cancer nodules. In recent times, DNA microarray technique has gained more attraction in both scientific and in industrial fields. The DNA microarrays are utilized in this paper for the purpose of identifying the presence of cancer. Statistical ranking has also been used for effective cancer classification. The most widely used ranking schemes are ANOVA, T-score and Enrichment Score. But, these existing techniques suffer from the drawbacks of less accuracy, complexity etc. This paper uses liver cancer data set for experimentation of the proposed technique. The classifier used here is SVM and FNN. The experimental results shows that the proposed technique has the ability to classify the cancer cells significantly when compared to the conventional methods of cancer classification.

Keywords
Microarray Dataset; Enrichment Score; Correlation Based Ranking;MAPSTD;SVM;FNN;

1. INTRODUCTION
Cancer is one of the dreadful diseases found in most of the living being, which is one of the challenging studies for research in the 20th century. There has been lot of proposals from various researchers on cancer classification and detailed study is still on in the domain of cancer classification.

In order to gain deep insight into the cancer classification problem, it is necessary to take a closer look at the problem, the proposed solutions and the related issues all together.

1.1 Challenges in Cancer Classification
There have been various investigations available in the literature on the classification problem by the statistical, machine learning and database research community.

However, gene classification as a new area of research has new challenges due to its unique problem nature. Some of the challenges are summarized below:

The unique nature of the available gene expression data set is the foremost challenge. Third challenge is the huge number of irrelevant attributes (genes). Fourth challenge arises from the application domain of cancer classification. Though Accuracy plays a vital factor in cancer classification, the biological relevancy is another key criterion, as any biological information exposed during the process can help in added gene function discovery and other biological examinations.

Micro array data analysis has been effectively applied in a number of investigations over a wide range of biological disciplines, which comprises of cancer classification by class detection and prediction, recognition of the unknown effects of a specific therapy, recognition of genes suitable to a certain diagnosis or therapy, and cancer diagnosis. Several algorithms have been established for recovering data because it is expensive and time consuming to repeat the experiment. In this research, efficient neural network techniques are used with effective learning algorithms for providing significant cancer classification.

1.2 Data Mining in Bioinformatics
Hand, Mannila and Smyth [31] defined Data mining as "the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner", while Han [32] called it "the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases". There are various other definitions for data mining from various researchers.

Bioinformatics is the application of molecular biology, computer science, artificial intelligence, statistics and mathematics to model, organize, understand and identify interesting knowledge associated with large-scale molecular biology databases. Classification is a major part of biology and thus classification techniques play a vital role in bioinformatics, often using similarities of structure to infer similarity of function. A wide range of such techniques are used, both deterministic and probabilistic.

Data mining in bioinformatics is vulnerable due to many facets of biological databases which includes their size, their number, their diversity and the lack of a standard ontology to aid the querying

of them, as well as the heterogeneous data of the quality and provenance information they contain. Ultimately, the possible financial value of, and the ethical considerations connected with, some biological data means that the data mining of biological databases is not always as easy to perform as is the case in some other areas.

The analysis of DNA microarrays poses a large number of statistical problems, including the normalization of the data. There are dozens of proposed normalization methods in the published literature; as in many other cases where authorities disagree, a sound conservative approach is to try a number of popular normalization methods and compare the conclusions reached.

A basic difference between microarray data analysis and much traditional biomedical research is the dimensionality of the data. Many analysis techniques treat each sample as a single point in a space with thousands of dimensions, then attempt by various techniques to reduce the dimensionality of the data to something humans can visualize.

DNA microarrays are now increasingly used to obtain data concerning gene expression in various organisms. There is a chance that after a period of basic methodological research they will make an important diagnostic tool in biological research and in medicine. It is now understood that this ingenious technology must be supplemented with appropriate statistical, computational and data storage facilities, in order to be useful for researchers. This paper provides a short review of existing statistical methods of microarray data processing and computational and statistical problems in microarrays. This paper is preceded in this order, sketching the problems connected with planning and analysis of microarray experiments.

1.3 Neural Network

Neural networks have been effectively applied across an astonishing range of problem domains like finance, medicine, engineering, geology, physics and biology. Artificial neural networks (ANNs) are non-linear data driven self adaptive technique. ANN is a dominant tool for modeling, particularly when the underlying data relationship is unknown. Correlated patterns between input data sets and corresponding target values can be recognized and studied by ANN. After training, the outcomes of new independent input data can be predicted by ANN. Thus they are very much suited for the modeling of agricultural data which are known to be complex and often non-linear. A key feature of these networks is their adaptive nature. Neural Networks (NNs) have been widely used for variety of applications. Some of them are classification problems, recognizing speech, and predicting the secondary structure of globular proteins. In time-series applications, NNs have been used in predicting stock market performance. These issues are solved through classical statistical methods, such as discriminant analysis, logistic regression, Baye's analysis, multiple regressions etc.

2. LITERATURE SURVEY

This chapter mainly discusses about the existing cancer detection techniques available in the literature. Several domains and concepts are used in the detection of cancer. The main domains used in this detection technique include neural networks, gene ranking, neuro-fuzzy, etc.

2.1 Cancer Classification using Neural Network

Although numerous traditional methods for identification of cancer in clinical practice can be frequently imperfect or ambiguous, molecular level diagnostics with gene expression profiles is capable of recommending the methodology of accurate, objective, and efficient cancer classification. Hong-Hee Won et al., [1] proposed the ensemble of neural network classifiers learned from negatively correlated characteristics to accurately categorize cancer, and methodically estimate the performances of the proposed technique with the use of three benchmark datasets. Experimental observation reveals that the ensemble classifier with negatively correlated characteristics provides the best recognition rate on the three benchmark datasets.

Hu et al., [2] analyzed the performances of cancer cell classification by using supervised and unsupervised learning methods. A single hidden layer feed forward NN with error back-propagation training is implemented for supervised learning, and c-means clustering approaches, fuzzy and nonfuzzy, are used for unsupervised learning. Network configurations with several activation functions, specifically sigmoid, sinusoid and Gaussian, are examined. A collection of characteristics, together with cell size, average intensity, texture, shape factor and pgDNA are preferred as the input for the network. These characteristics, specifically the texture data, are revealed to be extremely efficient in capturing the discriminate information in cancer cells. It is established that based on the information from 467 cell images from six cases, the neural network technique realizes a classification rate of 96.9% whereas fuzzy c-means scores 76.5%.

Bevilacqua et al., [3] presented a new technique to artificial neural network (ANN) topology optimization that makes use of the multi-objective genetic algorithm with the intention of discovering the most excellent network configuration for the Wisconsin breast cancer database (WBCD) classification problem. The WBCD is an openly available database contains 699 cases, each of which is characterized by 11 constraints. The initial 10 values of every record refer to geometrical characteristics of cells collected with FNA biopsy. The last constraint refers the characteristics of the tumor; two classes of tumor are taken into consideration in this database: benignant and malignant tumors. An intellectual scheme, IDEST, was intended and executed. At the middle of this system there's an Artificial Neural Network that is capable of categorizing cases. The design of such an ANN is a non trivial task and alternatives incoherent with the difficulty possibly will lead to unsteadiness in the network. Because of these reasons a permanent topology genetic algorithm (GA) was used to discover an optimal topology for the specified problem. In a subsequent step a multi-objective GA (MOGA) was developed and employed with the intention of refining the search in the "topology space".

A Brain Cancer Detection and Classification System has been intended and adopted by Joshi et al., [4]. The system makes use of computer based process to identify tumor blocks or lesions and categorize the kind of tumor with the use of Artificial Neural Network in MRI images of several patients with Astrocytoma kind of brain tumors. The image processing approaches like histogram equalization, image segmentation, image enrichment, morphological functions and feature extraction have been built for recognition of the brain tumor in the MRI images of the

cancer infected patients. The extraction of texture elements in the identified tumor has been accomplished by using Gray Level Co-occurrence Matrix (GLCM). These features are evaluated with the stored features in the Knowledge Base. At last a Neuro Fuzzy Classifier has been proposed to identify different kinds of brain cancers. The complete system has been examined in two stages, initially Learning/Training stage and subsequently Recognition/Testing stage. The recognized MRI images of affected brain cancer patients acquired from Radiology Department of Tata Memorial Hospital (TMH) were utilized to train the system. The unidentified samples of brain cancer affected MRI images are also collected from TMH and were utilized to test the system. The system was recognized to be very effective in classification of these samples and responds any irregularity.

Mammography is the modality of alternative for the premature recognition of breast cancer, mainly because of its sensitivity to the recognition of breast cancer. On the other hand, because of its elevated rate of false positive predictions, a huge number of biopsies of benign lesions result. Land et al., [5] investigates the use and compares the performance of two neural network hybrids as a support to radiologists in circumventing biopsies of these benign lesions. These hybrids present the potential to develop both the sensitivity and specificity of breast cancer diagnosis. The first hybrid, the Generalized Regression Neural Network (GRNN) Oracle, concentrates on enhancing the performance output of a set of learning approaches that function and are precise over the complete (defined) learning space. The second hybrid, an evolutionary programming (EP)/adaptive boosting (AB) dependent hybrid, sharply integrates the outputs from an iteratively called "weak" learning approach (one which executes at least to some extent enhanced than random guessing), with the intention of "boosting" the performance of the weak learner. The second part of this approach tells about modifications to enhance the EP/AB hybrid's performance, and additionally estimates how the use of the EP/AB hybrid may prevent biopsies of benign lesions (as evaluated to an EP only classification system), specified the necessity of missing few if any cancers.

In the modern world, in which mechanized identification is increasing its horizons in the area of medicine, breast cancer categorization is getting extensive concentration. In this approach, artificial neural networks have attained considerable recognition rates. On the other hand, to enhance the performance, a method is required to monitor the features of the input data, to obtain the significant ones and suppress those that are inappropriate. Even though neural networks have this ability, here Kermani et al., [6] revealed that by using a hybrid genetic algorithm and neural network (GANN), the feature extraction can be done more efficiently. An additional benefit of augmenting neural network training with a genetic algorithm is that the extracted features using genetic algorithm are clear and perceivable. Even though the authors estimated this approach by using the breast cancer data, the method is intended to handle any other kind of classification task.

Perfect classification of cancers based on microarray gene expressions is extremely essential for doctors to prefer a suitable treatment. Feng Chu et al., [7] used a novel radial basis function (RBF) neural network that permits for large overlaps between the hidden kernels of the similar class to this trouble. The author examined the developed RBF network in three standard data sets, i.e., the lymphoma data set, the small round blue cell tumors (SRBCT) data set, and the ovarian cancer data set. The experimental results in all the three standard data sets confirm that this RBF network is capable of realizing 100% accuracy with much smaller amount of genes than the existing approaches.

Cancer classification based on microarray gene expressions is a significant issue. Feng Chu et al., [8] used a t- test-based feature collection approach to select some essential genes from collection of genes. Subsequently, the authors categorize the microarray data sets by using the fuzzy neural network (FNN). This FNN integrates significant features of primary fuzzy model self-generation, parameter optimization, and rule-base generalization. FNN is used in three recognized gene expression data sets, i.e., the lymphoma data set (which contains 3 sub-types), small round blue cell tumor (SRBCT) data set (which contains 4 sub-types), and the liver cancer data set (which contains 2 classes, i.e., non-tumor and hepatocellular carcinoma (HCC)). The results in all the three data sets prove that the FNN can achieve 100% accuracy with a much lesser amount of genes. By considering the lesser amount of genes needed by the FNN and its elevated accuracy, it is to be concluded that the FNN classifier not only assists biological researchers to distinguish cancers that are complicated to be classified using conventional clinical techniques, but also assists biological researchers to concentrate on a lesser number of significant genes to discover the relationships among those significant genes and the development of cancers.

It is very vital for cancer diagnosis and treatment to perfectly recognize the site of foundation of a tumor. With the materialization and huge progression of DNA microarray approaches, constructing gene expression profiles for several cancer types has previously turn out to be a promising means for cancer classification. As well as investigation on binary classification, for instance, normal against tumor samples, which draws several efforts from a mixture of disciplines, the discrimination of multiple tumor types is also significant. In the meantime, the choice of genes which are appropriate to a certain cancer not only enhances the performance of the classifiers, but also offers molecular insights for treatment and drug production. Here, Rui Xu et al., [9] used semisupervised ellipsoid ARTMAP (SSEAM) for multiclass cancer discrimination and particle swarm optimization for revealing gene selection. SSEAM is a neural network structural design rooted in adaptive resonance theory and appropriate for classification tasks. SSEAM involves quick, steady, and limited learning and generates hyper ellipsoidal clusters, encouraging complex nonlinear judgment boundaries. PSO is an evolutionary algorithm-based method for overall optimization. A separate binary version of PSO is utilized to specify whether genes are selected or not. The efficiency of SSEAM/PSO for multiclass cancer diagnosis is confirmed by examining it on three openly available multiple-class cancer data sets. SSEAM/PSO attains aggressive performance on all these data sets, with outcome equivalent to or enhanced than those obtained by other classifiers.

Kocur et al., [10] concentrated on enhancing micro calcification classification by creating an effective computer-aided diagnosis system that obtains Daubechies-4 and biorthogonal wavelet characteristics. These wavelets were selected since they have been utilized in military target detection and fingerprint detection research with images illustrated by low contrast, comparable to

mammography. Feature selection methods are utilized to additionally develop the classification performance. The artificial neural network feature selection methods are complemented by a traditional decision boundary-based feature selection technique. The results obtained using the wavelet features are evaluated to more traditional measures of image texture, angular second moment, and Karhunen Loeve coefficients. The utilization of different signal processing to evaluate wavelet and neural approaches permits for a measure of the difficulty. It is concluded that development and involvements have been made with the introduction of two novel feature extraction techniques for breast cancer diagnosis, wavelets and eigenmasses. In addition, feature selection approaches are illustrated, evaluated, and validate, transforming sufficient discrimination power into promising classification results.

Schnorrenberg et al., [11] discusses the examination of nuclei in histopathological sections with a system that intimately simulates human experts. The estimation of immunocytochemically stained histopathological part presents a difficult problem because of many variations that are inherent in the methodology. In this respect, several portion of immunocytochemistry remain in doubt, in spite of the fact that results possibly will carry significant diagnostic, prognostic, and therapeutic information. In this approach, a modular neural network-based technique to the recognition and classification of breast cancer nuclei stained for steroid receptors in histopathological portions is described and estimated.

Development in molecular classification of tumors might play a major role in cancer treatment. A new method to genome expression pattern understanding is discussed and implemented by Azuaje et al., [12] to the identification of B-cell malignancies as a test set. With the use of cDNA microarrays data created by a earlier investigations, a neural network representation known as simplified fuzzy ARTMAP is capable of recognizing normal and diffuse large B-cell lymphoma (DLBCL) patients, Additionally, it finds the difference between patients with molecularly distinct forms of DLBCL not including preceding information of those subtypes.

A computer-based classification for diagnosing bladder cancer is investigated by Moallemi et al., [13]. In general, an object will come under any one of the two classes: Well or Not-well. The Well class comprises of the cells that will essentially be helpful for diagnosing bladder cancer; the Not-well class comprises of remaining cells. A number of descriptive features are obtained from all object in the image and then feed to a multilayer perceptron, which categorizes the cells as Well or Not-well. The perceptron's better classification capabilities decrease the number of computer misclassification errors to a level acceptable for clinical use. In addition, the perceptron's parallelism and additional aspects of this implementation lend it to very fast computation, thus offering precise classification at a satisfactory speed.

A characteristic microarray gene expression dataset is typically both very sparse and excessive. To choose multiple extremely informative gene subsets for cancer classification and diagnosis, a novel fuzzy granular support vector machine-recursive feature elimination algorithm (FGSVM-RFE) is intended by Yuchun Tang et al., [14]. Since a hybrid algorithm of statistical learning, fuzzy clustering, and granular computing, the FGSVM-RFE individually eradicates unrelated, unnecessary, or noisy genes in several granules at all stages and chooses extremely informative genes with potentially different biological functions in balance. Experimental studies on three open datasets reveal that the FGSVM-RFE outperforms all other techniques. Furthermore, the FGSVM-RFE can obtain multiple gene subsets on each of which a classifier can be modeled with 100% accurateness. The independent testing accurateness particularly for the prostate cancer dataset is considerably developed. The greatest result of the previous approach is 86% with 16 genes and the result of this approach is 100% with merely eight genes. The recognized genes are annotated by Onto-Express to be biologically significant.

Kai-Bo Duan et al., [15] proposes a novel feature selection technique that uses a backward removal process comparable to that used in support vector machine recursive feature elimination (SVM-RFE). Dissimilar to the SVM-RFE technique, at every step, the proposed technique calculates the feature ranking score from an arithmetic analysis of weight vectors of multiple linear SVMs trained on subsamples of the original training data. The proposed method is examined on four gene expression datasets for cancer classification. The experimental observation reveals that this feature selection technique chooses improved gene subsets than the original SVM-RFE and enhances the classification accuracy. A Gene Ontology-based similarity evaluation specifies that the chosen subsets are functionally different, additionally validating the gene selection technique. This examination also recommends that, for gene expression-based cancer classification, standard test error from multiple partitions of training and test sets can be suggested as a reference of performance quality.

Peng Qiu et al., [16] analyzes signal processing and modeling of genomic and proteomic data from two cutting edge methods, specifically microarray technique and mass spectrometry (MS) technique, since they are clear among the chief frontiers that can reshape cancer investigation. Initially, a review of a small number of major design techniques for cancer classification and prediction with the use of genomic proteomic data is done. Then developed an ensemble dependence model (EDM)-based structure and analyzed the conception of dependence network. The EDM network is implemented to both microarray gene expression and MS data sets in cancer investigation. The author also presented the performance-based design and dependence network-based scheme for biomarker recognition. The main objective is to present an extensive review of the current advances on model-based genomic and proteomic signal processing for cancer recognition and prediction.

The most significant application of microarray in gene expression examination is to categorize the unidentified tissue samples based on their gene expression levels with the assistance of recognized sample expression levels. Ghorai et al., [17] recommended a nonparallel plane proximal classifier (NPPC) ensemble that guarantees better classification accurateness of test samples in a computer-aided diagnosis (CAD) structure than that of a single NPPC model. For all data set only a small number of genes are chosen by using a mutual information condition. After that a genetic algorithm-based simultaneous characteristic and model selection system is used to train a number of NPPC expert models in multiple subspaces by exploiting cross-validation accuracy. The members of the ensemble are chosen by the performance of the trained models on a validation set. In addition the standard majority voting technique, the author has introduced

smallest average proximity-based decision combiner for NPPC ensemble. The efficiency of the NPPC ensemble and the proposed novel approach of integrating decisions for cancer diagnosis are examined and evaluated with support vector machine (SVM) classifier in an identical structure. Experimental observations on cancer data sets confirm that the NPPC ensemble provides similar testing accuracy to that of SVM ensemble with decreased training time on average.

Elevated dimensionality has been a most important setback for gene array-based cancer classification. It is very vital to recognize marker genes for cancer diagnoses. Jiexun Li et al., [18] proposed a structure of gene selection techniques based on previous investigations. This approach concentrates on most favorable search-based subset selection techniques since they estimate the group performance of genes and assist to identify global optimal set of marker genes. Particularly, this method is the first to initiate tabu search (TS) to gene selection from huge-dimensional gene array data. The comparative evaluation of gene selection techniques established the effectiveness of optimal search-based gene subset selection to recognize cancer marker genes. TS were revealed to be very capable tool for gene subset selection.

The goal of Toure et al., [19] work is to investigate the use of gene expression data in discriminating two categories of extremely comparable cancers-acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Classification outcome are reported in previous approaches using methods apart from neural networks. Here, this approach explored the responsibility of the feature vector in classification. Every feature vector includes 6817 elements that are gene expression data for 6817 genes. The authors demonstrated in this experiment that learning with the use of neural networks is promising when the input vector has the correct number of gene expression data. This result is very promising because of the nature of the data (available in large amounts and more new information becomes available with better technology and better understanding of the problem). Therefore, it is completely necessary to utilize an automated recognition system that has learning capability.

Sehgal et al., [20] proposed decision-based fusion models to categorize BRCA1, BRCA2 and Sporadic genetic mutations for breast and ovarian cancer. Several ensembles of base classifiers makes use of the stacked generalization approach have been proposed which contains support vector machines (SVM) with linear, polynomial and radial base function kernels. A generalized regression neural network (GRNN) is then implemented to forecast the mutation type in accordance with the outputs of base classifiers, and experimental output confirm that the novel proposed fusion methodology for choosing the most excellent and eradicating weak classifiers outperforms single classification models.

2.2 Gene Ranking

Hero [21] put forth a gene selection and ranking with microarray data. Over many years, an explosion in the quantity of genomic data available to biomedical researchers due to advances in biotechnology. For instance, by utilizing gene microarrays, it is now very easy to investigate a person's gene expression profile more than 30,000 genes of the person genome. Signals obtained from gene microarray experimentations can be associated to genetic features underlying disease, improvement, and aging in a population. This has significantly speeded up the gene detection. However, the enormous scale and investigational variability of genomic data makes removal of biologically important genetic information is very challenging. One of the biggest disputes is to recognize the affected genes that are participated in that specific disease based on a gene microarray research. The authors illustrated multi criterion approaches that are proposed for this gene selection and ranking difficulty.

Chen Liao et al., [22] presented a gene selection for cancer classification using Wilcoxon rank sum test and support vector machine. Gene selection is an important difficulty in microarray data processing. A novel gene selection approach derived from Wilcoxon rank sum test and support vector machine (SVM) is developed in this paper. Wilcoxon rank sum test is utilized to choose a subset is done first. Next step is to train and test the selected gene with the use of SVM classifier with linear kernel independently and genes with elevated testing accuracy rates are selected to form the last reduced gene subset. Leave-one-out cross validation (LOOCV) categorization outputs on two datasets namely ALL/AML leukemia and breast cancer to exhibit the implemented method can obtain good result with final reduced subset. The resulted genes are listed and their expression levels are outlined to illustrate that the selected genes can produce clear separation between two classes.

A statistical method for ranking differentially expressed genes was recommended by Broberg [23]. Current methods have evaluated gene selection approaches by utilizing ROC curves calculated by simulation. But, no effort has been made to evaluate selection accuracy as a function of population parameters. In particular, simulation investigations are limited unavoidably when the multiplicity is considered, as only a small subset of conditions can possibly be explored. The authors summarized a technique for predicting an optimal test statistic with which to rank genes with respect to differential expression. A test of this approach demonstrates that it permits generation of top gene lists that give a small number of false positives and a small amount of false negatives. Evaluation of the false-negative as well as the false-positive rate illustrates the main focus of this approach.

Identifying significant genes from microarray data was presented by Han-Yu Chuang et al., [24]. Microarray method is a latest improvement in investigational molecular biology which can produce quantitative expression magnitudes for large number of genes in a single, cellular mRNA sample. All these gene expression magnitudes outline a collective profile of the sample, which can be utilized to distinguish samples from dissimilar classes such as tissue types or treatments. In the gene expression profile data gathered in a specific evaluation, most likely only few genes will be differentially expressed among the classes, whereas many other genes have similar expression levels. Selecting a set of informative differential genes obtained from these large set of data is significant for microarray data analysis. In this paper, the authors illustrate a framework for selecting informative genes which is called as ranking and combination analysis (RAC). This proposed method provides the combination of many existing better informative gene selection approaches. To evaluate this method conducted many experiments using three data sets and six existing feature selection methods. The observations demonstrate that the RAC framework is a strong and effective method to categorize informative gene for

microarray data. In many cases, the combined approach on two selecting techniques provides a better performance compared to the efficiency of the individual technique. Significantly, the combined approach outperforms each of the individual feature selection approach when considering all three data sets together. All these experimentation result shows that RCA is an effective and useful approach for the microarray gene expression analysis.

Jin-Hyuk Hong Herold et al., [25] put forth a cancer classification with incremental gene selection based on DNA microarray data. Gene selection is an significant issue for cancer classification. Filter and wrapper techniques are extensively used for gene selection, where the filter technique is hard to measure the relationship among the genes and the wrapper technique requires lots of computation. The author proposed an innovative method called gene boosting which is used to select appropriate gene subsets by combining filter and wrapper approaches. This approach continuously selects a set of top-ranked informative genes using a filtering technique regarding to a chronological training dataset constructed according to the classification result for the original training dataset. Experimental conducted on few microarray benchmark datasets results shows that this technique is very efficient in predicting a appropriate gene subset. Aggressive performance was obtained with fewer genes in a reasonable time. This also led to the detection of some genes selected frequently as useful features.

Hengpraprohm et al., [26] presented a method for selecting informative features using K-Means clustering and SNR ranking. The performance of the proposed method was tested on cancer classification problems. Genetic Programming is employed as a classifier. The experimental results indicate that the proposed method yields higher accuracy than using the SNR ranking alone and higher than using all of the genes in classification. The clustering step assures that the selected genes have low redundancy; hence the classifier can exploit these features to obtain better performance.

Gene selection is an important problem in microarray data processing. A new gene selection method based on Wilcoxon rank sum test and support vector machine (SVM) is proposed by Chen Liao et al., [27]. First, Wilcoxon rank sum test is used to select a subset. Then each selected gene is trained and tested using SVM classifier with linear kernel separately, and genes with high testing accuracy rates are chosen to form the final reduced gene subset. Leave-one-out cross validation (LOOCV) classification results on two datasets: breast cancer and ALL/AML leukemia, demonstrate the proposed method can get 100% success rate with final reduced subset. The selected genes are listed and their expression levels are sketched, which show that the selected genes can make clear separation between two classes.

DNA microarray, which is one of the most important molecular biology technologies in post-genomic era, has been widely applied in medical field, especially for cancer classification. However, it is difficult to acquire excellent classification accuracy by using traditional classification approaches due to microarray datasets are extremely asymmetric in dimensionality. In recent years, ensemble classifiers which may obtain better classification accuracy and robustness have attracted more interests in this field but it is more time-consuming. Therefore, Yu Hualong et al., [28] proposed a novel ensemble classification method named as SREC (Simple Rule-based Ensemble Classifiers). Firstly, the classification contribution of each gene is evaluated by a novel strategy and the corresponding classification rule is extracted. Then we rank all genes to select some important ones. At last, the rules of the selected genes are assembled by weighted-voting to make decision for testing samples. It has been demonstrated the proposed method may improve classification accuracy with lower time-complexity than traditional classification methods.

Lipo Wang et al., [29] aimed at finding the smallest set of genes that can ensure highly accurate classification of cancers from microarray data by using supervised machine learning algorithms. Choose some important genes using a feature importance ranking scheme. In the second step, we test the classification capability of all simple combinations of those important genes by using a good classifier. For three "small" and "simple" data sets with two, three, and four cancer (sub) types, our approach obtained very high accuracy with only two or three genes. For a "large" and "complex" data set with 14 cancer types, we divided the whole problem into a group of binary classification problems and applied the 2-step approach to each of these binary classification problems. Through this "divide-and-conquer" approach, we obtained accuracy comparable to previously reported results but with only 28 genes rather than 16,063 genes. In general, our method can significantly reduce the number of genes required for highly reliable diagnosis.

Data mining algorithms are commonly used for cancer classification. Prediction models were widely used to classify cancer cells in human body. Saravanan et al., [30] focuses on finding small number of genes that can best predict the type of cancer. From the samples taken from several groups of individuals with known classes, the group to which a new individual belongs to is determined accurately. The paper uses a classical statistical technique for gene ranking and SVM classifier for gene selection and classification. The methodology was applied on two publicly available cancer databases. SVM one-against- all and one-against-one method were used with two different kernel functions and their performances are compared and promising results were achieved.

3. METHODOLOGY

The present research work illustrates that SVMs are very efficient in classifying and recognizing informative features or attributes (such as critically significant genes). Statistical ranking technique is used in the present research work for the better classification results.

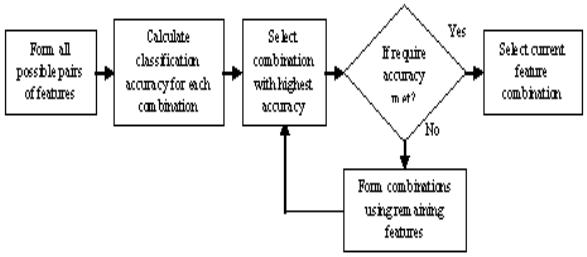

Figure 3.1: Proposed Feature Selection Method

The methodologies used in this thesis for cancer classification is presented in this chapter. The process is divided in tow phases. In the first phase, gene ranking is carried out with the help of ranking technique such as Enrichment Score, ANalysis Of Variance (ANOVA) and Correlation. In the second phase, classification is performed using Support Vector Machine (SVM) along with the class separability for classifying the cancer tumor exactly by means of microarray gene expression which be performed better than the usage of Fuzzy Neural Network (FNN) technique.

Initially, all the features are ranked using a feature ranking measure which are described later and the most important features alone are retained for next the step. After selecting some top features from the importance ranking list, the data set is attempted to classify with only one feature. In this proposed approach, the Support Vector Machine (SVM) classifier is used to test n-feature combinations.

3.1 Ranking Techniques

3.1.1 Enrichment Score

Enrichment Score is an annotation methodology that takes as inputs:

1. Genome-wide expression profiles consisting of p genes and n samples with each sample corresponding to one of two classes, $C_1$and C_2, the expression of the j-th gene in the i-th sample is x_j^i;

2. A database of m gene sets $\Gamma=\gamma_1, \gamma_2,\ldots\gamma_m$ where each gene set k is a list of genes (a subset of the p genes in the data set) belonging to a pathway or other functional or structural category;

3. A ranking procedure and correlation statistic that takes the expression data set and labels as inputs and produces correlation statistics for each sample that reflects the correlation of the p genes in that sample with respect to the distribution of expression in the two classes. The correlation statistics for the i-th sample would be $c_i = \{c_1^i \ldots c_p^i\}$ and produces as outputs:

An enrichment score for each sample in the data set with respect to each gene set in the database ES_i^k corresponds to the enrichment of i-th sample with respect to the k-th gene set;

A measure of confidence for enrichment score is given by a p-value with multiplicity taken into account by Family-wise error rate (FWER) p-values and a False Discovery rate (FDR) q-values.

Given the correlation statistics for the i-th sample

$$c_i = \{c_1^i \ldots c_p^i\}$$

And a gene setγ_k , we construct the following discrete random walk over the indices of the rank-ordered correlation statistic

$$v(\ell) = \frac{\sum_{j=1}^{\ell} |\mathcal{C}_{(j)}|^r I(g_{(j)} \epsilon \gamma_\kappa)}{\sum_{j=1}^{p} |\mathcal{C}_{(j)}|^r I(g_{(j)} \epsilon \gamma_\kappa)} - \frac{\sum_{j=1}^{\ell} I(g_{(j)} \notin \gamma_\kappa)}{p - |\gamma_\kappa|}$$

where $c_{(j)}$ is the rank-ordered correlation statistic, r is a parameter (in general r = 1), γ_k is the k-th gene set, $I(g_{(j)} \in \gamma_k)$ is the indicator function on whether the j-th gene (the gene corresponding to the j-th ranked correlation statistic) is in gene set γ_k , $|\gamma_k|$ is the number of genes in the k-th gene set, and p is the number of genes in the data set. The enrichment statistic for i-th sample with respect to the k-th gene set is the maximum deviation of the random walk from zero

$$ES_i^k = v[\arg\max_{\ell=1,\cdot,p} v(\ell)][$$

3.1.2 Correlation

A commonly used measure of correlation is provided by Pearson's product Moment Correlation Coefficient (PMCC). This is denoted by r and calculated from sample data using the formula

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}}$$

Where

$$S_{xx} = \sum (x_i - \bar{x})^2, S_{yy} = \sum (y_i - \bar{y})^2$$

$$\text{and } S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

The PMCC also provides a test statistic for the null hypothesis that there is no correlation between the two variables in the bivariate parent population from which the sample data were drawn. For the PMCC test, both the variables must be random. The notation for the population correlation coefficient is ρ, the equivalent Greek letter to r which is used for the test statistic.

It is usual to state the null hypothesis as

"H0: There is no correlation, ρ = 0".

Since ρ is a parameter of the (bivariate) population, the inclusion of the statement "ρ = 0"emphasis the point that, as is standard procedure for hypothesis testing, the test is being carried out on the parent population. It is good practice to emphasis this point by including the word "population" in the statement, making it

"H0: ρ =0, where ρ is the population correlation coefficient".

After the genes are ranked based on the ranking technique described above, classification is carried out to detect the occurrence of cancer. Support Vector Machine is the classifier used in this thesis which is described as below.

3.2 Classification using Support Vector Machines

Support Vector Machine (SVM) is a type of learning technique depending on the statistical learning theory, which employs the structural risk minimization inductive principle with the intention to acquire a good generalization from narrow data sets.

SVM produce black box models in the sense that they do not have the capability to explain, in a clear form, the process by means of which the exit takes place. To overcome this drawback, the theory created by either neural network or SVM could be transferred into a more logical representation; these conversion methods are known as rule extraction algorithms.

SVM seeks global hyper plane to separate both classes of examples in training set and avoid over fitting. SVM is a significant machine learning technique which is based on artificial intelligence. The exact working principle of SVM can be clearly understood from the Figure 3.2.

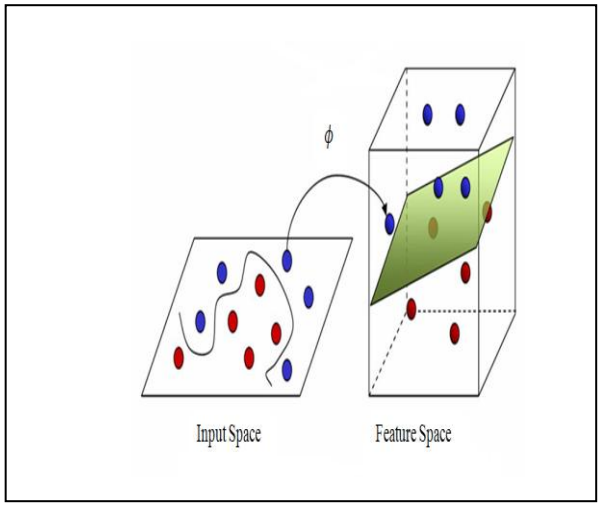

Figure 3.2: Principle of SVM

The mapping of the input-output functions from a group of labeled training data set is constructed by the supervised learning technique called SVM. SVMs in a high dimensional feature space, exploit a hypothesis space of linear functions which are trained with a learning algorithm from optimization theory that employs a learning bias derived from statistical learning theory.

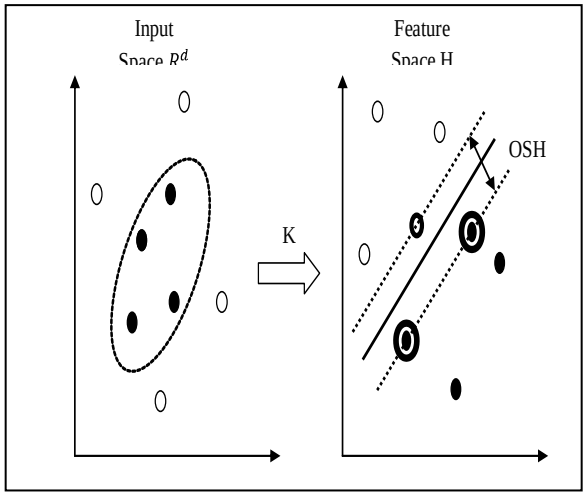

Figure 3.3: A separating hyper plane in the feature space corresponding to a non-linear boundary in the input space.

The above Figure 3.3 contains a separating hyper plane in the feature space corresponding to a non-linear boundary in the input space.

SVM is a new technique for training classifiers depending on various functions such as polynomial functions, radial basis functions, neural networks etc. In Support Vector machines, the classifier is created using a hyper-linear separating plane. SVM offers a most excellent solution for issues that cannot be linearly separated in the input space. The problem is resolved by making a non-linear transformation of the original input space into a high dimensional feature space, where an optimal separating hyper plane is found. A maximal margin classifier with respect to the training data is obtained when the separating planes are optimal.

The support vectors are the points which are at the margin and the solution depends only on these data points. This is the unique feature of this technique. Linear SVM can be extended to nonlinear SVM if a feature space uses a group of nonlinear basis function. The data points can be separated linearly in the feature space which are very high dimensional.

3.3 SVM algorithm

The fundamental idea of SVM can be described as follows:

Step 1: The inputs are formulated as feature vectors.

Step 2: By using the kernel function, these feature vectors are mapped into a feature space.

Step 3: A division is computed in the feature space to separate the classes of training vectors.

3.4 SVM Classifier

SVM is one of the best linear classification methods. The transformation of the samples space to high-dimension space is possible by the kernel mapping and the best linear classification surface of samples in this new space is obtained. This Non-linear transformation is achieved by suitable inner product function. The best linear classification surface function of characteristics space can be described by the following equation:

$$g(x) = \sum_{j=i}^{n} a_j y_i k(x, x_i) + b$$

Where (x_i, y_i) are the two types of sample collection divided in the sample space, b is the classification threshold, and $k(x, x_i)$ is being the nonlinear kernel function that replace characteristics space and meet Mercer conditions. The best linear classification surface function is obtained by striking the best resolve a_i where i = 1, 2… n of the following function Q (a).

$$g(x)g(x) = \sum_{j=i}^{n} a_j y_i k$$

$$\max_{a} Q(a) = \sum_{i=0}^{n} a_i - 0.5 \sum_{i=0}^{n} \sum_{j=0}^{n} a_i a_j y_i y_j k(x_i, x_j)$$

$$\sum_{j=i}^{n} a_j y_i = 0$$

$i = 1,2 \ldots . n$ And $0 \le a_i$

The above equation is solving of quadratic function extreme value on condition that inequality, Q (a) is convex function. Because its local optimal solution is global optimal solution, the solution is unique. Thus the best classification function of SVM is:

$$f(x) = \operatorname{sgn}(g(x)) = \operatorname{sgn}\left\{\sum_{j=1}^{n} a_j y_j k(x, x_i) + b\right\}$$

$$\operatorname{sgn}\left\{\sum_{j=1}^{n} a_j y_j k(x, x_i) + b\right\}$$

3.5 SVM Kernel Functions

A kernel function and its parameter have to be chosen to build a SVM classifier. Three main kernel functions have been used to build SVM classifiers. They are

Linear Kernel function, $K(x, z) = \langle x, z \rangle$

Polynomial kernel function, $K(x, z) = (\langle x, z \rangle + 1)^d$, d is the degree of polynomial.

Radial basis function, $K(x, z) = \exp\{-|x - z|^2 / 2\sigma^2\}$, σ is the width of the function.

4. EXPERIMENTAL RESULTS

This chapter provides the evaluation result for the proposed approach. The dataset used for evaluation is Liver Cancer Dataset. Initially, Correlation technique is used for ranking the important genes. Next, the classifier called Support Vector Machine (SVM) is used for classifying the occurrence of cancer.

An evaluation study of the proposed approach is presented in this chapter. The results of an extensive set of simulation tests are shown, in which the weather prediction approaches are compared under a wide variety of different scenarios.

MATLAB (MATrix LABoratory) is used for the computation of the numerical analysis and is considered as a fourth-generation programming language. MATLAB is a commercial Matrix Laboratory package which functions as an interactive programming environment. It is a foundation of the Mathematics Department software lineup and is also available for PC's and Macintoshes and may be found on the CIRCA VAXes. MATLAB allows matrix operations, plotting of functions and data, execution of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C^{++}, and FORTRAN.

The utilization of MATLAB to numerical experiments is very well adapted as the fundamental algorithms for MATLAB's built in functions and supplied m-files are based on the standard libraries LINPACK and EISPACK.

The main use of MATLAB is for numerical computation but an optional toolbox uses the MuPAD symbolic engine, allowing access to symbolic computing capabilities. Simulink is an additional package, adds graphical multi-domain simulation and Model-Based Design for dynamic and embedded systems.

The application of MATLAB is built around the MATLAB language. The easy way to execute MATLAB code is to type it in at the prompt, >>, in the Command Window, one of the elements of the MATLAB Desktop. Thus, MATLAB is considered as an interactive mathematical shell. Sequences of commands are saved in a text file, normally using the MATLAB Editor, as a script or encapsulated into a function, extending the commands available.

Thus MATLAB is an interactive tool for doing numerical computations with matrices and vectors. It can also present information graphically.

Hence for this research work MATLAB has been taken into consideration and all the three techniques have been implemented using this MATLAB. The platform used for these proposed approaches is Windows XP. The processor used is Pentium IV. The experimentation needs a system RAM of 2 GB.

4.1 Experimental Evaluation

The liver cancer data set [33] has two classes, i.e., the nontumor liver and HCC. The data set contains 156 samples and the expression data of 1,648 important genes. Among them, 82 are HCCs and the other 74 are nontumor livers. The data is randomly divided into 78 training samples and 78 testing samples. In this data set, there are some missing values. K-nearest neighbor method is used to fill those missing values.

The performance of the proposed approach is evaluated against the conventional techniques enrichment.

4.2 Enrichment Technique

TABEL 4.1 ACCURACY COMPARISON OF FNN AND SVM USING ENRICHMENT TECHNIQUE

No. of Fold	No of Genes	Gene Combination	FNN		SVM	
			Acc.	*Best Gene*	*Acc.*	*Best Gene*
5	5	2	100	(1,2),(1,3)	44.8	(1,2),(1,3)
5	7	2	100	(1,2),(1,3)	45	(1,2),(1,3)

4.3 Correlation Ranking Approach

TABEL 4.2 ACCURACY COMPARISON OF FNN AND SVM USING CORRELATION RANKING TECHNIQUE

No. of Fold	No of Genes	Gene Combination	FNN		SVM	
			Acc.	*Best Gene*	*Acc.*	*Best Gene*
5	5	2	91.48	(1,2),(1,3)	100	(1,2),(1,3)
5	7	2	95.45	(1,2),(1,3)	100	(1,2),(1,3)

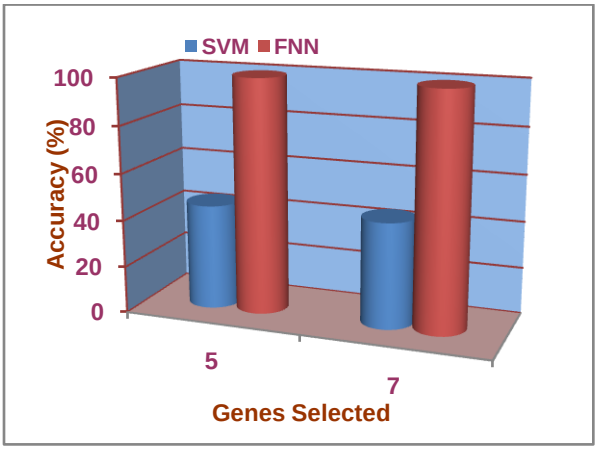

Figure 4.1: Accuracy Comparison of the FNN and SVM using Enrichment Technique

Figure 4.2 represents the execution time for classification by the use of enrichment technique.

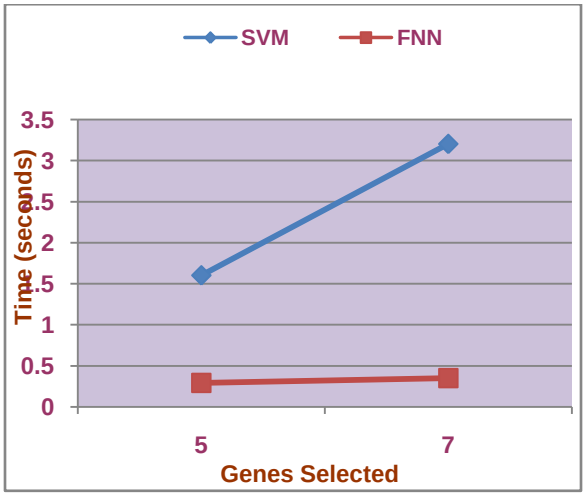

Figure 4.2: Learning Time for FNN and SVM using Enrichment Technique

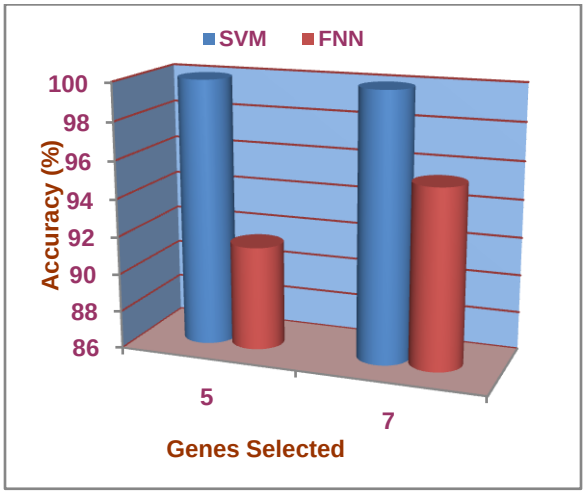

Figure 4.3: Accuracy Comparison of the FNN and SVM using Correlation Ranking

Empirical result shows the accuracy of the new approach is outstanding when evaluate against the Enrichment Score.

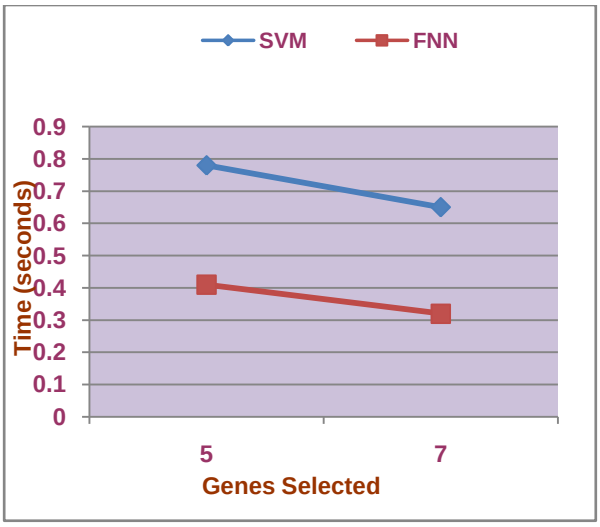

Figure 4.4: Learning Time for FNN and SVM using Correlation Ranking

4.4 MAPSTD

TABEL 4.3 ACCURACY COMPARISON OF FNN AND SVM USING MAPSTD

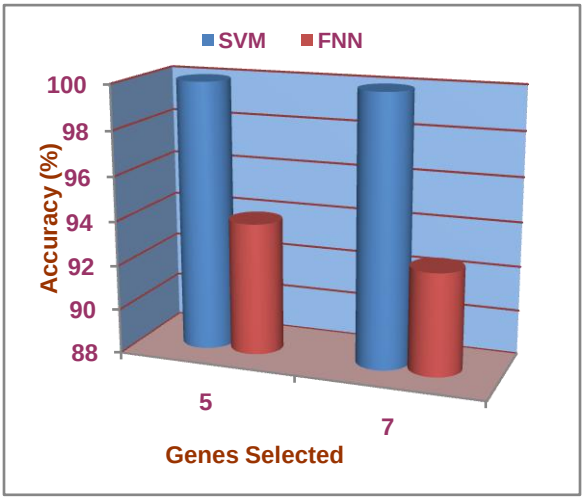

Figure 4.5: Accuracy Comparison of the FNN and SVM using MAPSTD

Hence, it is observed from the experimental results, the intended cancer classification approach that uses FNN and SVM with correlation ranking and MAPSTD provides utmost accuracy when compared among the other conventional approaches.

Figure 4.5 and 4.6 illustrate the accuracy and learning time while using MAPSTD.

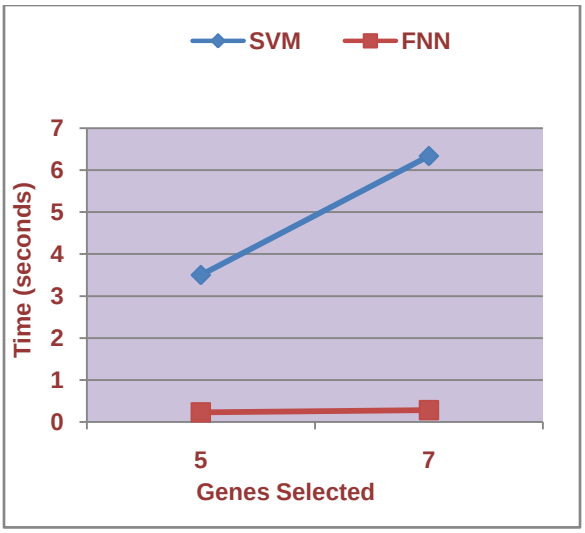

Figure 4.6: Learning Time for FNN and SVM using MAPSTD

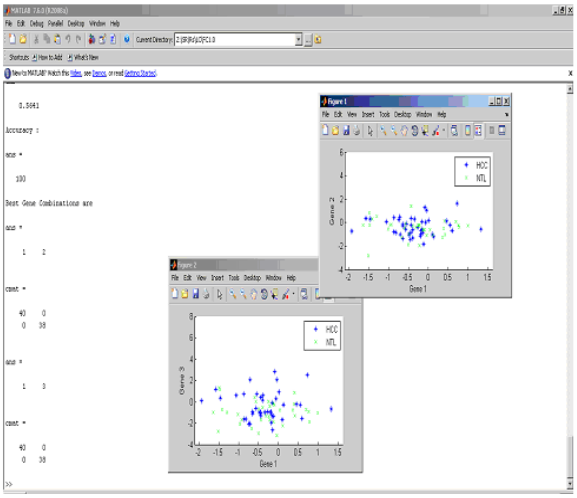

Figure 4.7: Accuracy of the proposed work

5. CONCLUSION

This research concentrates on the cancer classification by means of microarray data. It is obvious that DNA microarrays can be used to aid the severance of tumors with allied morphological appearance, guess patient result independently of predictable prognostic factors and select for response or resistance to specific anti-cancer therapies. As well as the DNA microarray technology will definitely result in better diagnosis and management of certain groups of patients with specific cancers. In this thesis, Fuzzy Neural Network is used for classifying the cancer cells. Correlation ranking technique and MAPSTD is used which provides significant ranking result of the cancer cells.

Support Vector Machine (SVM) is used better learning capability. As support vector machines are linear classifier that has the capability of finding the optimal hyper plane that increases the separation among patterns, this characteristic creates support vector machines as a potential means for gene expression data examination purposes. Thus the usage of SVM reduces the human involvement in choosing input weight.

This technique is guaranteed to revolutionize the understanding of cancer pattern and progression, identify new targets for treatment and provide a new generation of tumor markers for helping with cancer diagnosis and management. The usage of SVM helps in effectiveness of learning procedure. The proposed method is experimented with the help of liver cancer data set .The experimental result produces lesser mean test error when compared to the conventional methods. This clearly shows that the proposed technique classifies the cancer with better accuracy than the usage of fuzzy neural network for classification.

6. SCOPE FOR FUTURE ENHANCEMENT

As a result of the success of the research, more areas of investigation can be pursed. This could be in terms of improving the when large numbers of similar quantities are being estimated that is highest and lowest effects tend to be too extreme. This research which focuses on the cancer classification based on the FNN and SVM with effective correlation technique. In order to improve the learning capacity and to reduce complexity, this research requires some future enhancement.

In future, better neural network techniques can be incorporated with the present research work for less complexity and better learning capacity.

More over, better Neuro fuzzy techniques could also be used to improve the classification rate and accuracy.

Better machine learning techniques can also be adapted for better learning capability and speed.

7. REFERENCES

[1] Hong-Hee Won; Sung-Bae Cho; "Paired neural network with negatively correlated features for cancer classification in DNA gene expression profiles", Proceedings of the International Joint Conference on Neural Networks, Vol. 3, Pp. 1708 – 1713, 2003.

[2] Hu, Y.; Ashenayi, K.; Veltri, R.; O'Dowd, G.; Miller, G.; Hurst, R.; Bonner, R.; "A comparison of neural network and fuzzy c-means methods in bladder cancer cell classification", IEEE International Conference on Neural Networks, IEEE World Congress on Computational Intelligence, Vol. 6, Pp. 3461 – 3466, 1994.

[3] Bevilacqua, V.; Mastronardi, G.; Menolascina, F.; Pannarale, P.; Pedone, A.; "A Novel Multi-Objective Genetic Algorithm Approach to Artificial Neural Network Topology Optimisation: The Breast Cancer Classification Problem", International Joint Conference on Neural Networks (IJCNN '06), Pp. 1958 – 1965, 2006.

[4] Joshi, D.M.; Rana, N.K.; Misra, V.M.; "Classification of Brain Cancer using Artificial Neural Network", International Conference on Electronic Computer Technology (ICECT), Pp. 112 – 116, 2010.

[5] Land, W.H., Jr.; Masters, T.; Lo, J.Y.; McKee, D.W.; "Application of evolutionary computation and neural network hybrids for breast cancer classification using mammogram and history data", Proceedings of the 2001 Congress on Evolutionary Computation, Vol. 2, Pp. 1147 – 1154, 2001.

[6] Kermani, B.G.; White, M.W.; Nagle, H.T.; "Feature extraction by genetic algorithms for neural networks in breast cancer classification", IEEE 17th Annual Conference Engineering in Medicine and Biology Society, Vol. 1, Pp. 831 – 832, 1995.

[7] Feng Chu; Lipo Wang; "Applying RBF Neural Networks to Cancer Classification Based on Gene Expressions", International Joint Conference on Neural Networks (IJCNN '06), Pp. 1930 – 1934, 2006.

[8] Feng Chu; Wei Xie; Lipo Wang; "Gene selection and cancer classification using a fuzzy neural network", IEEE Annual Meeting of the Fuzzy Information Processing (NAFIPS '04), Vol. 2, Pp. 555 – 559, 2004.

[9] Rui Xu; Anagnostopoulos, G.C.; Wunsch, D.C.I.I.; "Multiclass Cancer Classification Using Semisupervised Ellipsoid ARTMAP and Particle Swarm Optimization with Gene Expression Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 4, No. 1, Pp. 65 – 77, 2007.

[10] Kocur, C.M.; Rogers, S.K.; Myers, L.R.; Burns, T.; Kabrisky, M.; Hoffmeister, J.W.; Bauer, K.W.; Steppe, J.M.; "Using neural networks to select wavelet features for breast cancer diagnosis", IEEE Engineering in Medicine and Biology Magazine, Vol. 15, No. 3, Pp. 95 - 102, 1996.

[11] Schnorrenberg, F.; Tsapatsoulis, N.; Pattichis, C.S.; Schizus, C.N.; Kollias, S.; Vassiliou, M.; Adamou, A.; Kyriacou, K.; "Improved detection of breast cancer nuclei using modular neural networks", IEEE Engineering in Medicine and Biology Magazine, Vol. 19, No. 1, Pp. 48 – 63, 2000.

[12] Azuaje, F.; "A computational neural approach to support the discovery of gene function and classes of cancer", IEEE Transactions on Biomedical Engineering, Vol. 48, No. 3, Pp. 332 – 339, 2001.

[13] Moallemi, C.; "Classifying cells for cancer diagnosis using neural networks", IEEE Expert, Vol. 6, No. 6, Pp. 8, 10 – 12, 1991.

[14] Yuchun Tang; Yan-Qing Zhang; Zhen Huang; Xiaohua Hu; Yichuan Zhao; "Recursive Fuzzy Granulation for Gene Subsets Extraction and Cancer Classification", IEEE Transactions on Information Technology in Biomedicine, Vol. 2, No. 6, Pp. 723 – 730, 2008.

[15] Kai-Bo Duan; Rajapakse, J.C.; Haiying Wang; Azuaje, F.; "Multiple SVM-RFE for gene selection in cancer classification with expression data", IEEE Transactions on NanoBioscience, Vol. 4, No. 3, Pp. 228 – 234, 2005.

[16] Peng Qiu; Wang, Z.J.; Liu, K.J.R.; "Genomic processing for cancer classification and prediction - Abroad review of the recent advances in model-based genomoric and proteomic signal processing for cancer detection", IEEE Signal Processing Magazine, Vol. 24, Vo. 1, Pp. 100 – 110, 2007.

[17] Ghorai, S.; Mukherjee, A.; Sengupta, S.; Dutta, P.K.; "Cancer Classification from Gene Expression Data by NPPC Ensemble", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 8, No. 3, Pp. 659 – 671, 2011.

[18] Jiexun Li; Hua Su; Hsinchun Chen; Futscher, B.W.; "Optimal Search-Based Gene Subset Selection for Gene Array Cancer Classification", IEEE Transactions on Information Technology in Biomedicine, Vol. 11, No. 4, Pp. 398 – 405, 2007.

[19] Toure, A.; Basu, M.; "Application of neural network to gene expression data for cancer classification", Proceedings International Joint Conference on Neural Networks (IJCNN '01), Vol. 1, Pp. 583 – 587, 2001.

[20] Sehgal, M.S.B.; Gondal, I.; Dooley, L.; "Support vector machine and generalized regression neural network based classification fusion models for cancer diagnosis", Fourth International Conference on Hybrid Intelligent Systems (HIS '04), Pp. 49 – 54, 2004.

[21] A.O. Hero, "Gene selection and ranking with microarray data," Seventh International Symposium on Signal Processing and its Applications, Vol. 1, pp. 457 – 464, 2003.

[22] Chen Liao, Shutao Li, Zhiyuan Luo, "Gene Selection for Cancer Classification using Wilcoxon Rank Sum Test and Support Vector Machine," International Conference on Computational Intelligence and Security, Vol. 1, pp. 368 – 373, 2006.

[23] P. Broberg, "Statistical methods for ranking differentially expressed genes," Genome Biology, Vol.4, No. 6, 2003.

[24] Han-Yu Chuang, Hong fang Liu Brown, S. McMunn-Coffran, C. Cheng-Yan Kao, D. F. Hsu, "Identifying significant genes from microarray data," Fourth IEEE Symposium on Bioinformatics and Bioengineering, pp. 358 – 365, 2004.

[25] Jin-Hyuk Hong, Sung-Bae Cho, "Cancer classification with incremental gene selection based on DNA microarray data," CIBCB '08, IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, pp. 70 – 74, 2008.

[26] Hengpraprohm, S.; Chongstitvatana, P.; "Selecting Informative Genes from Microarray Data for Cancer Classification with Genetic Programming Classifier Using K-Means Clustering and SNR Ranking", Frontiers in the Convergence of Bioscience and Information Technologies (FBIT), Pp. 211 – 218, 2007.

[27] Chen Liao; Shutao Li; Zhiyuan Luo; "Gene Selection for Cancer Classification using Wilcoxon Rank Sum Test and Support Vector Machine", International Conference on Computational Intelligence and Security, Vol. 1, Pp. 368 – 373, 2006.

[28] Yu, Hualong; Sen Xu; "Simple rule-based ensemble classifiers for cancer DNA microarray data classification", International Conference on Computer Science and Service System (CSSS), Pp. 2555 – 2558, 2011.

[29] Lipo Wang; Feng Chu; Wei Xie; "Accurate Cancer Classification Using Expressions of Very Few Genes",

IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 4, No. 1, Pp. 40 – 53, 2007.

[30] Saravanan, V.; Mallika, R.; "An Effective Classification Model for Cancer Diagnosis Using Micro Array Gene Expression Data", International Conference on Computer Engineering and Technology (ICCET '09), Vol. 1, Pp. 137 – 141, 2009.

[31] Hand, David, Heikki Mannila, and Padhraic Smyth, Principles of Data Mining, MIT Press 2001.

[32] Meyer T and Hart IR (1998) "Mechanisms of Tumour Metastasis", European Journal of Cancer 34: 214–221.

[33] "Gene expression patterns in liver liver cancers" http://genome-www.stanford.edu/hcc/