Mining Microarray Gene Expression Data using Machine Learning Techniques

Synopsis Submitted by

Shilpi Bose

Doctor of Philosophy (Engineering)

School of Education Technology
Faculty Council of Engineering &Technology
Jadavpur University
Kolkata-700032, India
2022

Mining Microarray Gene Expression Data Using Machine Learning Techniques

by

Shilpi Bose

Under the guidance of

Prof. Matangini Chattopadhyay

Professor

School of Education Technology

Jadavpur University

Kolkata, India

&

Dr. Chandra Das

Associate Professor

Department of Computer Sc. & Engg.

Netaji Subhash Engineering. College

Kolkata, India

School of Education Technology

Faculty Council of Engineering &Technology

Jadavpur University

Kolkata-700032,India

2022

Contents

1. Introduction

Due to intense research activities and wide use of high-throughput technology in the area of biological sciences, society is experiencing an explosion of biological data. As the size of the biological databases increases day by day, analyzing this enormous volume of biological data has become complicated. This analysis is very much crucial to elucidate several secrets of life and several aspects of medical sciences. The most efficient method to investigate these data is via laboratory experiments which involve lots of time, money, and manpower. So, effective and efficient computational tools are needed to store, analyze, and interpret these different types of biological data. In this regard, a new field called bioinformatics [1,2] has risen to overcome the above-mentioned issues.

Bioinformatics [1,2] is the conceptualizing biology in terms of molecules and applying informatics techniques to understand and organize the information with these molecules, on a large scale. It involves the creation and advancement of algorithms using techniques including machine learning, data mining, pattern recognition, applied mathematics, statistics, informatics, and biochemistry to store, analyze and interpret this vast amount of biological data [3]. Major research efforts in this field include sequence alignment and analysis, gene finding, genome annotation, protein structure alignment and prediction, classification of proteins, clustering and dimension reduction or feature selection from gene expression, protein-protein docking or interactions, and the modeling of evolution. Hence, in other words, bioinformatics can be described as the application of computational methods to make biological discoveries [1,2,3].

Machine learning [4,5] is one sub field of artificial/machine intelligence which is related to the study of computer algorithms that provides systems the ability to automatically learn and improve from experience. Machine learning algorithms allow the systems to make decisions autonomously without any external support. Such decisions are made by finding valuable underlying patterns within complex data. Machine learning techniques are divided into three broad categories: supervised learning, unsupervised learning, and reinforcement learning. These techniques are used in several ways in different fields like pattern recognition, image processing, data mining, natural language processing [4-12]etc.

Data mining [11, 12] is a branch of computer science where several hidden information are extracted from data using several mining techniques. Data mining techniques are widely used in medical data examination field for analyzing, extracting, transforming, interpreting and visualizing medical records stored in repositories. Medical data mining is very important for improvement of medical therapy and in parallel it is very challenging also, because diagnosis and prediction of diseases are directly related to a matter of life and death of patients. A wrong classification or prediction can be disastrous to the life of patients and their relatives. Data mining techniques consist of two types of techniques: data management techniques and data analysis techniques. Among the several data analysis techniques, machine learning techniques are widely used in medical data mining field to make decisions to easily and quickly diagnose and predict diseases.

In the several types of bio-technology, microarray technology is one of the most popular high-throughput bio-technology which is used to measure the expression level of a huge number of genes simultaneously in particular cells or tissues [13]. Results of microarray technology are large matrices known as gene expression data matrices where a row contains information about a gene; a sample/experiment is represented by a column and a cell contains information about a gene for a specific sample/experiment. Microarray data analysis has been successfully applied in a number of studies over a broad range of biological and medical disciplines, including identification of functions of novel genes, identification of pathway in gene regulatory network, cancer classification by class discovery and prediction, identification of unknown effects of a specific therapy, identification of genes relevant to a certain diagnosis or therapy, and cancer prognosis etc. [14-30]. So, microarray gene expression data analysis has important aspects in real life applications.

Due to several shortcomings of microarray experiments [31], considerable missing values (MVs) are introduced in the resultant matrix [31]. Sometimes, a large number of genes (up to 90%) are affected and contain missing values [31]. Such incomplete matrices pose a problem in analysis algorithms [14-30] as they need complete matrices. It is not feasible to repeat microarray experiments as they are overwhelmingly costly. So, designing algorithms for predicting these missing values accurately have become very important. Accuracy of these prediction methods can affect the results of analysis algorithms as these methods require complete gene matrices. So, missing value prediction in gene expression data is a mandatory preprocessing task before analysis.

The challenge is, therefore, to devise powerful machine learning methodology-based data mining techniques to preprocess and analyze gene expression data in more efficient ways. The systems should have the capability of flexible information processing to deal with real life ambiguous situations and to achieve tractability, robustness, and low-cost solutions. In the above background, the focus of the research undertaken in this thesis is presented next.

2. Aim and Works of the Dissertation

The major focus of this research work is to devise machine learning methodology based new data mining techniques to preprocess and analyze gene expression data, which are efficient in terms of prediction accuracy.

2.1 1st Work: Pre-processing on Microarray Gene Expression Data based on Clustering technique: A Framework for Neighborhood Configuration to Improve the KNN based Imputation Algorithms on Microarray Gene Expression Data

In view of the several technical problems associated with microarray experiments, a considerable number of entries are found missing in a typical microarray gene expression dataset. As a consequence, due to the unavailability of complete data, the effectiveness of the downstream analysis algorithms deteriorates. Different imputation techniques are employed to address this problem. These techniques are developed based on two approaches. One approach is weighted average based methods and second one is numerical approach-based

methods. Among these techniques, the numerical methods are more robust than the weighted average based methods but the weighted average based methods are widely used in several applications as these methods generate consistent results and are algorithmically simple, but these methods also suffer from some drawbacks that are seldom elaborated upon. These deficiencies have been pointed out in our first work. To solve these problems, in the first work we have first proposed a primary framework via proposing a new hybrid distance based a new version of the *K*-nearest neighbor imputation method (*K*NNimpute) named iterative sequential *K*-nearest neighbor imputation method (IS*K*NNimpute). The IS*K*NNimpute with the hybrid distance does not capable of solving all those deficiencies. So, we have introduced a new robust framework which is embedded in the *K*-nearest neighbor imputation method (*K*NNimpute), as well as in its different versions. The idea is to achieve better neighborhood formation, in order to improve the prediction accuracies. The new framework is developed using Euclidean distance, Pearson correlation coefficient, and mean square residue score.The new framework is tested on ten well-known microarray datasets. From the experimental results it has been found that in each and every case, the proposed modified methods significantly outperform their corresponding traditional versions and are also comparable with the existing robust numerical methods. Among all the versions, IS*K*NNimpute with the new framework is better than other *K*NN versions.

Experimental Results

The effectiveness of the proposed algorithm is certified by carrying out a large number of experiments over ten microarray gene expression datasets. For comparing the efficiency of the proposed methods, different versions of the proposed methods are compared with the well-known weighted average based and numerical methods based on existing missing value estimation techniques. The accuracy of the proposed methods in comparison with the above-mentioned existing techniques has been ensured using the following metrics:(a) normalized root mean squared error (NRMSE) [40] and (b) average distance between partitions error (ADBPE) [41].

In Figure 1 the different proposed versions of *K*NNimpute are compared with their corresponding existing versions with respect to different distances like Euclidean, Pearson, PEH distance in terms of NRMSE for ROS dataset. In Figure 2 and Figure 3 the proposed versions are compared with different existing popular methods LLSimpute [42], SVDimpute [43], BPCA[44] in terms of NRMSE and ADBPE respectively. In all cases the proposed versions show superior performance.

For our experimentation, different microarray gene expression datasets have been used. These datasets are classified into three categories: (1) time-series dataset (SP.AFA[45], SP.ELU[45], BAL[46]) (2) mixed data set. Mixed dataset comprise time-series data as well as non time-series data or multiple time-series data measured in different experimental conditions, YOS[47]) (3) non-time series dataset (GAS[48], ROS[49], GOL[50], Tymchuk[51] and HIR[52]). A synthetic dataset generated by SynTReN [53] is also considered here.

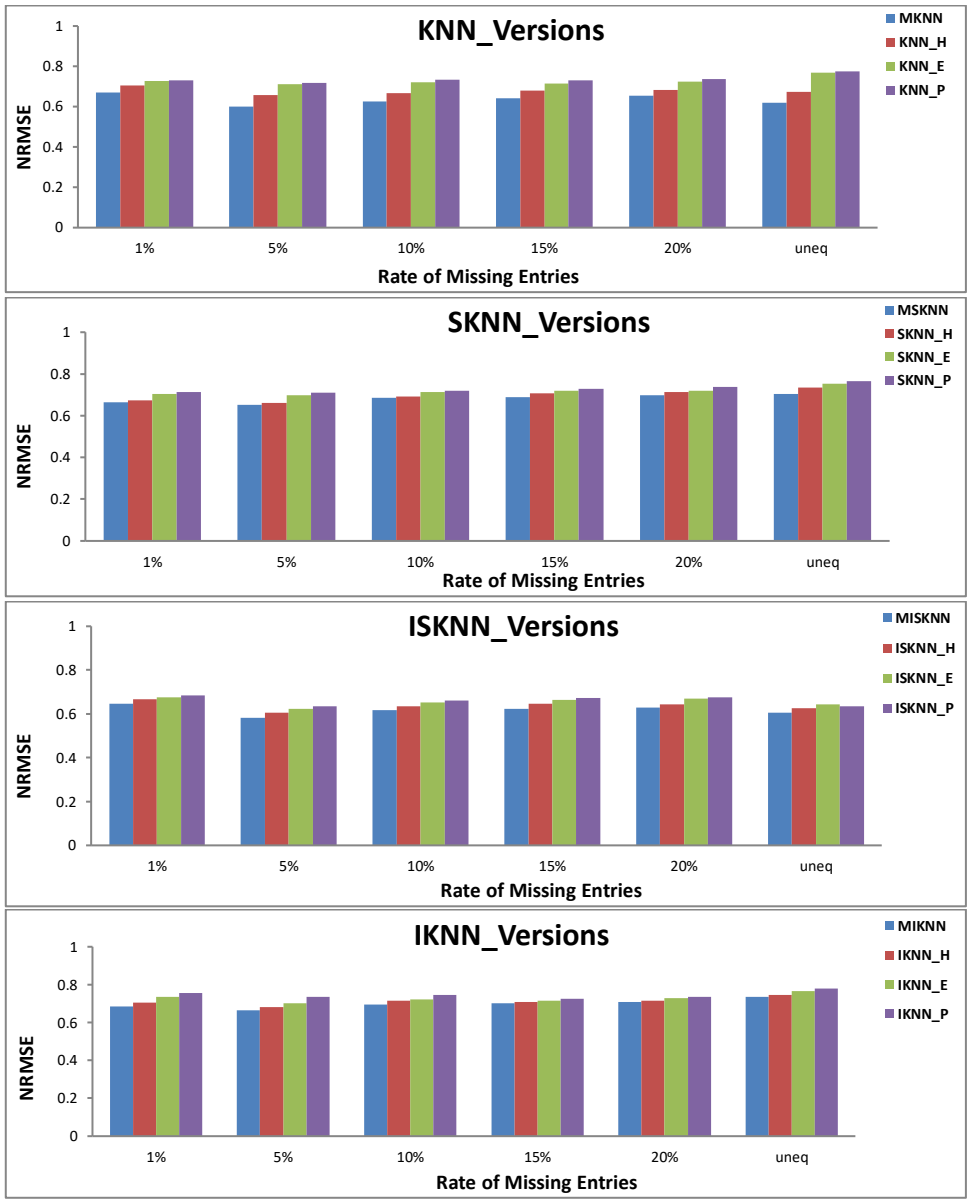

Figure 1 . Comparative performance analysis of different versions of the proposed method with its existing versions using different distances in terms of NRMSE for ROS dataset

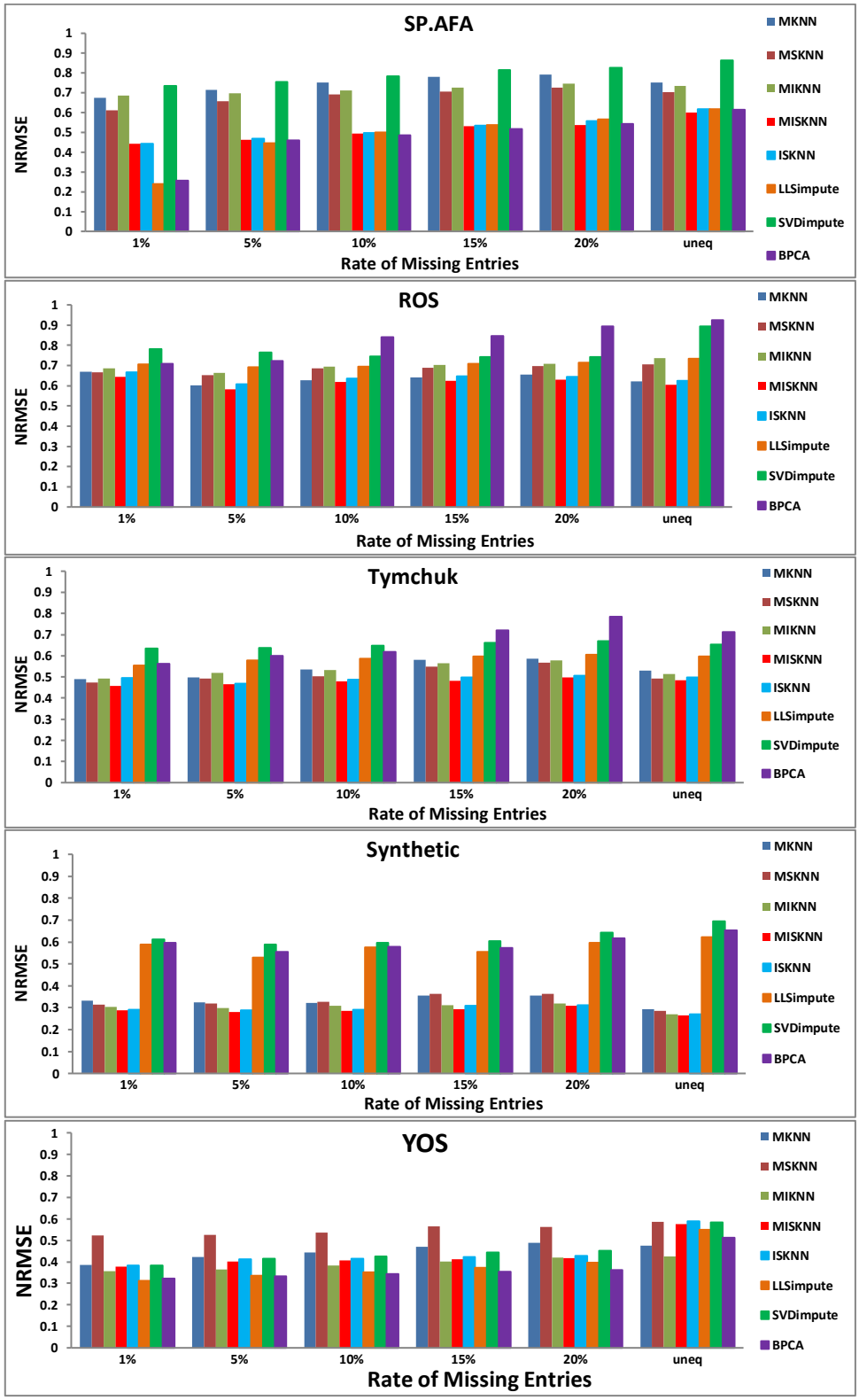

Figure.2. Comparative performance analysis of different versions of the proposed method for existing well-known imputation techniques in terms of NRMSE

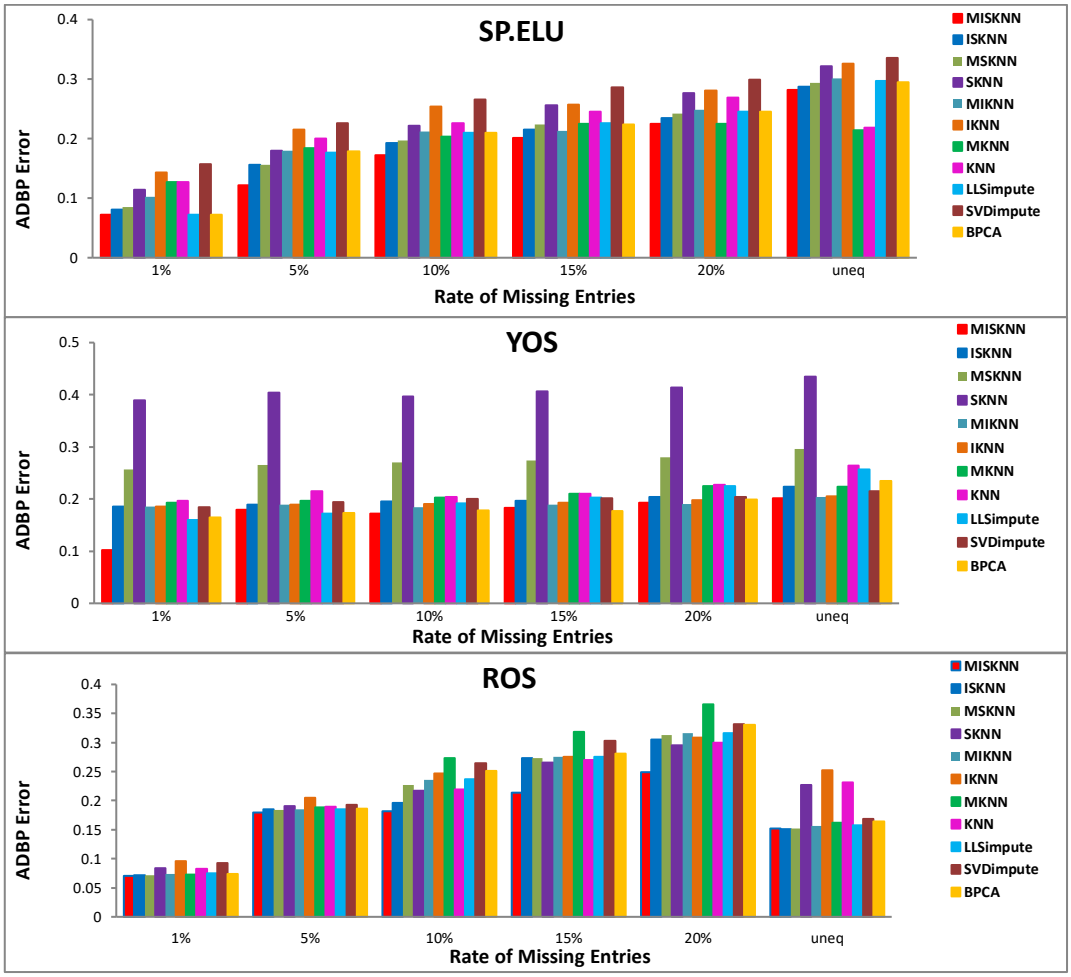

Figure 3. ADBP error of different methods for SP.ELU, YOS, and ROS datasets

Conclusion

In this work, the modified *K*NNimpute, and its different versions are proposed for improving the prediction accuracy of the traditional *K*-nearest neighbor rule-based estimation techniques. The motivation behind this work is that the neighborhood for each target gene will be constructed in such a way that the weighted average procedure will run only on the maximally positively co-expressed and magnitude wise closest genes.In all cases, the proposed methods give better performance than their corresponding traditional versions, and among them, the prediction accuracy of the modified IS*K*NN-impute is the best. For local structure-based datasets, this method significantly gives similar results compared to other numerical methods.

In the next work, we have proposed another missing value prediction technique for microarray gene expression data via integrating clustering and numerical approach.

2.2 2nd Work: Clustering and Numerical technique based Pre-processing on Microarray Gene Expression Data: Bi-clustering based Sequential Interpolation Imputation Technique for Missing Value Prediction in Microarray gene expression data

In the second work, we have given focus on developing a new imputation method via combining clustering and numerical approach to improve prediction accuracy. It has been already found that prediction accuracy of numerical methods is high but these methods sometime generate inconsistent results. Existing numerical methods are very complex and hard to implement. Due to unavailability of codes in the internet these methods are not used frequently in the applications. Considering this view, we have proposed a new imputation method which is a combination of bi-clustering and interpolation based numerical work. Here, using bi-clustering the neighborhood is first formed and then using interpolation missing values are imputed. The proposed work is simple compared to existing numerical methods and shows its superiority.

Experimental Results

In this paper, efficiency of the proposed BiSIimpute is evaluated by comparing it with a number of existing eminent imputation techniques, namely *K*NNimpute [43], S*K*NNimpute [55], LLSimpute [42], SVDimpute [43], Bayseian principal component analysis (BPCA)[44], NL[51], and bi-iLS[56]. Experiments have been carried out over nine different datasets. Accuracy of our algorithm is compared with other algorithms using two well-known metrics: i) normalized root mean squared error (NRMSE) and ii) average distance between partitions error (ADBPE) as discussed in the previous chapter. All the datasets are used here from previous experiment.

All the above mentioned methods are implemented in C using Linux environment in a machine with a 4 GB RAM, and 3.2 GHz core i3 processor.

Figures 4 and Figure 5 show the comparative experimental results to prove effectiveness of the proposed framework.

Conclusion

In this work, a bicluster-based sequential interpolation imputation method called BiSIimpute is proposed for estimation of missing values in DNA microarray data. The novelty of this method is that first time interpolation based imputation technique is applied in biclustering framework. Using NRMSE, and ADBPE metric, it is found that the proposed method outperforms all other methods mentioned here for different local structured based datasets. So, it is a new robust approach to estimate missing values in different local structured based microarray gene expression datasets. After data pre-processing several machine learning techniques are applied to mine gene expression data. One such mining task is to find functionally similar genes from microarray gene expression data which is discussed in the next work.

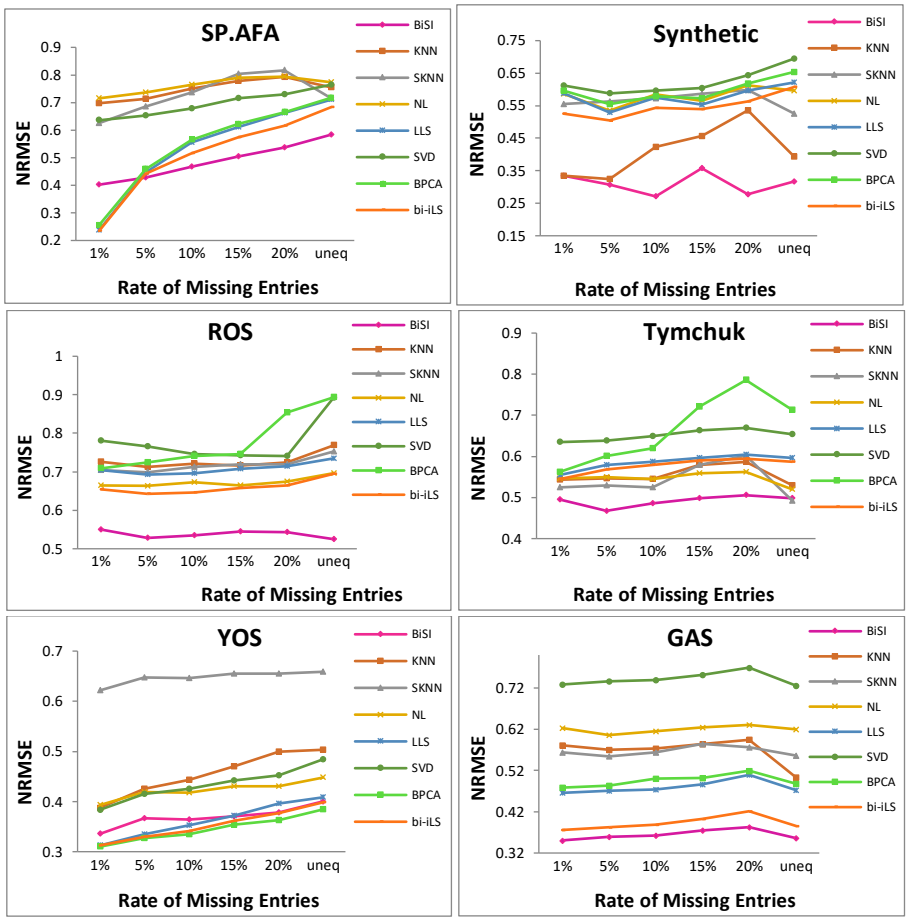

Figure 4. Comparative performance analysis of different methods based on NRMSE for different datasets with different percentage of missing entries

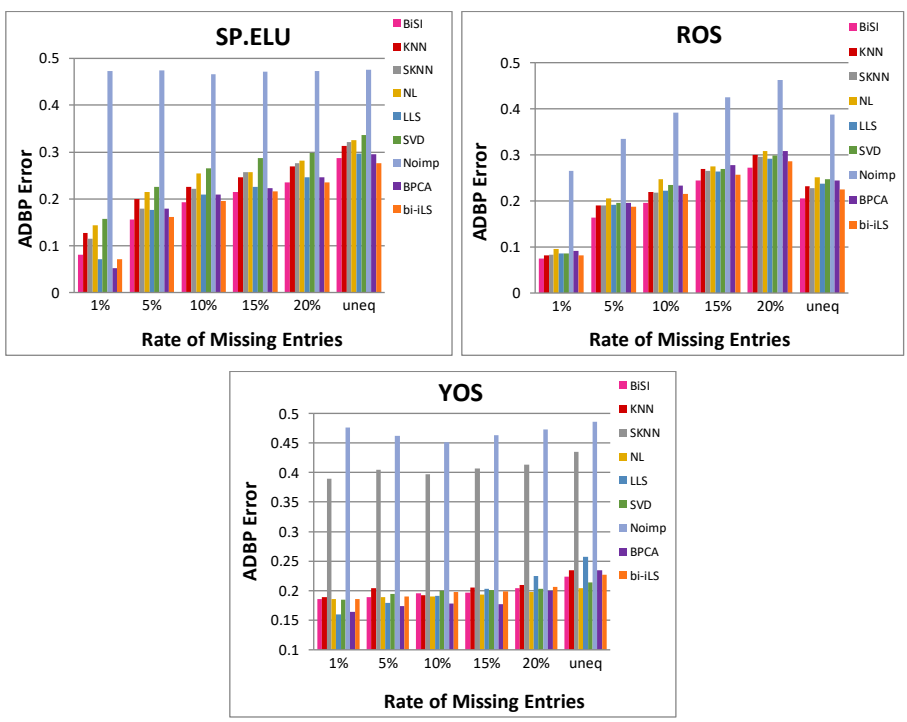

Figure 5.Comparative Performance Analysis of different methods with respect to ADBP error for SP.ELU, ROS, and YOS datasets

2.3 HCFPC: A New Hybrid Clustering Framework using Partition-Based Clustering Algorithms to Group Functionally similar Genes from Microarray Gene Expression Data

One important analysis task of microarray gene expression data is to find hidden patterns among genes present in this data to extract relevant information which will be beneficial for functional genomics. It has been already found that genes with similar expression patterns (co-expressed genes) may have similar biological functions. The information from these hidden patterns among genes may help in analysing functional enrichment of genes, understanding gene function of uncharacterized genes, understanding cellular processes, gene co-regulation or relation in functional pathways, and finding out information related to transcriptional regulatory networks. So, it is a great challenge to identify groups of genes based on similar patterns from this large voluminous gene expression data. Clustering techniques [10, 14,15] are widely used in gene expression data for clustering genes to partition genes among relevant functional groups. A huge number of different clustering techniques are already developed to solve this problem. However, the clustering of genes is an old problem but as gene expression data is very much noisy so proper noise deletion is an important task before clustering and is still challenging. Although tight clustering methods are developed, these methods have several computational limitations. Among the different category-based clustering methods, partition-based clustering methods are most popular but these methods are unable to eliminate noise. Here, in the third work, we have designed a novel framework using different partition-based clustering algorithms (mainly different versions of *K*-Means) to provide an intuitive model for eliminating noise and also generating functional gene clusters. The model is also capable of clustering genes without using any predefined *K* as *K* is automatically detected here.

Experimental Results

In this research work, the performance of HCFPC is compared with that of hard *k*-means (HKM) , fuzzy *k*-means (FKM) , possibilistic *k*-means (PKM) , cluster identification via connectivity kernels (CLICK) , and self-organizing map (SOM) [69] on different microarray gene expression data sets (with noise and without noise version). The major metrics for evaluating the performance of different algorithms are Silhouette index (SILH) [70], Davies-Bouldin index (DB) [71], Dunn index (DUNN) [71]. Also, the biological significance of the generated gene clusters generated by HCFPC algorithm is analyzed using the Gene Ontology Term Finder [72].

In this paper, four publicly available microarray times series gene expression datasets are taken for making comparative study. The description of the datasets is given in Table 1, which are downloaded from Gene Expression Omnibus (http://www.ncbi.nlm. nih.gov/geo/).

Table 1.Dataset Description

Dataset Name	Species	Number of Rows/Genes	Number of Columns/Time-Points
GDS2910(Noisy)	Yeast	2746(1900+846)	191
GDS1116(Noisy)	Yeast	1081(798+281)	131
GDS2002	Yeast	5617	30
GDS2003	Yeast	5617	30

Dataset Preparation:

Four datasets have been used here for study. Among these, GDS2002 and GDS2003 datasets required no pre-processing while GDS1116 and GDS2910 required pre-processing. Initially all rows containing missing values greater than 10% are deleted. The missing values are then filled with row-average values to get the complete data. After that artificially noise has been introduced. Among those deleted rows, rows with missing values above 75% are selected and those missing values are filled with random values outside the range of possible values (i.e. greater than the maximum possible value and less than the minimum possible value) in previously created complete data. These artificially created rows are then appended with the complete data to create the noisy datasets with 15 to 20% of noise elements.

Table 2. Performance Comparison of the Proposed Framework with HKM, FKM, and PKM on Normal Datasets

Dataset Name	Clustering Methods	Cluster No.	DB Index	Silhouette Index	Dunn Index
GDS2910	HCFPC framework	30	1.49	0.08	0.06
	HKM		2.47	0.05	0.049
	FKM		6.13	-0.06	0.045
	PKM		6.29	-0.17	0.037
GDS1116	HCFPC framework	5	1.91	0.12	0.042
	HKM		1.98	0.09	0.047
	FKM		3.97	-0.11	0.032
	PKM		4.65	-0.13	0.036

Table 3.Performance Comparison of the Proposed Framework with HKM, FKM, and PKM in Presence of Noise

Dataset	Evaluation Metric	HCFPC framework		HKM		FKM		PKM	
		Normal dataset with inherent noise	With noise	Normal dataset with inherent noise	With noise	Normal dataset with inherent noise	With noise	Normal dataset with inherent noise	With noise
GDS2910	DVB	1.49	1.55	2.47	2.76	6.13	8.43	6.29	12.54
	SILH	0.08	0.04	0.05	0.01	-0.06	-0.21	-0.17	-0.16
	DUNN	0.06	0.076	0.049	0.52	0.045	0.46	0.037	0.32
GDS1116	DVB	1.49	1.54	1.98	2.03	3.97	3.1	4.65	14.86
	SILH	0.12	0.11	0.09	0.13	-0.11	0.71	-0.13	-0.25
	DUNN	0.042	0.043	0.047	0.37	0.032	0.43	0.036	0.47

Table 4.Performance Comparison of the Proposed Framework with other Clustering Methods

Indices	Clustering Algorithms	GDS2002	GDS2003
		Normal dataset	Normal dataset
DB Index	CLICK	26.7	17.61
	SOM	13.41	15.22
	HCFPC framework	0.17	0.19
Silhouette Index	CLICK	-0.12	-0.09
	SOM	-0.05	-0.06
	HCFPC framework	0.89	0.93
Dunn Index	CLICK	0.03	0.05
	SOM	0	0.01
	HCFPC framework	4.1	2.43

Table 5.Significant GO Terms Obtained Using Proposed Algorithm for GDS2003

Ontology Aspects	Cluster Number	Gene Ontology term	Cluster frequency	Genome frequency	Corrected P-value	FDR	FALSE Positives
Biological Process	12	cytoplasmic translation	0.357	0.029	3.25E-117	0.00%	0
	3	mitochondrial translation	0.284	0.024	5.39E-73	0.00%	0
	18	ribosome biogenesis	0.538	0.067	3.87E-72	0.00%	0
	18	ribonucleoprotein complex biogenesis	0.543	0.08	1.15E-64	0.00%	0
	18	rRNA metabolic process	0.431	0.054	9.40E-55	0.00%	0
	12	peptide biosynthetic process	0.435	0.114	6.06E-54	0.00%	0
	12	organonitrogencompound biosynthetic process	0.551	0.187	2.85E-53	0.00%	0
	3	mitochondrion organization	0.284	0.04	3.02E-50	0.00%	0
Molecular Function	12	structural constituent of ribosome	0.315	0.033	4.86E-85	0.00%	0
	3	structural constituent of ribosome	0.219	0.006	4.15E-36	0.00%	0
	9	electron transfer activity	0.235	0.025	2.24E-20	0.00%	0
	9	active transmembrane transporter activity	0.324	0.025	1.56E-17	0.00%	0
	18	snoRNA binding	0.086	0.04	4.65E-17	0.00%	0
Cellular Component	3	mitochondrion	0.781	0.174	3.79E-125	0.00%	0
	12	cytosolic ribosome	0.329	0.023	3.36E-119	0.00%	0
	3	mitochondrial protein-containing complex	0.34	0.03	1.51E-87	0.00%	0
	12	ribonucleoproteincomplex	0.463	0.091	1.23E-79	0.00%	0
	18	preribosome	0.35	0.024	8.68E-64	0.00%	0
	12	intracellular non-membrane-bounded organelle	0.581	0.21	1.72E-53	0.00%	0

Comparative Performance Analysis

The performance of the proposed framework is compared with different existing partition based methods and other methods are compared on different datasets without noise and with noise. The proposed framework with the above-mentioned features performs significantly better than other partition-based clustering algorithms and other types of clustering algorithms for all microarray datasets (in presence of additional noise also) in terms of different quantitative indices and also provides biologically significant clusters. Results are given in Tables 2 to 5.

Conclusion

In this work, a new gene clustering framework is developed by integrating different partition-based clustering algorithms in a novel manner. The main novelty of this framework is that, it can work in presence of noise and after detecting noisy genes, it can eliminate them and generates good qualitative clusters with small set of significant genes. Apart from this instead of random centroid selection it selects centroid in a novel manner.

The proposed framework with the above-mentioned features performs significantly better than other partition-based clustering algorithms and other types of clustering algorithms for all microarray datasets (in presence of additional noise also) in terms of different quantitative indices and also provides biologically significant clusters.

In the next work, we have proposed a new ensemble machine learning model for cancer sample classification from gene expression data.

2.4 An Ensemble Machine Learning Model based on Multiple Filtering and Supervised Attribute Clustering Algorithm for Classifying Cancer Samples

Sample classification is one of the important downstream analysis based application of microarray gene expression data. The gene expression data matrix contains a huge number of genes compared to a limited number of samples and this is a most important problem for sample classification. Most classification algorithms suffer from such a high-dimensional input space. Also, a very small number of genes contain relevant information for sample classification. Secondly, the class imbalance problem is an overhead for sample classification also. In this regard, in the fourth work, a new ensemble machine learning classification model named Multiple Filtering and Supervised Attribute Clustering algorithm based Ensemble Classification model (MFSAC-EC) is proposed which can handle class imbalance problem and high dimensionality of microarray datasets. The MFSAC method is a supervised feature selection technique combining multiple filters with a new supervised attribute clustering algorithm. Using MFSAC method different sub datasets are formed based on different filtering measures and then for every sub dataset, a base classifier is constructed separately, and finally, the predictive accuracies of these base classifiers are combined using the majority voting technique forming the MFSAC-based ensemble classifier.

Experimental Results

To assess the performance of the proposed MFSAC-EC model, four well-known existing classifiers named K-Nearest Neighbor [86], Naive Bayes [86], Support vector machine [87], and Decision tree(c4.5) [86] are applied independently in this model and four different ensemble classification models are formed. To prove the superiority of the proposed model, it is compared with existing well-known filter methods (used here) and existing recognized gene selection methods [76, 88-89] and also with different existing ensemble classifiers [90-95]. To analyze the performance, the methods are applied to different publicly available cancer and other disease-related gene expression datasets. The major metrics used here for evaluations of the performance of the proposed classifier are the Cross-validation method (LOOCV, 5-fold, and 10-fold), ROC Curve, and Heat map.

Description and Preprocessing of the Datasets

The experimentation has been carried out over ten publicly available different gene expression binary class and multi-class datasets. Among these datasets, eight datasets are cancer datasets and two arthritis datasets. The eight cancer datasets are Leukemia [79], Colon [96], Prostate [97], Lung [98], RBreast [99], Breast [100], MLL [101], and SRBCT [102]. To show the accuracy of the proposed model with respect to other than cancer datasets here two arthritis datasets RAHC [103] and RAOA [103] are also considered.

Comparison of MFSAC-EC Model with Well-Known Existing Gene Selection Methods

In Figure 6, the MFSAC-EC model with different existing classifiers as base classifiers are compared with existing well-known supervised gene selection methods named mRMR

(minimum redundancy maximum relevance framework) [76], MSG (mutual information based supervised gene clustering algorithm) [88], CFS (Correlation-based Feature Selection) [89], and FCBF(Fast Correlation-Based Filter) [89] with respect to different classifiers using 10-fold cross-validation method. From these results, it has been found that the proposed model outperforms in most of the cases.

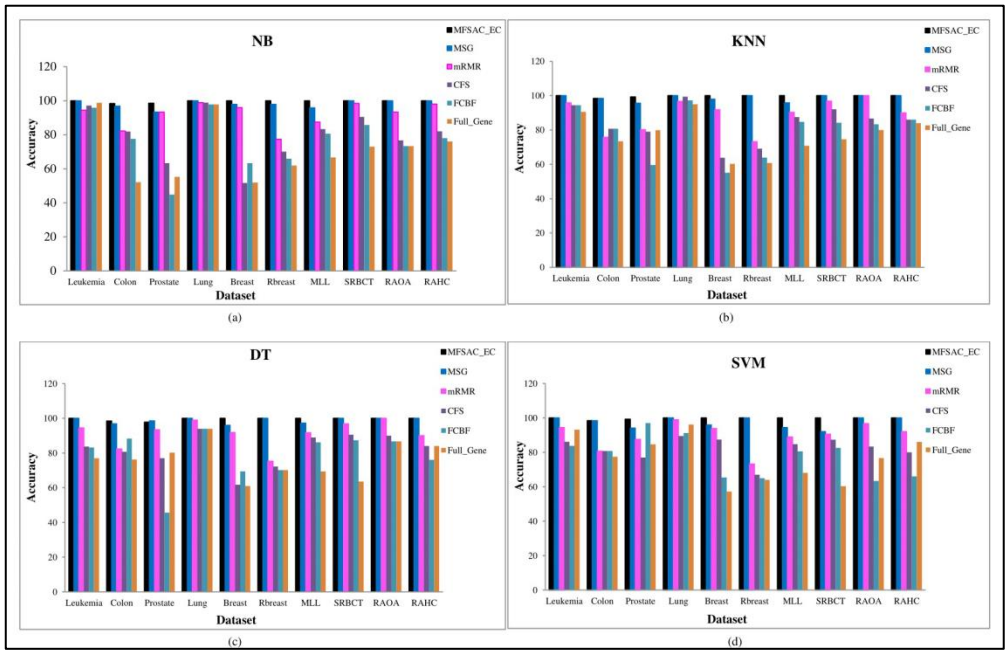

Figure 6. Comparative performance analysis of MFSAC-EC model with respect to different existing gene selection methods in terms of 10 fold cross validation

Comparison of MFSAC-EC Model with Well-Known Existing Ensemble Classification and DEEP learning Models

In the below Tables 6, 7 and 8, MFSAC-EC is compared with different ensemble and deep learning models. In all cases the proposed method shows its superiority.

Biological Significance Analysis

The top 8 genes selected by the MFSAC-EC model for Colon cancer and Leukemia are listed in Table 9. For every gene, the name and symbol of the gene as well as the Accession number of the Affymetrix chip are listed. Apart from this information, to validate those genes, biomedical literature of the genes is searched and for every gene, the corresponding reference about its role and significance for a particular disease is provided.

Conclusion

Many machine learning and statistical learning-based classifiers for sample classification already exist in the literature, but these methods are prone to suffer from overfitting due to small sample size problems, class imbalance problems, and the curse of the high dimensionality of microarray data. Although some of the existing methods can mitigate these issues to quite an extent, the problems have still not been satisfactorily overcome. Due to this reason, here a novel feature selection-based ensemble classification model named

Table 6. Comparison of MFSAC-EC using DT with different existing Ensemble Classifiers using DT in terms of 10-Fold Cross Validation

Dataset	**MFSAC -EC**	**PCA-based RotBoost**	**ICA-based RotBoost**	**AdaBoost**	**Bagging**	**Arcing**	**Rotation Forest**	**EN-NEW1**	**EN-NEW2**
Colon	**98.39**	95.48	96.1	94.97	94.92	69.35	95.21	79.03	83.87
Leukemia	**100**	98.75	98.77	98.22	97.47	Not Found	97.97	Not Found	Not Found
Breast	**100**	94.39	97.88	98.89	92.74	80.41	98.6	94.85	95.88
Lung	**100**	98.11	99.54	96.3	97.08	97.24	97.56	98.34	99.45
Prostate	**97.79**	Not Found	Not Found	90.44	94.12	87.5	Not Found	94.85	97.06
MLL	**100**	98.86	99.31	97.63	97.11	91.67	97.61	93.06	98.61
SRBCT	**100**	99.5	99.59	98.16	96.46	Not Found	97.44	Not Found	Not Found

Table 7. Comparison of MFSAC-EC using DT, KNN, NB, SVM with different existing Ensemble Classifiers using DT, KNN, NB, SVM in terms of 10-Fold Cross Validation

Dataset	MFSAC-EC				Bagging			Boosting			Stacking			HBSA		SD_Ens	Meta_Ens
	DT	NB	KNN	SVM	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	KNN	SVM		
Leukemia	**100**	**100**	**100**	**100**	94.12	88.23	73.53	91.18	88.24	75.53	91.18	91.18	91.18	88.46	88.46	92.45	94.12
Colon	98.39	98.39	**100**	98.39	95.16	66.13	90.32	98.39	87.1	91.94	98.39	93.59	93.59	75	85	94.4	99.21
Prostate	97.79	**99.26**	**99.26**	**99.26**	26.47	26.47	38.24	26.47	26.47	52.94	26.47	26.47	52.94	85.29	97.06	52.94	52.94
Lung	**100**	**100**	**100**	**100**	91.28	96.64	97.32	81.88	95.3	97.99	97.99	97.99	96.64	Not Found	Not Found	81.88	97.99
Breast	**100**	**100**	**100**	**100**	78.95	36.84	68.42	68.42	36.84	68.42	68.42	68.42	68.42	Not Found	Not Found	73.49	79.87

Table 8. Comparison of MFSAC-EC using SVM and KNN with respect to different existing deep learning Classifiers using random splitting

Dataset	SVM			KNN		
	MFSAC-EC	Folded Autoencoder	Autoencoder	MFSAC-EC	Folded Autoencoder	Autoencoder
Colon	**100**	90.15	73.11	**98.39**	81.09	56.97
Prostate	**96.81**	84.16	64.3	**97.87**	76.48	52.1
Leukemia	**100**	93.62	84.12	**100**	85.24	77.13

Table 9. List of genes selected by MFSAC-EC model for Colon and Leukemia cancer Datasets

Dataset	Gene Name	Accession Number	Description	Validation of Genes
Colon	TPM1	Hsa.1130	Human tropomyosin isoform mRNA, complete cds.	[103], [104], [105]
	IGFBP4	Hsa.1532	Human insulin-like growth factor binding protein-4 (IGFBP4) gene, promoter and complete cds.	[106], [107], [108]
	MYL9	Hsa.1832	Myosin Regulatory Light Chain 2, Smooth Muscle Isoform (Human); contains element TAR1 repetitive element	[109], [110]
	ALDH1L1	Hsa.10224	Aldehyde Dehydrogenase, Mitochodrial X Precursor (Homo sapiens)	[111], [112]
	KLF9	Hsa.41338	Human mRNA for GC box binding protein/ Kruppel Like Factor 9, complete cds	[113], [114], [115]
	MEF2C	Hsa.5226	Myocyte-Specific Enhancer Factor 2, Isoform MEF2 (Homosapiens)	[116], [117], [118]
	GADPH	Hsa.1447	Glyceraldehyde 3-Phosphate Dehydrogenase	[119], [120]
	TIMP3	Hsa.11582	Metalloproteinase Inhibitor 3 Precursor	[121], [122]
Leukemia	TXN	X77584_at	TXN Thioredoxin	[123], [124], [125]
	CSF3R	M59820_at	CSF3R Colony stimulating factor 3 receptor (granulocyte)	[126], [127], [128], [129]
	MPO	M19508_xpt3_s_at	MPO from Human myeloperoxidase gene	[130], [131], [132], [133]
	LYZ	M21119_s_at	LYZ Lysozyme	[134], [135]
	CST3	M27891_at	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	[136]
	ZYX	X95735_at	Zyxin	[136], [137]
	CTSD	M63138_at	CTSD Cathepsin D (lysosomalaspartyl protease)	[134]
	CD79A/ MB-1 gene	U05259_rna1_at	MB-1 membrane glycoprotein	[134]

MFSAC-EC is proposed. It has been shown that the proposed model can handle the above mentioned issues present in existing models. From the experimental results, it has been found that the proposed model outperforms all other well-known existing classification models combined with the different recognized feature selection methods and also the newly developed ensemble classifiers for all types of cancer datasets mentioned here. Apart from this classification task, the proposed model can also rank informative attributes according to their importance. The efficiency of the proposed model in this task is vindicated by finding the most informative genes for Colon cancer and Leukemia cancer datasets using this model.

3. Conclusion and Future Directions

The main objective of this thesis is to develop machine learning based some classification and clustering methodologies, which preprocess and analyze gene expresssion data more accurately. In this regard, certain problems of gene expression data and the solutions of these problems using the proposed methodologies are discussed in this thesis.

In the first work we have developed a novel framework for better neighbourhood formation in KNNimpute and its several versions to improve their prediction accuracy. From experimental results it has been found that prediction accuracy of the *K*NN and its several versions has been greatly improved after using this framework. This method can be applied in RNA expression data, protein expression data for prediction of missing values in future.

In the second work we have developed another imputation method via integrating clustering and numerical method as it is already known that numerical methods are robust although these methods has several limitations. In this work we have tried to overcome all these drawbacks. This method can be applied in RNA expression data, protein expression

data and also in other areas apart from bioinformatics for prediction of missing values in future.

In the third work, we have proposed a new framework based on partition based clustering for groupingin unsupervised manner functionally similar genes from microarray gene expression data. This is the first framework where we have eliminated noise using different partition based clustering methods and tries to overcome the drawbacks of different partition based clustering methods. Experimental results show superioty of the proposed method. One limitation of this method is that to check gene gene similarity we have used Euclidean distance as in gene expression data it is already known that not value-wise only pattern based similar and expression value wise closer genes are also functionally similar. Using Euclidean distance value wise closer genes can only be considered. In future work we will solve this problem.

In the fourth work we have proposed a new ensemble classification model named MFSAC-EC for sample classification in microarray gene expression data. The proposed model has two components. One is gene/feature selection to reduce feature dimenstion and second one sample classification. For feature selection we have applied multi filters based supervised attribute/gene clustering algorithm and for classification we have applied a modified bagging model. The proposed model shows its superioty for different cancer datasets. In future we will modify this model via applying deep learning techniques and apply it for other disease based microarray datasets.

Finally, it can be concluded that different classification and clustering schemes reported in this thesis can be extended to model other complex problems of bioinformatics and data mining.

Reference

1. W J S Diniz,F Canduri,"REVIEW-ARTICLE Bioinformatics: an overview and its applications", Genetics and molecular research: GMR16(1), March 2017.DOI:10.4238/gmr16019645
2. N.M. Luscombe, D. Greenbaum, M. Gerstein, "What is bioinformatics? An introduction and overview", Yearbook of Medical Informatics 2001.
3. P. Baldi and S. Brunak. Bioinformatics: The Machine Learning Approach. Cambridge, MA: MIT Press, 1998.
4. Iqbal H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions", SN Computer Science, March, 2021.
5. Shreena Angra, SAchinh Ahuja, "Machine learning and its applications: A review", 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC).
6. Ayon Dey. Machine Learning Algorithms: A Review. International Journal of Computer Science and Information Technologies, Vol. 7 (3) , 2016, 1174-1179
7. Robert P. W. Duin. SPR, 2002. Structural, Syntactic, and Statistical Pattern Recognition. Joint IAPR International Workshops SSPR 2002 and SPR, 2002.
8. Batyrkhan Omarov, Young Im Cho, "Machine learning based pattern recognition and classification framework development", 2017 17th International Conference on Control, Automation and Systems (ICCAS), 2017.
9. A.W.Olthof,P.Shouche, F.F.A.IJpma, R.H.C.Koolstra, V.M.A.Stirler, P.M.A.vanOoijen, L.J.Cornelissen, "Machine learning based natural language processing of radiology reports in orthopaedictrauma",Computer Methods and Programs in Biomedicine, 8, 2021.

10. Ayush Pratap, Neha Sardana, "Machine learning-based image processing in materials science and engineering: A review", Materials today Proceedings, 2022.
11. Shuja Mirza, Dr.Sonu Mittal, Dr.Majid Zaman, "A Review of Data Mining Literature", International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 11, November 2016.
12. Jiawei Han, M. Kamber, "Data Mining Concepts and Techniques", Elsevier, 2000.
13. H. Causton, J. Quackenbush, and A. Brazma, Microarray Gene Expression Data Analysis: A Beginner's Guide. Wiley-Blackwell, 2003.
14. Zhang A. Advanced analysis of gene expression microarray data. Singapore: World Scientific; 2006.
15. Selvaraj, S. &Natarajan, J. "Microarray data analysis and mining tools". *Bioinformation, 6* (3), 95, 2011 .
16. A. Brazma and J. Vilo, "Minireview: Gene Expression Data Analysis," Federation of European Biochemical Societies Letters, vol. 480, no. 1, pp. 17-24, 2000.
17. D. Jiang, C. Tang, and A. Zhang, "Cluster Analysis for Gene Expression Data: A Survey," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 11, pp. 1370-1386, Nov. 2004.
18. E. Domany, "Cluster Analysis of Gene Expression Data," J. Statistical Physics, vol. 110, nos. 3-6, pp. 1117-1139, 2003.
19. Jelili Oyela de, Itunuoluwa Isewon, Funke Oladipupo, Olufemi Aromolaran, Efosa Uwoghiren, Faridah Ameh, Moses Achas, and Ezekiel Adebiyi, "Clustering Algorithms: Their Application to Gene Expression Data", Bioinform. Biol. Insights. 2016; 10: 237–253.
20. S. Tavazoie, D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church, "Systematic Determination of Genetic Network Architecture," Nature Genetics, vol. 22, no. 3, pp. 281-285, 1999.
21. Daniel P Berrar,C Stephen Downes,Werner Dubitzky, "Multiclass cancer classification using gene expression profiling and probabilistic neural networks", PAC SYMP BIOCOMPUTING, 2003.
22. Heping Zhang et al., Cell and tumor classification using gene expression data: Construction of forests", PNAS, 2003.
23. Ramachandro Majji et al., "Jaya Ant lion optimization-driven Deep recurrent neural network for cancer classification using gene expression data", Medical & Biological Engineering & Computing volume 59, pages1005–1021, 2021.
24. Javed Khan et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", Nature Medicine, 2001.
25. Singh, D., et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell, 1* (2), 203–209, 2002.
26. Shedden, K. , et al. (Gene expression–based survival prediction in lung adenocarci- noma: A multi-site, blinded validation study. *Natural Medicine, 14* (8), 822, 2008.
27. John D Hainsworth, Mark S Rubin, David R Spigel, Ralph V Boccia, Samuel Raby, Raven Quinn, F Anthony Greco, "Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: a prospective trial of the Sarah Cannon research institute", Journal of
28. Kalifa Manjang et al., "Prognostic gene expression signatures of breast cancer are lacking a sensible biological meaning", Nature 2021.
29. Luke Kumar et al., "Gene expression based survival prediction for cancer patients—A topic modeling approach", PLoS One. 2019; 14(11): e0224446.
30. Shuang Liu et al., "Identification of Potential Key Genes for Pathogenesis and Prognosis in Prostate Cancer by Integrated Analysis of Gene Expression Profiles and the Cancer Genome Atlas", Frontiers in Oncology, 2020.
31. Liew AW, Law NF, Yan H. Missing value imputation for gene expression data, Computational techniques to recover missing data from available information. Brief Bioinform 2011; 12(5): 498-513.
32. Moorthy K, Mohamad M.S., Deris S. A review on missing value imputation algorithms for microarray gene expression data. CurrBioinform 2014; 9:18-22.
33. Tibshirani, R., Hastie, T., Narasimhan, D., Chu, G.: Diagnosis of multiple cancer types by shrunken centroids of gene expression, Proc. Natl. Acad. Sci. U. S. A. 99, 6567–6572, 2002.
34. Kerr, M.K., Martin, M., Churchill, G.A.:Analysis of variance for gene expression microarray data, J. Comput. Biol. 7, 819–837, 2000.

35. Anindya Bhattacharya et al.: Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles,Bioinformatics 2008, 11, 1359-1366.
36. Das C, Bose S, Chattopadhyay M, Chattopadhyay S. A novel distance based iterative sequential KNN algorithm for estimation of missing values in microarray gene expression data. IJBRA 2016; 12(4): 312-42.
37. Shilpi Bose, Chandra Das, Kuntal Ghosh, Matangini Chattopadhyay, Samiran Chattopadhyay, "A Framework for Neighborhood Configuration to Improve the KNN based Imputation Algorithms on Microarray Gene Expression Data", International Journal of Bioinformatics Research and Applications (IJBRA), Inderscience, 2021.
38. Y. Cheng and G.M. Church, Biclustering of Expression Data. Proc. Eighth Int'l Conf. Intelligent Systems for Molecular Biology (ISMB '00), pp. 93-103, 2000.
39. Jesus S and Aguilar-Ruiz. Shifting and scaling patterns from gene expression data. Bioinformatics 2005, 21:3840-3845.
40. Bras LP, Menezes JC. Improving Cluster-based Missing Value Estimation of DNA Microarray Data. Biomolecular Engineering 2007; 24:273–282.
41. Tuikkala J, Elo LL, Nevalainen OS, Aittokallio T. Missing value imputation improves clustering and interpretation of gene expression microarray data. BMC Bioinformatics 2008; 9: 202. doi: 10.1186/1471-2105-9-202.
42. Kim H, Golub GH, Park H. Missing value Estimation for DNA Microarray Expression Data: Local Least Square Imputation. Bioinformatics 2005; 21: 187–198.
43. Tryosanka O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Bostein D, Altman RB: Missing Value Estimation Methods for DNA Microarrays, Bioinformatics, 2001, 17:520–525.
44. Oba S, Sato MA, Takemasa I, Monden M, Matsubara K, Ishii S. A Bayseian Missing Value Estimation Method for Gene Expression Profile Data. Bioinformatics 2003; 19: 2088-96.
45. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. MolBiol Cell 1998; 9: 3273-97.
46. Baldwin DN, Vanchinathan V, Brown PO, Theriot JA. A gene expression program reflecting the innate immune response of cultured intestinal epithelial cells to infection by Listeria monocytogenes. Genome Biol 2003; 4(1): R2.
47. Yashimoto H, Saltsman K, Gasch A, et al. Genome-wide analysis of gene expression regulated by the calcineurin/Crz1p signaling pathway in Saccharomyces cerevisiae. J BiolChem 2002; 277: 31079-88.
48. Gasch AP, Spellman PT, Kao CM, et al. Genomic expression programs in the response of yeast cells to environmental changes. MolBiol Cell 2000; 11: 4241-57.
49. Ross DT, Scherf U, Eisen MB, etal.Systematic variation in gene expression patterns in human cancer cell Lines. Nat Genet 2000; 24(3): 227-35.
50. Golub TR, Slonim DK, Tomayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999, 286:531–537.
51. Yu T, Peng H, Sun W. Incorporating nonlinear relationships in microarray missing value imputation. IEEE/ACM Trans Comput. Biol. Bioinform. 2011; 8(3): 723-31.
52. Kerr, M.K., Martin, M., Churchill, G.A.:Analysis of variance for gene expression microarray data, J. Comput. Biol. 7, 819–837, 2000.
53. Van den Bulcke T, Van Leemput K, Naudts B, et al. SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinformatics 2006; 7: 43.
54. Chandra Das, Shilpi Bose, Samiran Chattopadhyay, Matangini Chattopadhyay, Alamgir Hossain, "A Bicluster-based Sequential Interpolation Imputation Method for Estimation of Missing Values in Microarray Gene Expression Data", Current Bioinformatics, Bentham Science, 12(2), pp. 118-130, 2017.
55. Kim KY, Kim BJ, Yi GS: Reuse of Imputed Data in Microarray Analysis Increases Imputation Efficiency, BMC Bioinformatics 2004, 5(160).
56. Cheng KO, Law NF, Siu WC. Iterative bicluster-based least square framework for estimation of missing values in microarray gene expression data. Pattern Recog 2012; 45(4): 1281-9.

57. Pradipta Maji and Sushmita Paul, “Rough-Fuzzy Clustering for Grouping Functionally Similar Genes from Microarray Data”, IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 10, NO. 2, MARCH/APRIL 2013.
58. D. Jiang, J. Pei, and A. Zhang, “DHC: A Density-Based Hierarchical Clustering Method for Time-Series Gene Expression Data,” Proc. IEEE Third Int’l Symp. Bioinformatics and BioEng., pp. 393-400, 2003.
59. Bikram Karmakar et al., “Tight clustering for large datasets with an application to gene expression data”, Scientific Reports, Nature Research, volume 9, article number: 3053 (2019).
60. Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*. 2001;17(2):126–36. doi:10.1093/bioinformatics/17.2.126.
61. Lu Y, Lu S, Fotouhi F, Deng Y, Brown SJ. FGKA: A fast genetic k-means clustering algorithm. In: *Proceedings of the* 2004 *ACM Symposium on Applied Computing*. Vol ACM; 2004:622–3.
62. A.P. Gasch and M.B. Eisen, “Exploring the Conditional Coregulation of Yeast Gene Expression through Fuzzy K-Means Clustering,” Genome Biology, vol. 3, no. 11, pp. 1-22, 2002.
63. Fu L, Medico E. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics*. 2007;8(1):1.
64. Nasser S, Alkhaldi R, Vert G. A modified fuzzy k-means clustering using expectation maximization. In: *2006 IEEE International Conference on Fuzzy Systems*. Vol Vancouver, BC, Canada: IEEE; 2006:231–5.
65. Luis Tari et al., “Fuzzy c-means clustering with prior biological knowledge”, Journal of Biomedical Informatics, 2009.
66. Raghu Krishnapuram et al., “A Possibilistic Approach to Clustering”, IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 1, NO. 2, MAY 1993.
67. Shamir R. and Sharan R. Click: A clustering algorithm for gene expression analysis. In In Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB ’00). AAAI*Press.*, 2000.
68. Basel Abu-Jamous et al., “Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data”, Genome Biology, (2018) 19:172.
69. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub, “Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation,” Proc. Nat’l Academy of Sciences USA, vol. 96, no. 6, pp. 2907-2912, 1999.
70. J.P. Rousseeuw, “Silhouettes: A Graphical Aid to the Interpration and Validation of Cluster Analysis,” J. Computational and Applied Math., vol. 20, no. 1, pp. 53-65, 1987.
71. J.C. Bezdek and N.R. Pal, “Some New Indexes for ClusterValidity,” IEEE Trans. System, Man, and Cybernetics, Part B:Cybernetics, vol. 28, no. 3, pp. 301-315, June 1988.
72. E.I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J.M. Cherry, andG. Sherlock, “GO::Term Finder Open Source Software for Accessing Gene Ontology Information and Finding Significantly Enriched Gene Ontology Terms Associated with a List of Genes, ”Bioinformatics, vol. 20, no. 18, pp. 3710-3715, 2004.
73. Tabares-Soto Reinel, Orozco-Arias Simon, Romero-Cano Victor, Segovia Bucheli Vanesa, Rodríguez-Sotelo José Luis , Jiménez-Varón Cristian Felipe. A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. PeerJ Computer Science. 2020.
74. Chin A J, Mirzal A, Haron H, Hamed H N A. 2015. Supervised, Unsupervised and Semi-supervised Feature Selection: A Review on Gene Selection. IEEE Transactions on Computational Biology and Bioinformatics.
75. Dashtban M. and Balafar M. 2017. Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. Genomics.109(2): 91-107.
76. Ding C. and Peng H. 2005. Minimum redundancy feature selection from microarray gene expression data. J. Bioinf. Comput. Biol. 3(2): 185–205.
77. Elyasigomari V, Lee D. A., Screen H. R.C, Shaheed M.H. 2017. Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification. Journal of Biomedical Informatics.

78. Furey T.S., Cristianini N, Duffy N, Bednarski D.W, Schummer M, and Haussler D. 2000. Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. Bioinformatics. 16(10): 906-914.
79. Golub T. R., Slonim D. K., Tamayo P, Huard C, Gaasenbeek M, Mesirov J.P., Coller H, Loh H.L., Downing J. R., Caligiuri M. A., Bloomfield C. D., and Lander E. S.. 1999. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science. 286(5439):531-537.
80. Nada A and Alshamlan H. 2019. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification", IEEE Access.
81. Błaszczyński Jerzy, StefanowskiŁ Jerzy, Idkowiakukasz.2013. Extending Bagging for Imbalanced Data. Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013.
82. Gu Q., Li Z., Han J.. Generalized Fisher Score for Feature Selection. 2011. UAI'11: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence.
83. Zhou Nina and Wang Lipo. 2007. A Modified T-test Feature Selection Method and Its Application on the HapMap Genotype Data. Geno. Prot. Bioinfo.
84. Das C et al. 2019. Comparative Performance Analysis of Different Measures to Select Disease Related Informative Genes from Microarray Gene Expression Data. International Conference on Innovation in Modern Science and Technology (ICIMSAT-2019), Springer.
85. Leung Yukyee and Hung Yeungsam. 2010. A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification", IEEE/ACM Transactions on Computational Biology and Bioinformatics. 7(1).
86. R.O. Duda and P. E. Hart. Pattern Classification and Scene Analysis. New York: Wiley, 1973.
87. Vapnik V. 1995. The Nature of Statistical Learning Theory. New York: Springer-Verlag.
88. Maji P and Das C. 2012. Relevant and Significant Supervised Gene Clusters for Microarray Cancer Classification. IEEE Transactions on Nanobioscience. 11(2).
89. Ruiz R.,Riquelme J.C., Aguilar-Ruiz J.S. 2006. Incremental wrapper-based gene selection from microarray data for cancer classification. Journal of Pattern Recognition. 39(12): 2383–2392.
90. Bolo´ n-Canedo V. et al. 2012. An ensemble of filters and classifiers for microarray data classification. Pattern Recognition.
91. Nagi Sajid et al.. 2013. Classification of microarray cancer data using ensemble approach. Netw Model Anal Health Inform Bioinformatics.
92. Wang Ching Wei. 2006. New Ensemble Machine Learning Method for Classification and Prediction on Gene Expression Data. Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA.
93. Wang Shu-Lin et al.. 2012. Finding minimum gene subsets with heuristic breadth-first search algorithm for robust tumor classification. BMC Bioinformatics.13:178.
94. Osareh Alireza et al. 2013. An Efficient Ensemble Learning Method for Gene Microarray Classification. Hindawi Publishing Corporation BioMed Research International.
95. Alon U, Barkai N, Notterman D. A., Gish K., Ybarra S., Mack D., and Levine A. J. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by Oligonucleotide arrays. Proc. Natl. Acad. Sci. USA. 96(12): 6745–6750.
96. Singh D., Febbo P.G., Ross K., Jackson D. G., Manola J., Ladd C., Tamayo P., Renshaw A. A., Amico A. V. D, Richie J. P. , Lander E. S., Loda M., Kantoff P.W., Golub T.R., and Sellers W.R..2002. Gene expression correlates of clinical prostate cancer behavior. Cancer Research..1: 203–209.
97. Gordon G. J., Jensen R. V., Hsiao L.-L., Gullans S. R., Blumenstock J. E., Ramaswamy S., Richards W. G., Sugarbaker D. J., and Bueno R. 2002. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Research. 62: 4963–4967.
98. Veer L. J. V., Dai H., Vijver M. J. V. D., He Y. D., Hart A. A.M., Mao M., Peterse H. L., Kooy K. v. d., Marton M. J., Witteveen A. T., Schreiber G.J. , Kerkhoven R.M., Roberts C. ,Linsley P. S., Bernards R., and Friend S. H..2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 415:530–536.

99. West M., Blanchette C., Dressman H., Huang E., Ishida S., Spang R., Zuzan H. , Olson J. A., Marks J. R., and Nevins J. R.. 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl. Acad. Sci. USA. 98(20):11 462–11 467.
100. Armstrong S, Staunton J, Silverman L, Pieters R, den Boer M, Minden M, Sallan S, Lander E, Golub T and Korsmeyer S. 2001. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics. 30:41–47.
101. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, et al. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Medicine. 7(6):673–679.
102. Pouw Kraan T.C.T.M. van der, Wijbrandts C.A., Baarsen L.G.M. van, Voskuyl A.E., Rustenburg F., Baggen J.M., Ibrahim S.M., Fero M., Dijkmans B.A.C., Tak P.P., and Verweij C.L. 2007. Rheumatoid Arthritis Subtypes Identified by Genomic Profiling of Peripheral Blood Cells: Assignment of a Type I Interferon Signature in a Subpopulation of Pateints. Annals of the Rheumatic Diseases. 66: 1008-1014.
103. Gardina Paul J et al.2006. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC GENOMICS.
104. Thorsen Kasper et al. 2008. Alternative Splicing in Colon, Bladder, and Prostate Cancer Identified by Exon Array Analysis. Molecular & Cellular Proteomics. 7: 1214-1224.
105. Botchkina Inna L. et al. 2009. Phenotypic Subpopulations of Metastatic Colon Cancer Stem Cells: Genomic Analysis. Cancer Genomics & Proteomics. 6: 19-30.
106. Durai Rajaraman et al. 2007. Role of insulin-like growth factor binding protein-4 in prevention of colon cancer. World Journal of Surgical Oncology. 5:128.
107. Singh Pomila et al. 1994. Episomal Expression of Sense and Antisense Insulin-like Growth Factor (IGF) binding Protein-4 Complementary DNA Alters the Mitogenic Response of a Human Colon Cancer Cell Line (HT-29) by Mechanisms That Are Independent of and Dependent upon IGF-11. Cancer Research. 54: 6563-6570.
108. Yu Herbert et al. 2000. Role of the Insulin-Like Growth Factor Family in Cancer Development and Progression. Journal of the National Cancer Institute. 92(18).
109. Yan Zhi et al. 2012. Identification of candidate colon cancer biomarkers by applying a random forest approach on microarray data. Oncology Reports, SPANDIDOS Publications.
110. Zhu Kongxi et al.. 2019. Long non-coding RNA MBNL1-AS1 regulates proliferation, migration, and invasion of cancer stem cells in colon cancer by interacting with MYL9 via sponging microRNA-412-3p. Clinics and Research in Hepatology and Gastroenterology, ELSEVIER.
111. Feng Hailiang et al. 2018. ALDH1A3 affects colon cancer in vitro proliferation and invasion depending on CXCR4 status. British Journal of Cancer. 118: 224–232.
112. Waals Lizet M. van der et al.2018. ALDH1A1 expression is associated with poor differentiation, 'right-sidedness' and poor survival in human colorectal cancer. PLOS ONE.
113. Brown Adam R.et al.. 2015. Krüppel-like factor 9 (KLF9) prevents colorectal cancer through inhibition of interferon-related signaling. Carcinogenesis. 36(9): 946–955.
114. Ying Mingyao et al. 2014. KLF9 Inhibits Glioblastoma Stemness through Global Transcription Repression and Integrin-α6 Inhibition. Journal for Biochemistry and Molecular Biology.
115. Simmen Frank A et al.. 2008. The Krüppel-like factor 9 (KLF9) network in HEC-1-A endometrial carcinoma cells suggests the carcinogenic potential of dys-regulated KLF9 expression. Reproductive Biology and Endocrinology.
116. Chen Xiao et al. 2017. MEF2 signaling and human diseases.Oncotarget. 8(67).
117. Giorgio Eros Di et al.2018. MEF2 and the tumorigenic process, hic suntleones. BBA - Reviews on Cancer.
118. Su Li et al. 2016. MEF2D Transduces Microenvironment Stimuli to ZEB1 to Promote Epithelial–Mesenchymal Transition and Metastasis in Colorectal Cancer. Molecular and Cellular Pathobiology.
119. Zhang Jin-Ying , ZhangFan, Hong Chao-Qun, Giuliano Armando E., Cui Xiao-Jiang, Zhou Guang-Ji , Zhang Guo-Jun, Cui Yu-Kun. 2015. Critical protein GAPDH and its regulatory mechanisms in cancer cells. Cancer Biol Med 12:10-22.

120. Tang Zhenjie et al.. 2012. Over-expression of GAPDH in human colorectal carcinoma as a preferred target of 3-Bromopyruvate Propyl Ester" Published in final edited form as: J BioenergBiomembr. 44(1): 117–125.
121. Su Chun-Wen et al. 2019. TIMP-3 as a therapeutic target for cancer. Therapeutic Advances in Medical Oncology.
122. Bai YX et al. 2007. Clinicopathologic significance of BAG1 and TIMP3 expression in colon carcinoma. World Journal of Gastroenterology.
123. Kamal Amany M. et al. 2016. Expression of thioredoxin-1 (TXN) and its relation with oxidative DNA damage and treatment outcome in adult AML and ALL: A comparative study. Hematology.
124. Karlenius Therese Christina et al.2010. Thioredoxin and Cancer: A Role for Thioredoxin in all States of Tumor Oxygenation. Cancers (Basel). 2(2): 209–232.
125. Léveillard Thierry et al. 2017. Cell Signaling with Extracellular Thioredoxin and Thioredoxin-Like Proteins: Insight into Their Mechanisms of Action. Oxidative Medicine and Cellular Longivity.
126. Zhang Yang MD et al.2018. CSF3R Mutations are frequently associated with abnormalities of RUNX1, CBFB, CEBPA, and NPM1 genes in acute myeloid leukemia. Cancer.
127. Ritter Malte et al.. 2020. Cooperating, congenital neutropenia–associated Csf3r and Runx1 mutations activate pro-inflammatory signaling and inhibit myeloid differentiation of mouse HSPCs. Annals of Hematology.
128. Klimiankou Maksim et al. 2019. Ultra-Sensitive CSF3R Deep Sequencing in Patients With Severe Congenital Neutropenia. Front. Immunol.
129. Lance Amanda et al. 2020. Altered expression of CSF3R splice variants impacts signal response and is associated with SRSF2 mutations. Leukemia.
130. Szuber Natasha et al. 2018. Chronic neutrophilic leukemia: new science and new diagnostic criteria. Blood Cancer Journal.
131. Kim Yundeok et al. 2012. Myeloperoxidase Expression in Acute Myeloid Leukemia Helps Identifying Patients to Benefit from Transplant. Yonsei Med J. 53(3): 530–536.
132. Lagunas-Rangel Francisco Alejandro et al. 2017. Acute Myeloid Leukemia—Genetic Alterations and Their Clinical Prognosis. Int J Hematol Oncol Stem Cell Res. 11(4): 328–339.
133. Handschuh Luiza et al. 2019. Not Only Mutations Matter: Molecular Picture of Acute Myeloid Leukemia Emerging from Transcriptome Studies. Journal of Oncology.
134. Wang, H. et al. 2013. Dynamic transcriptomes of human myeloid leukemia cells. Genomics. 102:250–256.
135. Tong Dong Ling and Ball Graham R. 2014. Exploration of Leukemia Gene Regulatory Networks Using A Systems Biology Approach. 2014 IEEE International Conference on Bioinformatics and Biomedicine.
136. Austin H Chen, Yin-Wu Tsau and Ching-Heng Lin. 2010. Novel methods to identify biologically relevant genes for leukemia and prostate cancer from gene expression profiles. BMC Genomics 2010.
137. Yunsong Q et al. 2013. Interval-valued analysis for discriminative gene selection and tissue sample classification using microarray data. Genomics.