Cancer Prediction with Gene Expression Data

G Sivagamasundari[1], Latha Parthiban [2]

***Abstract*: With continuous growth in technology and quantum of data, many data mining algorithms are developed that uses micro array data to classify the genes and expressions in normal and disease conditions. There are many clustering algorithms that help to classify the genes and there is conflict in large pool of genes and their expression characters. The proposed system takes the input from multiple sources produces associate storage, cluster the information and classify it.**

***Index Terms*: Data mining, Gene expression data, Classification.**

I. INTRODUCTION

Biological research includes sequences of genes and proteins, gene expressions, biological functions, pathways etc. There are several techniques that can generate large data of genes, proteins, and their expressions. Different data mining methods are used to predict the gene expression data, and various algorithms are used to predict the genes and their functions The microarray data of both normal and disease conditions have several conditions of data that use cluster ways and microarray knowledge to classify. The microarray data of both normal and disease conditions have several conditions of data that use cluster ways and microarray knowledge to classify the genes. Based on the literature and different statistical approaches, there are several clustering algorithms that are used to built gene expression data and classification, however little attention has been paid to uncertainty within the results obtained

To get a transparent image on the sight of a angle, a transparent cancer classification associate analysis system has to be pictured followed by a scientific approach to analyse international organic phenomenon that provides an optimized answer for the known drawback space. Molecular medical speciality provides a possibility of human cancer classification, however these tests don't seem to be wide applied as a result of characteristic molecular markers for many solid neoplasm save nevertheless to be known.

Biological research includes sequences of genes and proteins, gene expressions, biological functions, pathways etc. There are several techniques that can generate large data of genes, proteins, and their expressions. Different data mining methods are used to predict the gene expression data, and various algorithms are used to predict the genes and their functions The microarray data of both normal and disease conditions have several conditions of data that use cluster ways and microarray knowledge to classify. The microarray data of both normal and disease conditions have several conditions of data that use cluster ways and microarray knowledge to classify the genes. Based on the literature and different statistical approaches, there are several clustering algorithms that are used to built gene expression data and classification, however little attention has been paid to uncertainty within the results obtained

To get a transparent image on the sight of a angle, a transparent cancer classification associate analysis system has to be pictured followed by a scientific approach to analyse international organic phenomenon that provides an optimized answer for the known drawback space. Molecular medical speciality provides a possibility of human cancer classification, however these tests don't seem to be wide applied as a result of characteristic molecular markers for many solid neoplasm save nevertheless to be known.

Revised Manuscript Received on July 05, 2019.

G Sivagamasundari Research Scholar, School of Engineering and Technology, Pondicherry University, India

Latha Parthiban, Department of Computer Science, Pondicherry University CC, Puducherry, India **Latha Parthiban,** Department of Computer Science, Pondicherry University CC, Puducherry, India **Second Author name**, His Department Name, University/ College/ Organization Name, City Name, Country Name.

II. BACKGROUND WORK

Many researchers presented data clustering that discovers the similarity or dissimilarity between groups of items in a dataset. Automated weighted method for attribute selection, generate the initial cluster centers rather than random or user specified cluster centres. An initialization problem can be resolved by binary search based initialization method to initialize cluster points for k-means algorithm. K-Means algorithm is simple, efficient and provides easy convergence. But it has drawbacks like initial cluster centres, stuck in local optima. Lack of knowledge, lack of universal method, and no information about number. of clusters.

Anirban Mukhopadhyay et al 2013 stated that data clustering is an important task of datasets, where a set of data is grouped on its similarities. But there occurs an optimization problem with the clusters[1]. To overcome this multiple genetic based algorithm is applied on clusters characteristics such as compactness, separation and connectivity. The proposed approach Interactive Genetic Algorithm methodology uses a set of objective functions to arrive with the best results. But biological information is taken into account for validating the clusters. .

In [3], the authors described the gene clustering with a valid proximity measure which is useful in achieving accurate results for microarray data. The approach proposed in [4] takes into account 16 proximity measures in 52 data sets for the clustering of genes using Intrinsic Biological Separation Ability (IBSA) methodology. Result proposes Rank-Magnitude correlation coefficient as a measure for cancer and YS1 for time-course experiments. Proximity measurement measures the similarity between two data objects or gene expression patterns. There is no proper advice for selecting a proper proximity measure for the

cluster, as it is very useful for the clustering of gene expression time course data.

.Balasundar I. Raju et.al 2013[2] proposed a new method called gene therapy is an experimental practice to treat or prevent disease using a person's genes. Ultrasound-mediated Delivery (USMD) methodology is used in clinical practice for safety and effectiveness. The USMD delivers the gene to the liver via tail vein site, using the process named sonoporation. This proves to be successful as the circulation time of Plasmid DNA is increased and the energy is delivered deep into the tissue.

Authors in [5] described the various tools that are required to integrate microarray gene expression (MAGE) data, as it is an important factor of gene-expression analysis. So the task is to identify the differences in batch execution effects.

Authors in [6-7] proved that cancer identification and classification tasks are important process during tumor discovery. Micro-array based gene expression proved to produce fruitful cancer identification results. The proposed approach uses support vector machines (SVM) as an iterative procedure and employs unlabled data and produces experiential results. But the situation is corrupted when introducing labeled data in the scene. The potential gene markers for each cancer subtype has been identified, which helps in successful cancer diagnosis. It manages an unlabeled gene expression data and achieved better empirical success .It is time consuming, expensive and tested with limited samples. In [8-9], the authors presented the DNA micro arrays and methods for gene ranking or gene expression data that includes T-score and ANOVA(Selection of important genes)

III. EXPERIMENTAL RESULTS

Microarray technology has made it possible to study the expression levels of thousands of genes simultaneously. In this work, real data collected from the hospitals of different patients have been used. The Patients data have been compared with the expert's documental data as shown by heat map solution. Proper Examination of gene expression data points to cancer identification and classification, that helps in proper treatment selection.

The implementation contains four stages: Data Set Input, Gene Knowledge Extraction, Ontological Mapping and Gene Expression Design as shown in fig 2 to fig 5. Fig 6 shows the ontological mapping implementation, fig 7 the heat map, fig 8 the knowledge data

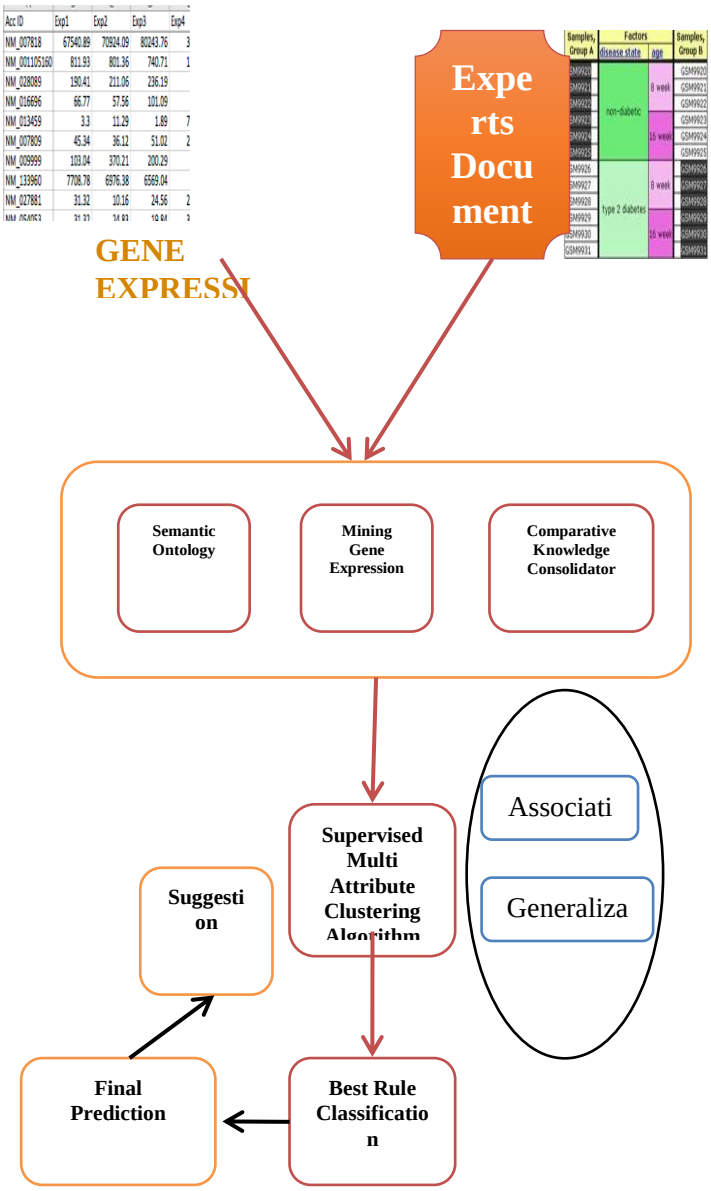

Fig 1 Architecture Diagram (KEOPS method)

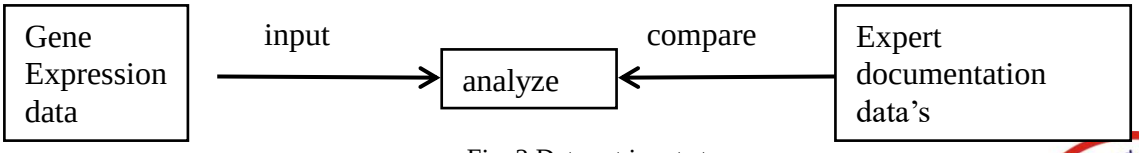

Fig 2 Data set input stage

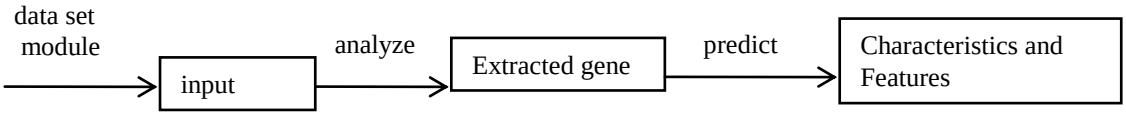

Fig 3 Gene Knowledge extraction

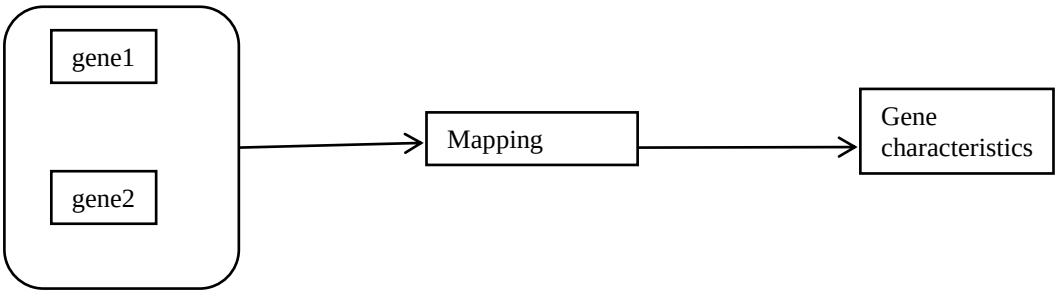

Fig 4 Ontological Mapping

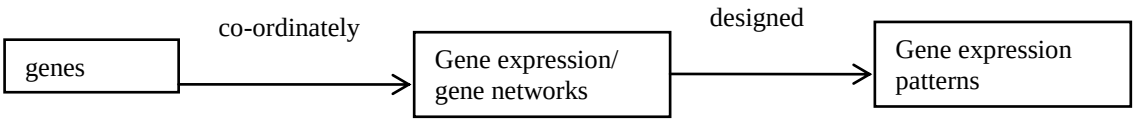

Fig 5 Gene Expression Design

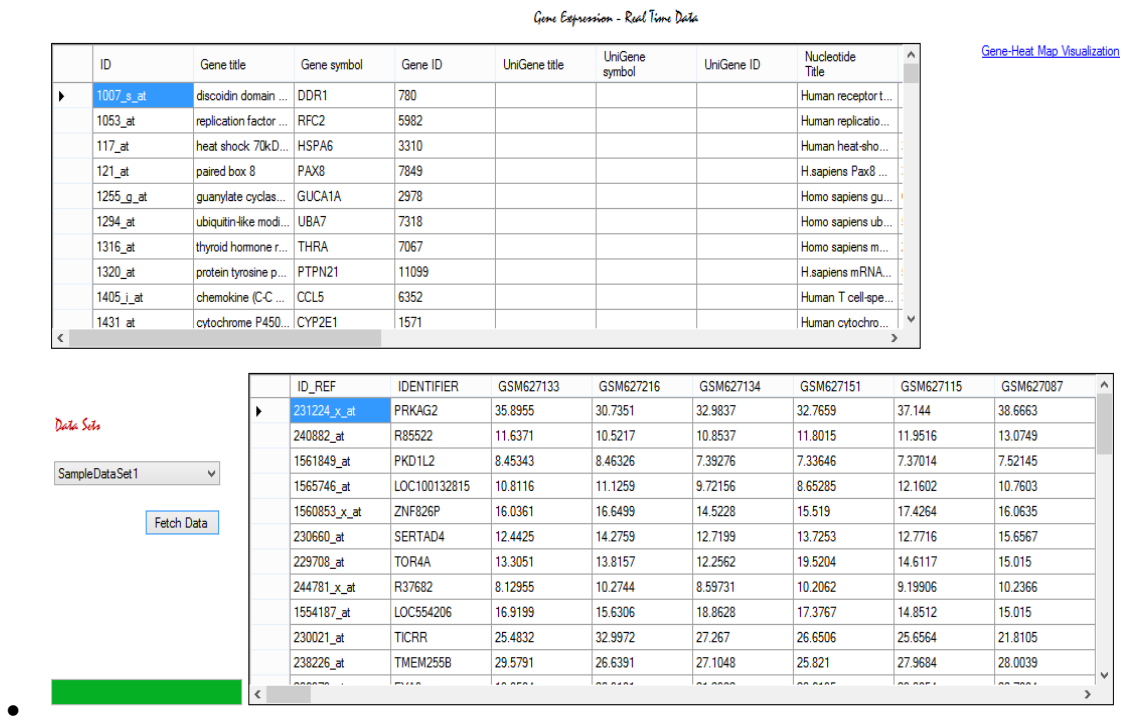

Fig 6 Ontological mapping

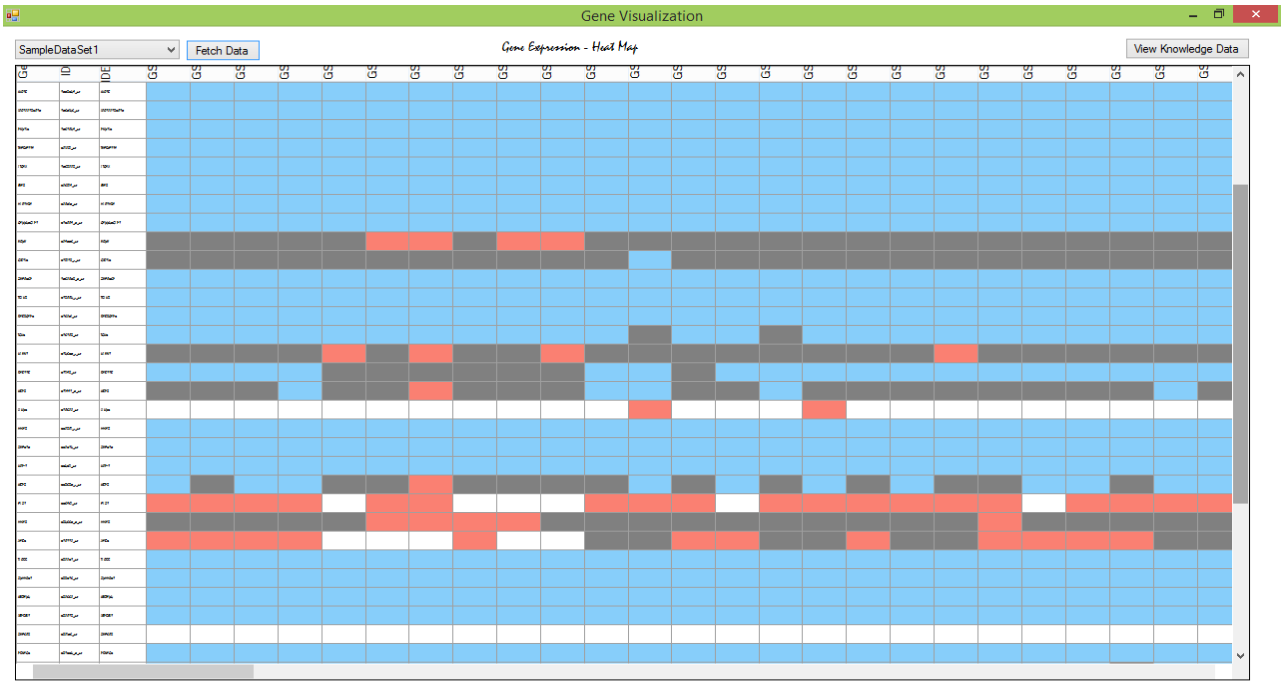

Fig 7 Heat map

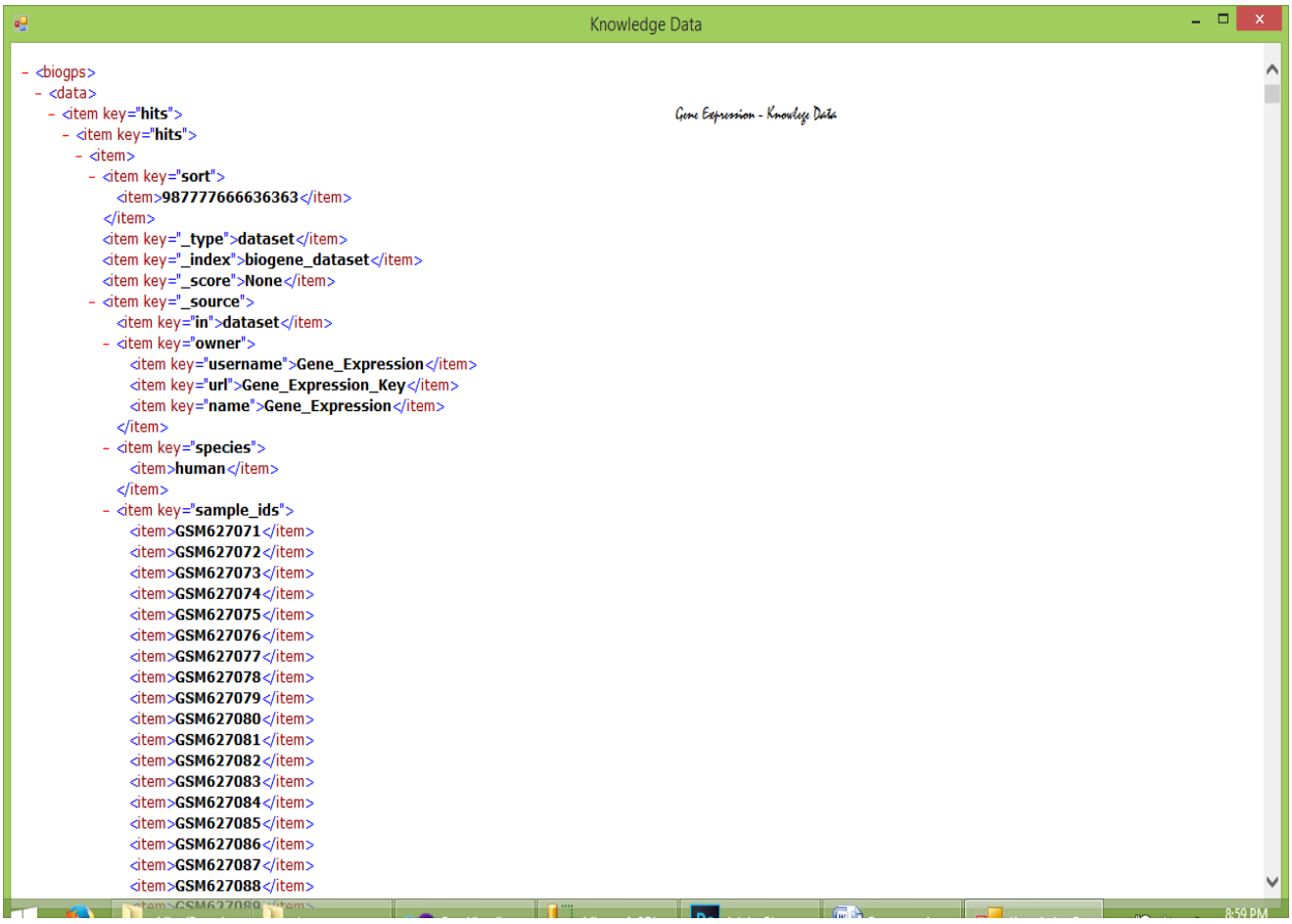

Fig 8 Knowledge data

Table 1 Data set input

ID_REF	IDENTIFIER	GSM627133	GSM627216	GSM627134	GSM27151	GSM27115
231224_at	PRKAG2	35.8955	30.7351	32.9837	32.7659	37.144
240882_at	R85522	11.6371	10.5217	10.8537	11.8015	11.9516
1561849_at	PKD1L2	8.45343	8.46326	7.39276	7.33646	7.37014
1565746_at	LOC100132815	10.8116	11.1259	9.72156	8.65285	12.1602
1560853_at	ZNF826P	16.0361	16.6499	14.5228	15.519	17.4264
230660_at	SERTAD4	12.4425	14.2759	12.7199	13.7253	12.7716
229708_at	TOR4A	13.3051	13.8157	12.2562	19.5204	14.6117
244781_at	R37682	8.12955	10.2744	8.59731	10.2062	9.19906
1554187_at	LOC554206	16.9199	15.6306	18.8628	17.3767	14.8512
230021_at	TICRR	25.4832	32.9972	27.267	26.6506	25.6564
238226_at	TMEM255B	29.5791	26.6391	27.1048	25.821	27.9684

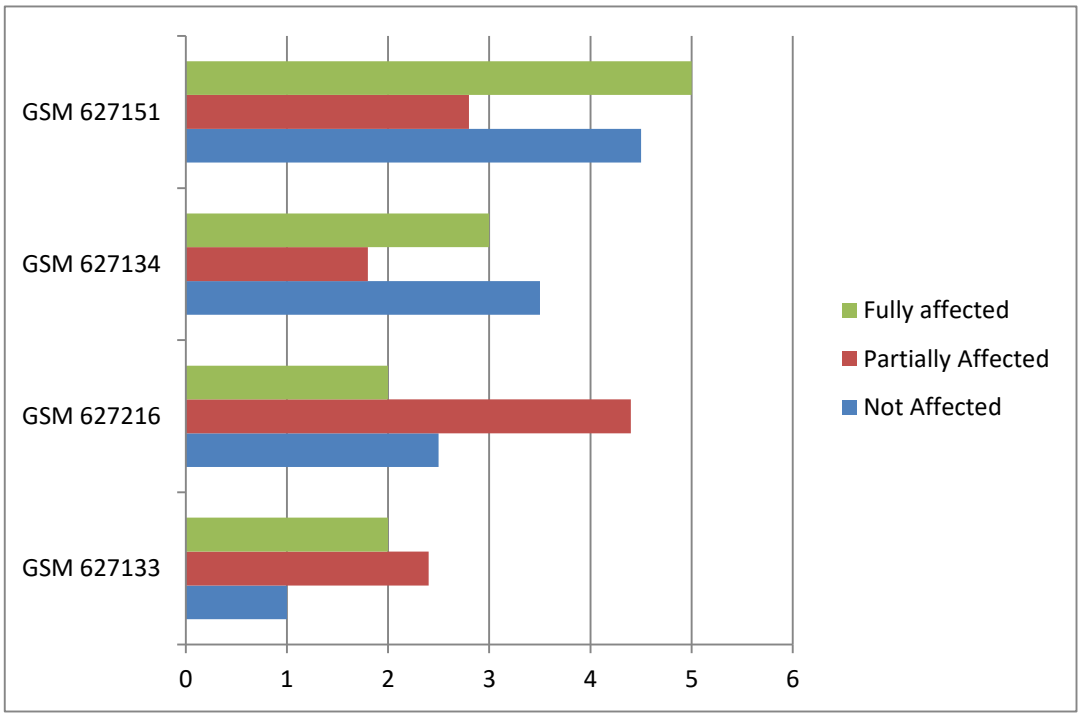

Fig 9 Heat map showing the affected levels

IV. CONCLUSIONS

DNA Microarrays and heat maps are used to examine less gene and not entire genome. Prediction can be done easily when the heat map shows the affected levels. Sequencing in future can help in locating faulty genes as well as varied type of disease with better accuracy with the help of DNA chromosome sequence mutation. Where the mutation sequence can help to identify the cancer causing tissues.

REFERENCES

1. Anirban Mukhopadhyay,ujjwal maulik and sanghamitra Bandyopadhyay (2013) 'An Interactive approach to multiobjective clustering of gene expression patterns' IEEE transactions on biomedical engineering,volume.60 No.1.
2. Balasundar I.Raju,Evgeniy Leyvi,Ralf seip,Shriram sethuraman,Andrew bird,Songtao Li,and Dwight koberi (2013) 'Enhanced Gene Expression of Systematically Administered Plasmid DNA in the Liver with Therapeutic Ultrasound and Micro bubbles' IEEE transactions on Ultrasonics,Ferroelectricals,and Frequency control, volume 60,No.1.
3. Bharathi And Dr.A.M.Natarajan (2010) 'Cancer Classification Of Bioinformatics Data Using ANOVA ' International Journal Of Computer Theory And Engineering, Vol. 2, No. 3, 1793-8201.
4. Colin molter,Robin Duque,Hugues Bersini, and Ann Nowe(2013) 'GENESHIFT: A Nonparametric Approach for Integrating MicroArray Gene Expression Data Based on the Inner product as a distance Measure between the distributions of genes',IEEE/ACM transactions on computational biology and bioinformatics,volume 10,No.2.
5. Dimitris Maroulis, Dimitris Iakovidis, Ilias Flaounas and Stavros Karkanis(2006) 'A gene expression analysis system for medical diagnosis', IFIP International Federation for

Information Processing Volume 4 No 2
6. Huimin Zhao (2007) 'A multi-objective genetic programming approach to developing Pareto optimal decision trees', -ELSEVIER decision support systems 43 809-826.
7. Jaskowik, Ricardo, J.G.B.Campello,and Ivan G.Costa(2013) 'Proximity Measures for Clustering Gene Expression Microarray data:A Validation Methodology and a Comparative Analysis' IEEE transactions on biology and bioinformatics, Volume 10,No.4
8. Rui Xu, Anagnostopoulos, G.C. And Wunsch, D.C.I.I(2007) 'Multiclass Cancer Classification Using Semi supervised Ellipsoid ARTMAP And Particle Swarm Optimization With Gene Expression Data' IEEE/ACM Transactions On Computational Biology And Bioinformatics, Volume.4, No.1, pp. 65-77.
9. Shaurya Jauhari ans S.A.M.Rizvi(2014) 'Mining Gene Expression Data focusing cancer Therapeutics:A Digest' IEEE transactions on computational biology and bioinformatics,Volume.11 No.3.