Stage classification of clear cell renal cancer based on gene expressions

Akshita Sawhney - MT17143

Supervisor: Dr. Debajyoti Bera

Department of Computational Biology
Indraprastha Institute of Information Technology, Delhi

This dissertation is submitted for the degree of
Master of Technology

August 2019

I would like to dedicate this thesis to my family and my mentor …

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

Akshita Sawhney - MT17143
August 2019

Acknowledgements

I sincerely acknowledge and thank my supervisor Dr.Debajyoti Bera (IIIT Delhi) for giving me the opportunity to work on this project and also for his supervision and valuable guidance during the course of this thesis.
I would also like to thank my family and friends for their consistent support, motivation and trust in my abilities.

Abstract

Tumor development is rooted at genetic level with abnormalities in gene expressions as an important biomarker. Clear cell renal cell carcinoma, a histological subtype of Renal cell carcinoma is the one of the most common form of adult kidney cancer. It shows resistance to conventional chemo-therapies and radio therapies, due to which it is important to continue its intrinsic understanding and identify more molecular markers that can improve the diagnosis outcomes. Analysis of gene level variations for insights into cancer detection is a common practice with gene expression data as its basis. Several attempts have been made to use basic statistical measures to identify genes with differential patterns. The aim of this study is to uncover the information hidden beneath gene expressions by a.) exploiting advanced statistical techniques, b.) analysing structural form (gene correlation network) and c.) discovering relevant segments from the distribution of gene expressions using frequent itemset mining. These approaches have been modelled with an idea to reflect upon the unseen aspects of gene expressions and put them to use to achieve better and robust renal cell cancer classification . Average classification accuracy of '79.5' % is reported on unseen test data. Inferred results were mapped back to the literature and evidences validate the relevance of the proposed feature engineering strategies.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Renal Cell Carcinoma(RCC)

Renal Cell Carcinoma make up about 80-85% of kidney cancer, so it is the most common type of kidney cancers. It happens in the tubules inside the nephrons. There are four kinds of RCC : Clear cell, Papillary, Chromophobe, and Collecting duct RCC[46]. Clear Cell RCC(ccRCC) is the most common type and it occurs in about 75 to 85% of cases. It occurs more commonly in men than women by about 50% and the median age for those diagnosed with ccRCC is about 64 years. There are 4 stages in this cancer each defined by the size of the cancer and the area it encompasses. Stage I is when tumors are confined to the kidney and it's size is less than 7cm. Stage II is when the tumor size increases by 7cm , whereas is Stage III the tumor extends to the vessel named IVC or the lymph nodes immediately surround the kidney. In the Stage IV the tumor moves to the adjacent orans or express the lymph nodes that are distant or can even move through the blood vessels distally, making this stage to be the most advanced form of kidney cancer. So staging is the most important because of the implication for the prognosis and treatment of the patients who are diagnosed with RCC. Fig 1.1 shows the stage based survival for patients with RCC.

1.2 Problem Statement

This study is an attempt to propose an effective classification mechanism between early and late stages of renal cell cancer being a gene expression dataset. Identification of distinguishing patterns for cancer stage classification has been a challenging task in research community and the non-linear behaviour amongst thousands of genes makes it even more complex. High dimensionality of the feature set poses computational challenges in solving

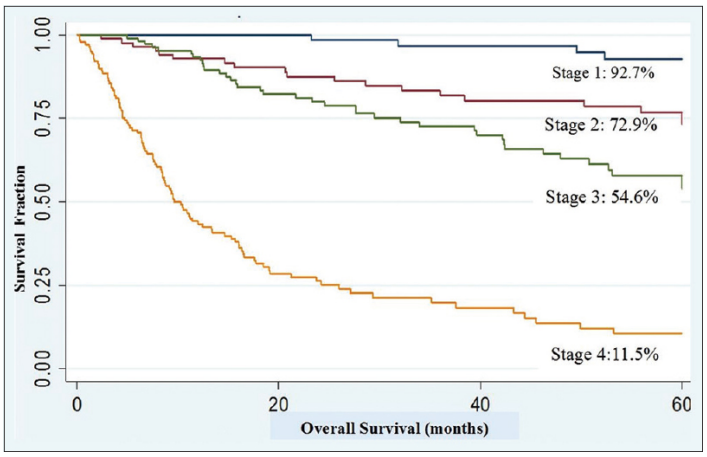


Fig. 1.1 Stage based survival of patients with RCC

such prediction problems. A wide array of statistical and structural techniques were put in place and exhaustive validation was performed in order to justify the findings.

1.3 Outline of the thesis

The body of this thesis is structured around 3 different kinds of feature selection techniques: Rest of this chapter discusses about the origin of the data and its preparation for further analysis. Chapter 2 talks about identifying distinguishing features using different statistical techniques. In Chapter 3 instead of single gene features , groups of genes are identified belonging to specific ranges as item sets. Chapter 4 moves to another non-linear dimension of the feature space looking at various interaction information among the genes to select significant features. Chapter 5 then reports result obtained using the proposed methods , ensemble of all the techniques and an analysis of all the results. Finally, chapter 6 concludes the thesis and provides directions for future research.
Fig 1.2 shows the structure of the thesis:

1.4 Dataset

The gene expression data was extracted from the web resource, CancerCSP [4]. The data provided here was normalized and was divided in 4 parts that are positive training data(early stage), negative(late stage) training data, positive validation data, and negative validation data. It was further transformed with the goal of removing bias from the existing split. The data collected was merged as a whole and randomly shuffled so that the training and validation

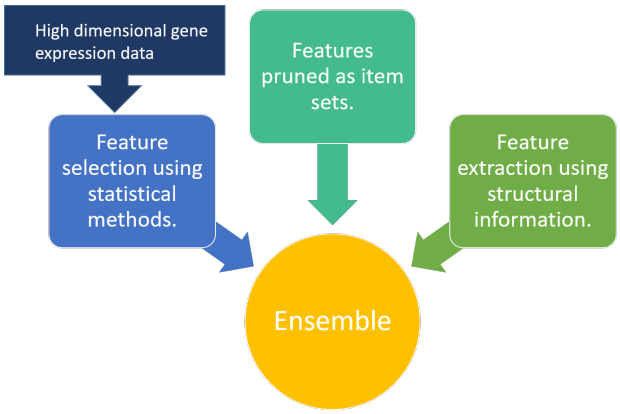


Fig. 1.2 Structure of thesis

splits are representative of the overall distribution of the data. Following this, the data set was split(80:20) into 5 sets (train and test) while making sure that the ratio of patients in early and late stage(55:45) is preserved during training.

1.4.1 Data pre-processing

The pre-processed data set present at CancerCSP was originally obtained from TCGA data portal.[47] with their clinical information in the form of Biospecimen Core Resource (BCR) IDs for patients from the Biotab utility. It contained gene expression values obtained from experiments in the form of RNA-Seq by Expectation Maximization (RSEM)[34] values for 20,531 genes of 523 patients. The data set contained information of patients who had malignant tumors at different stages. Stage I and II constituted early stage whereas stage III and IV collectively meant late stage. Due to a broad spectrum of values, the RSEM values were converted to $\log_2$ scale adding 1 as a constant to each RSEM value. The genes with low variance of 0.025 were removed reducing to 19,166 genes. Following this the $\log_2$ values for each gene were transformed into z scores for normalization[4]. Table 1 shows the formulae that were used for data pre-processing.

Transformed values	formula
log scale	$\mathbf{x} = \log_2(\mathbf{RSEM} + \mathbf{1})$
$\mathbf{Z_{Score}}$	$\mathbf{z} = \frac{\mathbf{x} - \bar{\mathbf{x}}}{\mathbf{std}}$

Table 1.1 Formulae used for data normalization.

1.4.2 Data preparation

The pre-processed data set (z-score transformation) as described in 1.4.1 was transformed using the following steps:

1. Since the data set downloaded from CancerCSP was statically split in train and test, they were combined to form one complete data set which contained 523 patients and 19166 genes. This was done so as to eliminate bias and introduce stochasticity in the splits which would make the discovery of useful patterns easier.

2. After this, it was observed that data was unbalanced in terms of early and late stage. Early stage had 317 patients whereas the late stage had 206 patients.

3. The more frequent class(early) was down-sampled by randomly picking 250 patients and the other class was joined as it is, so as to create a data set which is nearly balanced. This is done 5 times and 5 such data sets were created to ensure that every possible variation is captured and the results are robust.

4. These 5 data sets were then randomly split into stratified train and validation sets(80:20).

1.4.3 Data visualization

Visualization plays an important role in understanding the variance across classes.We first conducted a study to discover differentiating patterns in the distribution of gene expressions, and this study guided us towards some of the genes that could be potentially used for classification. A snapshot of few such genes is shown in Fig. 1.3.

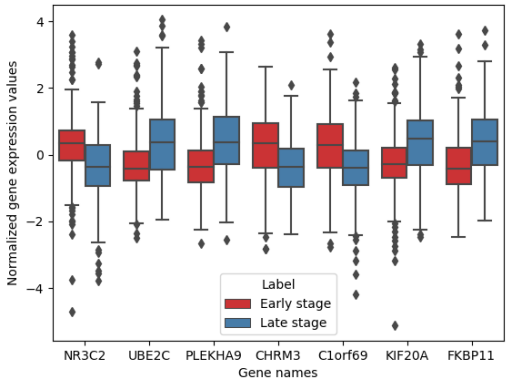


Fig. 1.3 A box plot diagram representation of distribution of a few genes showing differential expression in early and late stage of ccRCC.

1.4.4 Training classifiers on the whole data

Looking at the visualization, it could be observed that the genes expression values in early and late stage had discriminating characteristic . Therefore, we trained the whole data, taking all the genes as it is to check how the data was being classified without doing any feature selection. It was considered as a baseline , to move forward with. Fig1.4 shows the results that were observed, after training various classifiers.

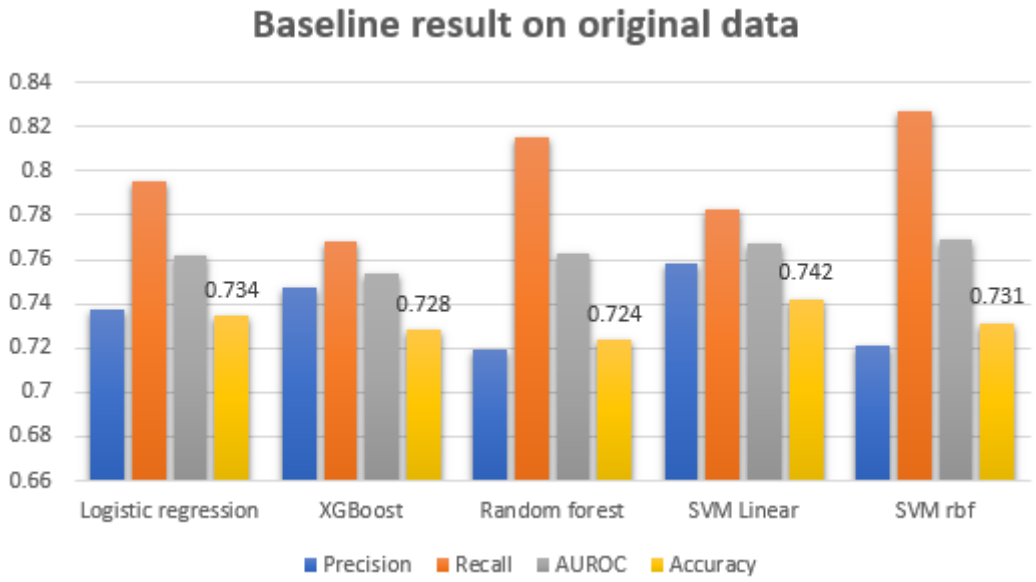


Fig. 1.4 Different metric comparison among 5 classifiers as a baseline

Chapter 2

Feature Selection using statistical methods

Complex machine learning problems involve high dimensional feature engineering and dealing with a large number of features may introduce noise into the system. Some problems of biological origin involve thousands of genes for hundreds of patients and renal cell cancer classification is not an exception. Subjecting 19166 gene expression values corresponding to numerous patients to machine learning environment poses computational challenges and subjects the classifier to noisy data as well as shown in section 1.4.4 . Hence, coming up with a reduced feature set without compromising in terms of information loss is our essential step in a stage prediction pipeline. To achieve this, in this chapter, a battery of basic statistical and complex machine learning strategies were put in place.

Following is the structure of the chapter:

1. Motivation behind using statistical techniques
2. Related Work
3. Feature Selection
4. Training the Classifiers
5. Results and Key findings

2.1 Motivation behind using statistical techniques

Many cancer types are analyzed using gene expression values by identifying discriminating genes becoming potential markers. To identify these markers , many feature selection methods are used. But the problem faced by many methods are that they do not handle extremely high dimensional data very well. The search strategy that is used by them is computationally prohibitive and non effective for further classification. Therefore, the approach that has been taken forward in this chapter, looks more on the statistical inference methods, that effectively decides for each gene whether it can be a contributor for discrimination or is noise. A feature is considered more important if the distribution of the classes are far apart.The major advantage is that it does not follow any iterative process or an elimination technique to decide the importance of the feature.

2.2 Related Work

A bunch of statistical techniques have been used before to identify the relevant features amongst a huge set of genes in the micro-array data. MRMR(minimum redundancy — maximum relevance) approach has been used[11], in which the genes that have less correlation and high mutual information are selected which was high on computation requirement. Statistical significance test like t-test and ANOVA are also used to identify the distinguishing genes[22].B.Chandra and ManishGupta have also described a technique that uses effective ranges of the gene expressions to identify genes that have a smaller overlapping area[5]. This technique was further improved and a more efficient approach was proposed by Wang, Jianzhong, et [51]. A comprehensive review of the approaches used for feature selection was also described by Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga[42]. A threshold techniques also has been experimented to find the top biomarkers from the gene expression data[4].

2.3 Feature Selection

The pipeline used in here is illustrated in Fig 2.1.

Starting with KS test and KL divergence, the complexity of the data was reduced to give a set of genes that could assist in classification, and thereby proposing a mechanism to ensure early detection of renal cell cancer. The elaborate description is as follows:

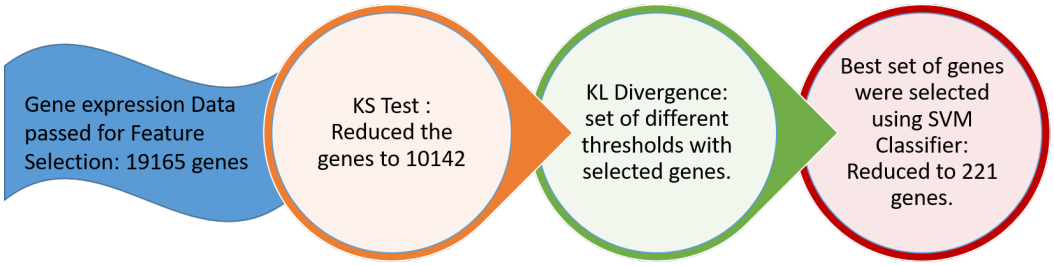


Fig. 2.1 Pipeline for feature selection

2.3.1 KS Test

KS test is a statistical feature selection technique which uses the mean of the cumulative distribution of two samples to test if they come from the same distribution.In a 2 sample ks test the empirical distributions of the two samples are compared with each other. This test was computed on each gene for the two classes: early and late stage. All of them were then sorted on the basis of their KS distance metric(d). The null hypothesis assumed here states that both the distributions are taken from the same continuous distribution. The substitute hypothesis is that they are drawn from different distribution. This test is based on the below relation[31]:

$$\mathbf{KS} = |\mathbf{F_1}(\mathbf{x}) - \mathbf{F_2}(\mathbf{x})| \tag{2.1}$$

where $F_1(x)$ and $F_2(x)$ are the two cumulative distributions for the selected gene for all the patients belonging to early and late stage respectively.

Fig 2.2 shows two distributions, Fig 2.2(left) shows distribution of a gene with low KS metric and low p-value and Fig 2.2(right) shows distribution of a gene with high KS metric and high p value. It can be observed looking at the blue line representing early stage and red line representing late stage that the distributions are far apart.

There are two statistics that explain the result of the KS test that are KS distance metric(d) and p-value. If the d is low or the p-value is high, then we cannot reject the hypothesis that the two class distributions belong to the same distribution.
The genes with a p-value less than 0.05 and d-metric greater that 0.2 were selected. The respective p-value was selected because in majority of hypothesis analysis it is considered that p-value less that 0.05 means that the null hypothesis turns out to be false and vice versa .

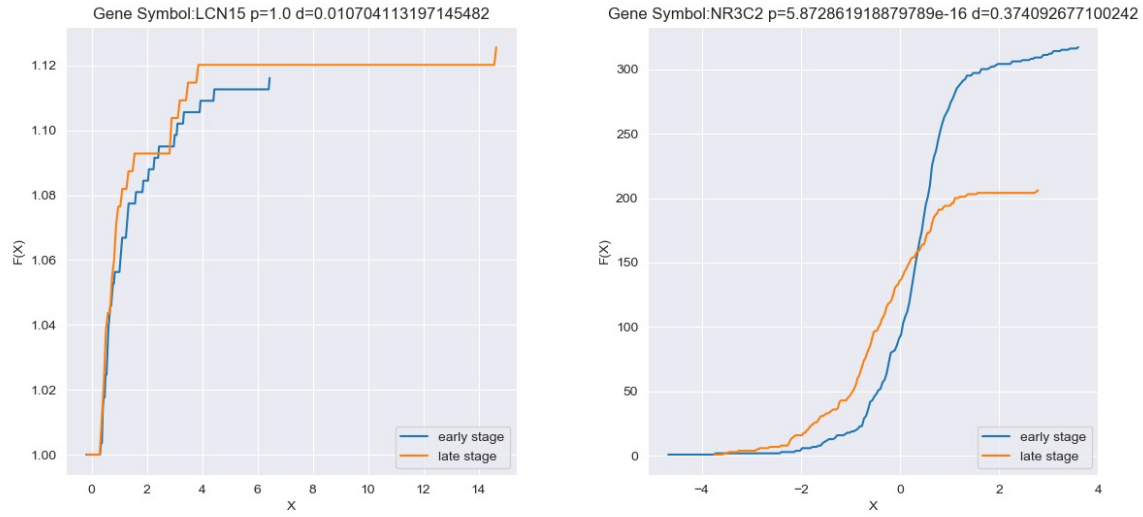


Fig. 2.2 The cumulative distribution functions for two classes (early and late) for a non-significant gene (left) and for a significant gene (right) in Kolmogorov–Smirnov test, where x shows the range of normalized expression value and F(x) shows the cumulative distribution of the gene.

The other cutoff for the 'd' which deals with how far the distributions are from each other was decided in such a way that sufficient genes are left, such that there was no information loss while the noise was removed. The number of genes that were left after applying this test were 10142 out of 19165 genes.

Fig 2.3 shows the plot of the KS metric of all the genes sorted in descending order of the metric. It clearly shows that many genes were there that were fairly discriminating based on their cumulative distributions.

2.3.2 Kullback-Leibler Divergence(KL Divergence)

The common goal is to identify features such that the distance between the 2 classes is maximized so that the discrimination becomes efficient. Another approach that was used to further scale down the dimensions is the Kullback-Leibler Divergence(KL Divergence). KL Divergence uses maximization of probability densities to find the distinguishing features. It is an information theory approach that uses relative entropy to measure the relative dissimilarity between two probability distribution functions(pdfs).

Let the two pdfs be p(x) and q(x) over the random variable X, then the KL Divergence is defined as:

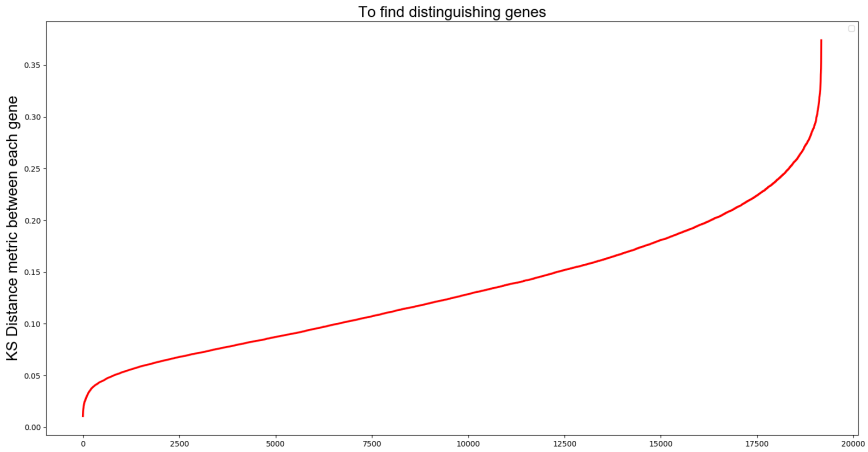


Fig. 2.3 Plot of the KS metric of all the genes sorted in descending order

$$\mathbf{D(p||q)} = \sum_{\mathbf{x}\epsilon\mathbf{X}} \mathbf{p(x)log_2}\frac{\mathbf{p(x)}}{\mathbf{q(x)}} \tag{2.2}$$

where D(p||q) is the KL Divergence. It takes only positive values and and is zero if and only if p=q. It is also non symmetric and does not follow the triangle inequality.

2.3.3 KL Divergence application

To apply Kl divergence to the current dataset, probability distributions of every gene using the expression values of the two class was determined as follows:

1. The range of each gene was binned to obtain 10 bins.
2. The frequency of samples belonging to each bucket was calculated of for each class.
3. This frequency obtained in Step 2 were divided by the total number of patients in that stage, so that the values are now probabilities.
4. On these probability distributions the KL divergence was calculated for each gene
5. A range of thresholds from 0.05 to 0.60 was taken and for each threshold , the genes that were common from the ones in KS test and the ones from this kl divergence threshold were stored to study on them further.

Fig 2.4 shows the distribution of the number of genes for each threshold that was selected. We can observe that with increasing threshold the number of genes are decreasing.

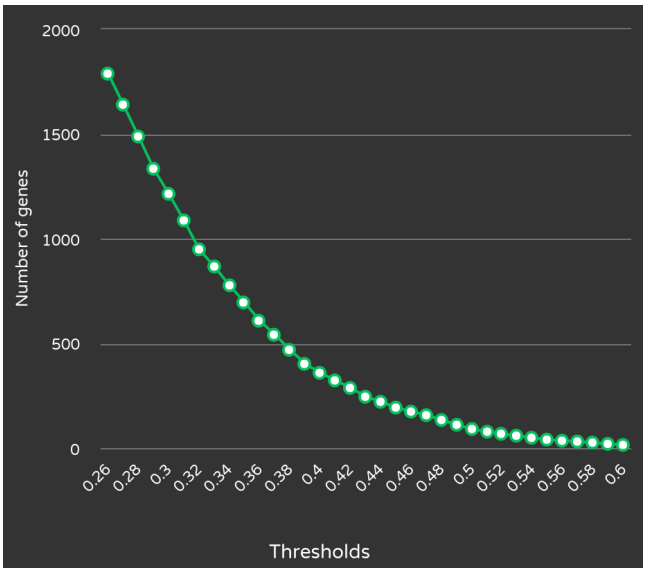


Fig. 2.4 Plot of the threshold vs number of genes

2.3.4 Biological Relevance of using above tests

Selecting distinguished genes from cancer expression dataset has always been a challenge. In this dataset the genes that would serve as prognostic genes from early to late stage were to be diagnosed. According to the biological nature of genes expressing themselves, there is a significant difference in the expression values from one phenotype to another. Thus various statistical methods are used to identify the genes with significant difference between genotypes .The Kolmogorov-Smirnov (K-S) test is a nonparmetric statistical methods used to compare distribution of samples. This method is very sensitive to the difference of the distribution of two sample types. It has been successfully applied in the analysis of ovarian cancer gene data, recognition, and other fields. However, an independent nonparametric test method does not take into account the redundancy of the genes in the selection of genes with discriminatory power[45]. Therefore another test that uses mutual information among the gene named as Kullback-Leibler Divergenc test(KL Divergence test)can efficiently select subsets of genes that are not redundant and do not add noise.

2.3.5 Classification to select the best k

To select the best threshold after KL divergence, machine learning model is used. All the k's(threshold) and a different set of genes respectively are used to transform the 5 split train data and this data is then trained on SVM model. After training, the set of genes giving the maximum average accuracy on the cross-validation test split of train data is selected as the final k. The whole training procedure that was followed is explained in section 2.4 below, which remains alike always.

2.4 Training the Classifiers

A standard machine learning pipeline is illustrated in Fig.2.5.

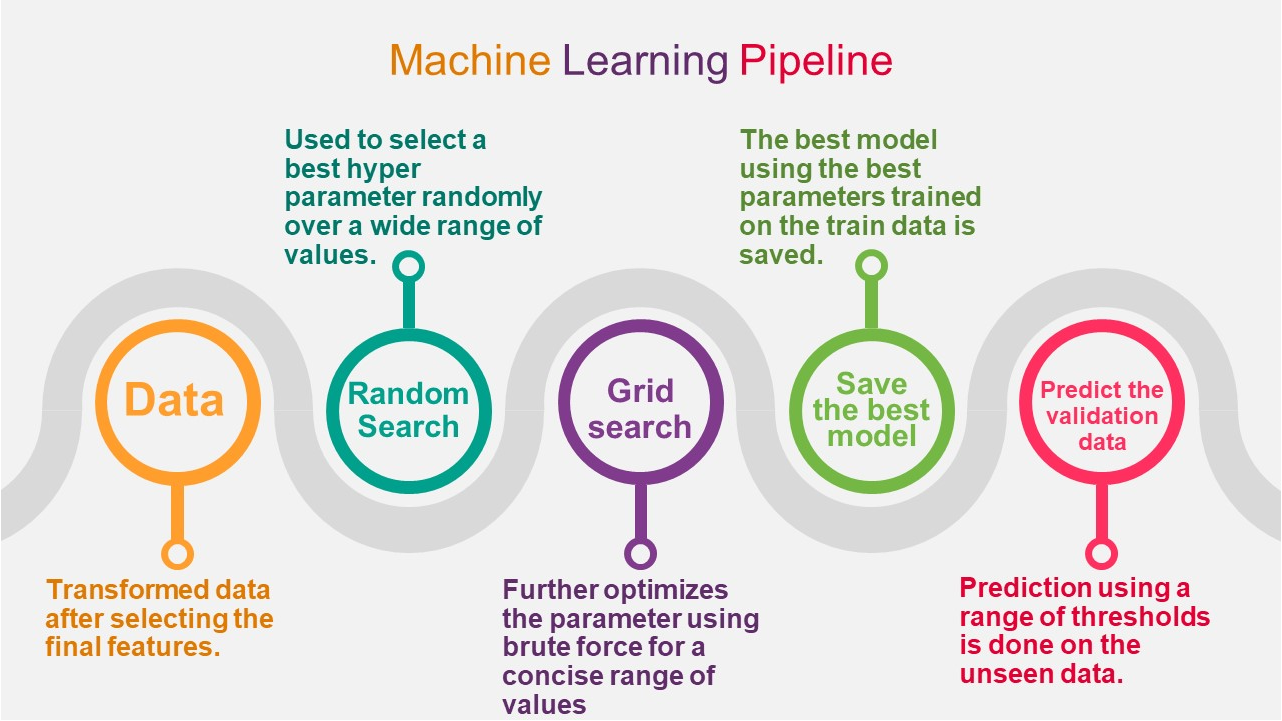


Fig. 2.5 Machine Learning pipeline

The following pipeline was used for training a model:

1. Random Search with the scoring function and cross validation.
2. Grid Search with the scoring function and cross validation.
3. Saving the best trained model.

4. Predicting the validation data(unseen data) and storing the prediction probabilities.

5. Computing all the validation metrics using the threshold method.

2.4.1 Random Search and Grid Search with cross validation

Apart from the parameters that are learned during the training of the model , there are some hyper-parameters(HPs) of every model that are to be learned to train the data and the ones that performs best on the test data are selected.
Grid search is a well known way of finding the optimized value of HP by exhaustively trying out all the combinations available. Even though a very thorough testing is done but with an increasing values of combination the computation cost increases. This limitation prevents from getting to the region of the parameter that would give the highest accuracy. This is where Random search comes into picture. In random search the HPs are sampled from a huge distribution that can cover a big amount of the HP space. It randomly chooses the combinations of HPs covering the whole space. Getting a hold of this space then Grid search can be exhaustively applied to reach to the most optimized HP set.
To make sure that the Hps that are selected are appropriate without any bias, either toward a class or the split of the training data itself, cross validation on the train dataset is done. The folds that have been chosen for our training is 8, this is because the number of instances in the train dataset were not much and selecting 8 as the k fold value would leave sufficient variation instances in train split of the folds. The folds are also stratified so that there is a balance in the the two classes. A limitation common to both these techniques is that the Hps that are chosen can be the parameters that are highly specific to the training data (overfitted), rather than giving a generic best estimator.To make sure that the HP that was being selected is not an overfitted parameter , a new scoring function was introduced that is explained in 2.4.2 .

2.4.2 Scoring function

A good fit of a model refers to how well the target function has been approximated. Overfitting happens when the models is trained extremely well on the train data. This means that the model learns each minute detail and also the noise in the training data to the extend that it impacts the performance of the model negatively whereas the purpose of training the model is to keep it generic . To make sure that the model that is being trained is not overfitted, the difference in train and the test accuracy should be less.

Therefore, a scoring function was made to make sure that the hyper-parameters that are selected are such that they give maximum accuracy on the test split and there is not much difference in the train and the test accuracy. The formula used in the scoring function is shown below

$$\textbf{acc\_diff} = (\textbf{train\_acc} - \textbf{test\_acc}) * (-\textbf{1})$$
$$\textbf{score} = \textbf{wdiff} * \textbf{acc\_diff} + \textbf{wtest} * \textbf{test\_acc}$$

It shows that the difference in the train and test accuracy(acc_diff) is made negative. This was done to make sure that more the difference in the train and test accuracy(overfitting) more negative will be this term, leading to decrease in the overall score. Whereas the test accuracy should be higher, for which it is positive , leading to a higher score for a higher test accuracy. In the formula their are two other terms that are wdiff and wtest, these were weights given to the acc_diff and the test accuracy .If by giving equal weights, a larger difference in the train and test accuracy is observed then wdiff was increased else if the difference is not high but the test accuracy is on a lower side then the wtest can be higher. These weights decide the importance that needs to be given to either the problem of overfitting or the test accuracy.

Fig 2.6 shows the distribution of the scores for each dataset when the scoring function was added in grid search for each classifier. Also on the right , we can see the average cross validated train and test accuracy of the 5 splits for the same classifiers. Looking at both the graph we can observe that the difference in the train and the test accuracy is significantly less and also the Logistic Regression that shows best test accuracy have a higher set of scores compared to the rest.

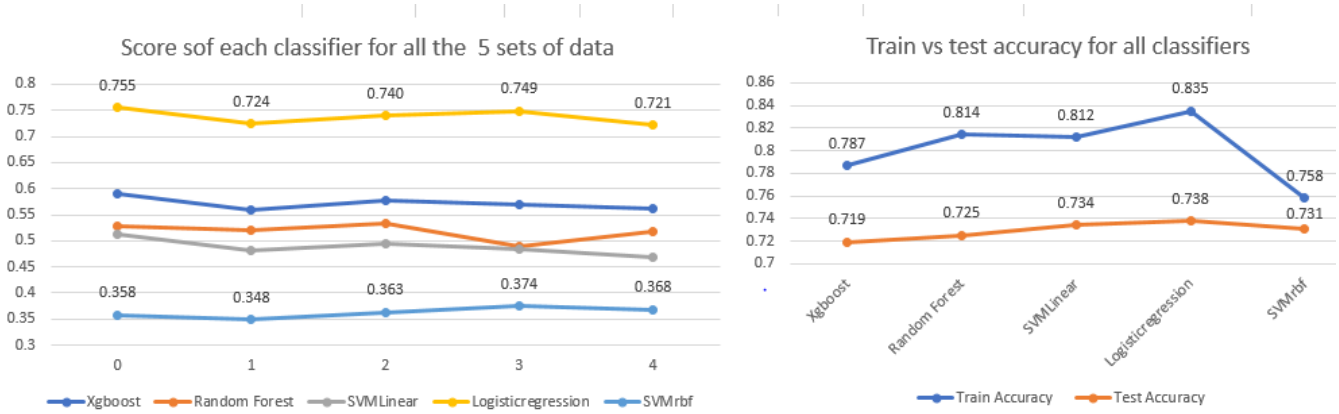


Fig. 2.6 (Left) Distribution of the scores for all the classifiers. (Right) Cross-validated Train and test accuracy averaged for the 5 split of datasets for all the classifiers

2.4.3 Saving the best trained model

Once the appropriate Hps were obtained using the above mentioned search techniques and a new scoring function, the best trained models were then stored.

2.4.4 Predicting the test

The trained classifier was then used to make predictions for the unseen test dataset. The predicted probabilities of the test data is stored separately.

2.4.5 Computing all the validation metrics using the threshold method

When a direct prediction is done on the test data, the prediction probabilities use 0.5 as the threshold to predict the samples. In the current pipeline, a range of thresholds are used to give separate predictions for each threshold.The predicted labels were then compared against the true labels and accuracy was computed. Similarly all the other metrics are computed for each predictions. The metrics that were calculated and reported can be referred to in the Results chapter

2.5 Models selected for training

2.5.0.1 Support Vector Machine(SVM)

A Support Vector Machine is a classifier that uses a hyperplane to separate different classes. In two dimensional space this is line separating the plane in two parts. Whenever the data cannot be separated by a line, then SVM uses the concept of kernels, where it transforms the data to higher dimensions and separate it using the hyperplane. To separate the classes there can be many hyperplanes that can be used, but the aim of SVM is to choose the one that has maximum margin, that is, the maximum distance from the support vectors(Fig 2.7) . Support Vectors are data points that are closest to to the hyperplane and are a major influence on the orientation of the hyperplane. This distance can lead to a tradeoff between allowing miss classifications to the model or making it more generic. A tuning Parameter C is introduced that defines the violation of the margin allowed.

2.5.0.2 Logistic Regression(LR)

Logistic Regression is a statistical technique that is used for classification. It models the probabilities for binary classification problems. It is an extension to the linear regression

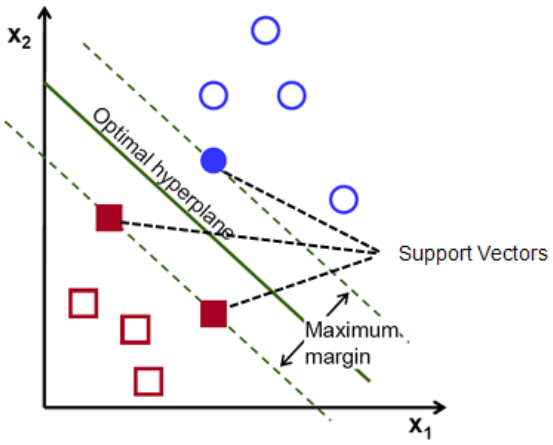


Fig. 2.7 Support Vector Machine classifier with a optimal hyper-plane separating the two classes.

models for classification purposes. The problem that is faced if linear regression is used for classification is that it does not have bounds, which means it does not provide probabilities to predict the classes but rather just a linear interpolation between points. Therefore a solution to this problem comes with the introduction of the logistic function. This function is used to squeeze the output of the linear equation between 0 and 1 . The logistic function is a sigmoid function defined as:

$$\textbf{logistic}(\eta) = \frac{\mathbf{1}}{\mathbf{1}+\exp(-\eta)} \tag{2.3}$$

Hence, to make the linear equation give outputs between 0 and 1, it is wrapped in the logistic function.

2.5.0.3 eXteme Gradient Boosting(Xgboost)[6]

XgBoost is an algorithm that has lately been a dominating algorithm in machine learning. It is a boosted algorithm for decision trees to provide better performance and speed. Gradient boosting is a technique used in the field of machine learning where an ensemble of weak classifiers is done to improve the performance. The ensemble works by giving weight to all the weak classifiers , adding more weight to the difficulty of classifying. Each time it tarins a classifier it uses the weights of the previous classifier, again giving more weights to the next set of misclassifications. Predictions are then made by majority vote of the weak learners' predictions, weighted by their individual accuracy. XgBoost uses the ensemble

of decision trees model that consist of a few regression and classification tress(CART). In CART, the leaves of the tree not only contains the decison values but also have real scores associated with the leaves, giving a better interpretation. It is fast because it provides parallel tree boosting. It is now available as a software library in many languages and can be easily installed on any machine.

2.5.0.4 Random Forest(RF)

Random Forest is another classification technique that uses the ensemble of decision trees to give the predictions. Whereas the difference is that it is not a boosting techniques rather a bootstrapping + aggregation technique . It does a random sampling with replacement of

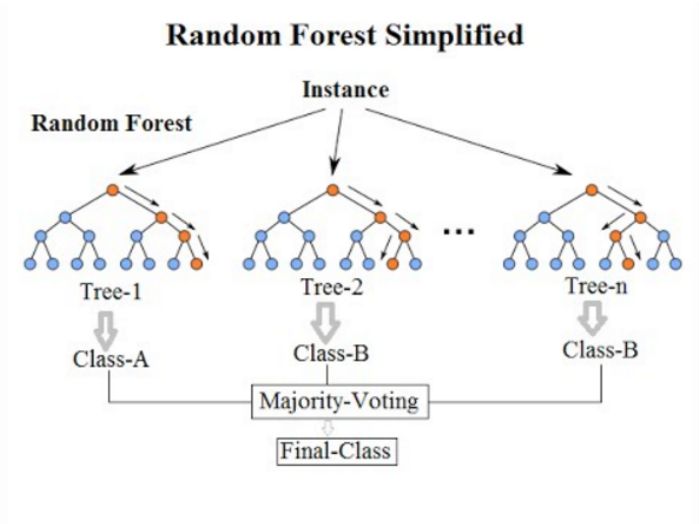


Fig. 2.8 Ensemble of a forest of decision trees: Random Forest

a few data points and trains a decision tree on them. With the hep of this it builds a forest of decision trees and gives the overall prediction by voting (in classification)(Fig2.8) and averaging(regression) all results of the trees that were modeled. So the basic idea behind Random Forest is to use the wisdom of the crowds. Also since decision trees have a tendency to overfit on the data, Random Forest also corrects this tendency.

Chapter 3

Feature Extraction using Item Set Mining

In this chapter we introduce another strategy of extracting a new set of features using item set mining and then using them for classification. The motivation behind using this techniques is to introduce groups of genes as features and considering these genes to be bounded by a specific range in both the stages of cancer. Following is the structure of the chapter.:

1. Motivation behind using itemset mining
2. Related Work
3. Feature Selection
4. Training the Classifiers
5. Results and Key findings

3.1 Motivation behind itemset mining

Cancer is a result of combination of non-linear changes that take place inside the body. As a result of this complicated causal behaviour, genes tend to upregulate / downregulate other gene expressions. It is extremely essential to capture these variations to make sure that early cancer detection becomes a possibility. This led to the idea of looking at a group of genes to be discriminating markers rather that single gene biomarkers. Also it was observed that the potential markers selected from the previous technique had statistically different range of values in the two stages of cancer. Therefore, this observation was also used to capture effective ranges for every gene which could help in identifying certain cuttoffs for each gene

in a group to classify whether a patient is in early or late stage. For example if G1, G2 are identified as a discriminative group, with this it will also say that if a patient P has G1 in range -1 to 0 and G2 in range 1 to 4 then he/she has RCC in early stage.
To identify these groups of co-occuring genes was a motivation to dive into pattern mining techniques. Pattern mining techniques have always been an important part in the domain of bioinformatics. There a variety of such techniques that help in making complex data more feasible to handle and help in identifying patterns that are not very visible to human eye. Frequent itemset mining is a popular group of these pattern mining techniques. The conventional usage of this technique has mainly come into picture during market based analysis, where it is used to identify the items that the customer tends to buy together. Therefore, the shopkeeper then can use these interesting patterns to place the items in such a way that it increases the overall sale. The application of this technique used in bioinformatics can answer many questions like to identify the pattern of genes that co-occur together in a specific stage of cancer. Also to identify discriminative patterns that occur in significantly different frequencies in the two classes. A variety of algorithms have been used to identify these frequent patterns focusing to increase the computational cost and performance[39]. Frequent itemset mining can be an efficient way to capture interesting patterns in complex data and have proven its value in biological data[2],[23].

3.2 Related Work

Feature extraction has always been a challenge for microarray data. Also the features that are selected using different statistical techniques , do not capture the co-occurence of the genes and also miss out on the patterns that can be identified. The application of frequent item set mining is highly applied in gene expression data where it helps in selecting genes that share a common pattern among them. These genes selected have a higher chance of sharing similar biological properties[39]. Researchers have used this technique in making medicinal decisions for frequently occurring diseases[27]. But one of the problems that is faced while using this technique is that a huge number of rules are discovered , which then need to be filtered out making it difficult for biologists to choose appropriate thresholding techniques that can filter out important patterns. There are algorithms like FARMER[7] that are used to reduce the itemsets but still thousands of groups are left. Also another challenge that is faced is to convert the biological data in the form of transactions in such a way that the itemsets that will be extracted makes sense and answer the question that was initially asked. In the approach that is used here, there are two binning approaches used to convert the expression

data into transactions and a more thoughtful technique is used to select the set of items sets that would create a difference in classification.

3.3 Feature Selection

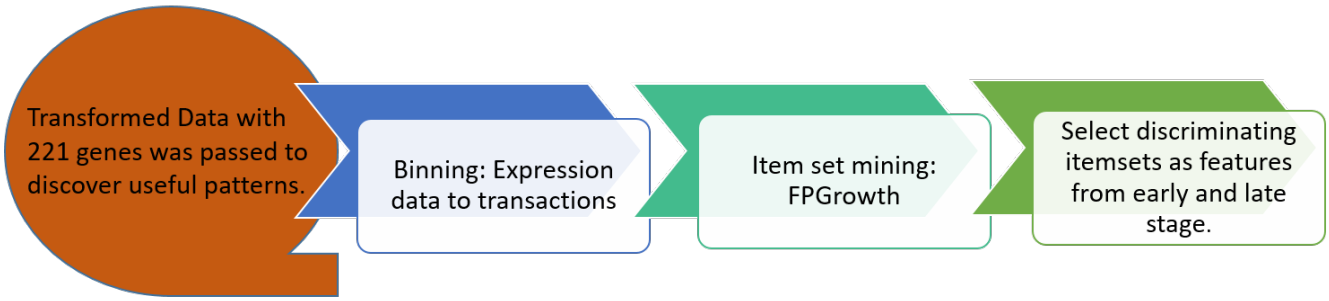


Fig. 3.1 Pipeline for selecting itemsets as features

The pipeline used here is illustrated in Fig.3.1.

3.3.1 Binning techniques to convert the data in the form of transactions

A challenge that is faced when itemset mining is to be applied to gene expression data is to convert the matrix of gene expression values for each patient in the form of transactions. An attempt was made to capture the correlation amongst genes in the form of frequency based item sets. Standard item set mining algorithms took discrete data as input and this was achieved by introducing the concept of bins. A bin is a window from a distribution with a range of values associated with it. For a gene, 'n' bins were obtained, which gave rise to 'n' gene-bin pairs for every gene. Each pair was referred to as an item and a transaction had several such items which could take up binary(0/1) values (indicating whether if the gene expression value for that gene lies in that bucket). The distribution to be bucketed is decided in three ways:

1. Considering the min-max range and fixing the bin count within that range. For example a gene 'PLCL2' had whole range from -3.82 to 4.33 , so the whole range was divided into 10 buckets making each bucket of size: 0.81

PLCL2	Min:-3.825 and Max:4.328	#buckets: 10	Bucketsize: 0.815

Fig. 3.2 Binning 1

2. Fixing the bin size such that different number of bins are obtained for every gene. For example the same gene PLCL2 was divided with a bucket size of 0.432 fixed leading to 19 buckets to be formed.

PLCL2	Min:-3.825 and Max:4.328	#buckets: 19	Bucketsize: 0.42

Fig. 3.3 Binning 2

3. Applying K-means clustering on each gene and each cluster-id was considered as a bin. For example PLCL2 forms 3 clusters , making the bucket number to b 0,1,2.

Important terms used	**Definition**
Itemset	A colletion of one or more items
Support Count	Frequency of an itemset occurring in the transactions
Support Fraction	Fraction of itemsets that contain an itemset $\frac{\textbf{Occurrence}}{\textbf{Total transaction count}}$
Frequent Itemset	An itemset appearing in at least minsup(minimum support) transactions from the transaction database, where minsup is a parameter given by the user

Table 3.1 Terms related to Mining Frequent Itemsets

3.3.2 Apply item set mining

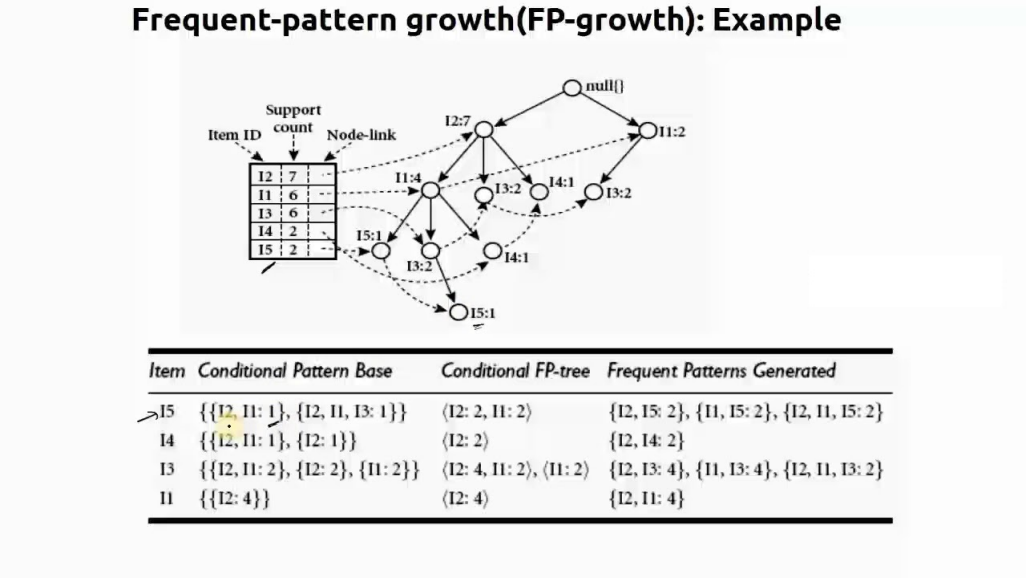
The transaction data now in the form gene name and bucket id as shown in figure 3.2 is then passed through an itemset mining algorithm. The algorithm used here for discovering frequent itemsets is FPGrowth .

Itemsets are discovered for both early and late stage separately with a threshold of 10% so that maximum number of itemsets are generated. FPGrowth is an advanced version of the

```
9PLCL2 5SYNJ2BP 8SEMA3G 18LAMB2 7ANKRD56 8C2orf55 9LRP5 2RCHY1 10DI
5PLCL2 10SYNJ2BP 11SEMA3G 20LAMB2 11ANKRD56 11C2orf55 19LRP5 10RCHY
6PLCL2 3SYNJ2BP 4SEMA3G 15LAMB2 9ANKRD56 3C2orf55 13LRP5 6RCHY1 6DI
10PLCL2 7SYNJ2BP 10SEMA3G 15LAMB2 11ANKRD56 7C2orf55 20LRP5 7RCHY1
9PLCL2 2SYNJ2BP 9SEMA3G 11LAMB2 10ANKRD56 7C2orf55 16LRP5 7RCHY1 8I
2PLCL2 2SYNJ2BP 5SEMA3G 14LAMB2 7ANKRD56 7C2orf55 12LRP5 3RCHY1 5DI
10PLCL2 8SYNJ2BP 12SEMA3G 14LAMB2 13ANKRD56 5C2orf55 18LRP5 9RCHY1
8PLCL2 2SYNJ2BP 8SEMA3G 12LAMB2 11ANKRD56 6C2orf55 16LRP5 7RCHY1 1(
9PLCL2 6SYNJ2BP 10SEMA3G 12LAMB2 10ANKRD56 7C2orf55 15LRP5 9RCHY1
15PLCL2 15SYNJ2BP 9SEMA3G 12LAMB2 14ANKRD56 5C2orf55 12LRP5 12RCHY
6PLCL2 5SYNJ2BP 8SEMA3G 13LAMB2 7ANKRD56 7C2orf55 16LRP5 3RCHY1 6DI
8PLCL2 6SYNJ2BP 10SEMA3G 9LAMB2 10ANKRD56 7C2orf55 14LRP5 9RCHY1 8I
8PLCL2 4SYNJ2BP 10SEMA3G 14LAMB2 9ANKRD56 8C2orf55 15LRP5 6RCHY1 9I
11PLCL2 7SYNJ2BP 8SEMA3G 15LAMB2 13ANKRD56 4C2orf55 18LRP5 10RCHY1
```

Fig. 3.4 Transaction database in the form of gene-name and bucketid

Apriori algorithm[1]. It uses an extended prefix-tree structure for storing compressed and crucial information about frequent patterns, named frequent-pattern tree (FP-tree).



Item	Conditional Pattern Base	Conditional FP-tree	Frequent Patterns Generated
I5	{{I2, I1: 1}, {I2, I1, I3: 1}}	⟨I2: 2, I1: 2⟩	{I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2}
I4	{{I2, I1: 1}, {I2: 1}}	⟨I2: 2⟩	{I2, I4: 2}
I3	{{I2, I1: 2}, {I2: 2}, {I1: 2}}	⟨I2: 4, I1: 2⟩, ⟨I1: 2⟩	{I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2}
I1	{{I2: 4}}	⟨I2: 4⟩	{I2, I1: 4}

Fig. 3.5 Using FP-tree to discover itemsets[24]

SPMF, which is an open-source data mining library[18] was used to run the FPGrowth algorithm. This library is specialized for pattern mining and offers 171 data mining algorithms.

3.3.3 Reduce the number of itemsets and select the important ones

As mentioned before, wisely reducing the large number of itemsets obtained to a number that can be further used for analysis is a challenge. To come up with a solution to this, the itemsets were selected on the basis of their discriminating characteristics between the two

classes. For this, the itemsets from one class were traced among the itemsets of the other class, and the ones that had a percentage difference of a particular threshold, were selected. For example an item set from early stage, 7ZNF55,6NR3C2 where ZNF55 gene belonged to the range of 7th bucket and NR3C2 belonged to 6th bucket had a support of 37.5% in the early stage. this itemset was then searched through the late stage transaction database, with every transaction that had ZNF55 and NR3C2 in their respective bucket was counted, leading to a 15% in the late stage.Therefore, with the threshold difference of 22% in the early and late stage itemsets, this item set was accounted.
This threshold was set in such a way that it was as high as giving a few itemsets. This was done for both the class transactions. This helped in selecting those itemsets that were highly present in one class and not so prominent in the other. Another purpose that was fulfilled was to find such itemsets that belonged to specific ranges of the gene expressions giving a bound to the values.

3.3.4 Prepare the data for classification

Once the itemsets were selected from both the early and the late stage, the gene expression data set is modified.
For each selected feature set, the value is set to a binary 1, if the sample values lie in the same buckets of the genes in the feature else it is set to 0. This is done for all the 5 data splits that were explained in the data preparation above in Section 2.1.2 . The final dataset in the form of binary 0's and 1's is then passed for training the models. Figure shows an example of how the data transformation takes place:

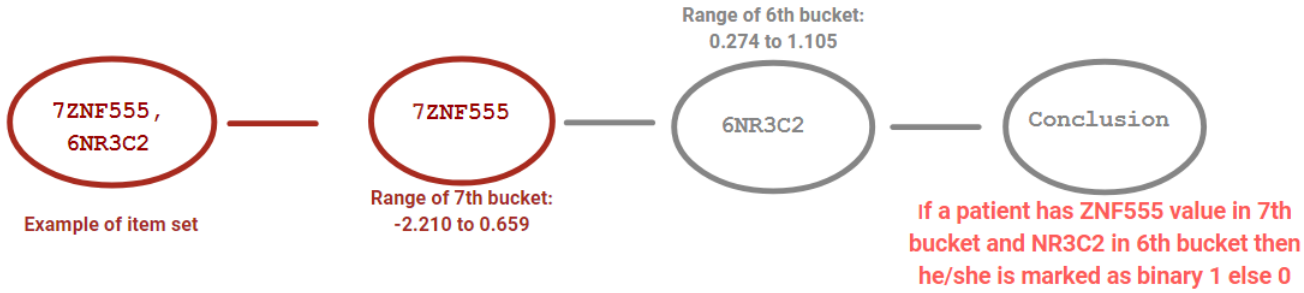


Fig. 3.6 Transfroming dataset

3.4 Training the Classifiers

Refer section 2.4 and 2.5

3.5 Results and Key findings

It was observed that the three kinds of binning techniques played a major role in deciding the right set of item sets. The first binning technique where 10 bins were fixed , gave 51 itemsets that were at about a difference of 20% in both early and late stage. On these 51 itemset various classifiers were trained to observe the classification results on them. It was observed that the maximum accuracy was 75% . The group of genes that were selected contained: '6MUT,7TJP2 : where MUT and TJP2 lied in their 6th(range: -0.32 to 0.14) and 7th(range: 0.12 to 0.34) bucket respectively, 6THRB,8KBTBD4 etc. Another problem that was observed with this technique was that the recall was higher than precision by 15%, indicating that the early stage patients were getting classified better that the late stage.

In the second binning technique where the bin size was same for each gene , leading to different number of buckets in different genes. This deduced 40 set of item sets that were again at a difference of 20% difference in early and late stage. This technique was not able to classify the 2 stages well, the accuracy for classification was 69.8%.

Finally the last technique of clustering had different cluster ids as the bucket ids. It selected 25 set of itemsets which were at a difference of 40% from each other. Classification performed here was better than the previous technique giving a maximum of 73.5% accuracy. Whereas it was also observed that in this technique that precision and recall were close to each other with a difference of 2%.

Chapter 4

Feature selection using structural information

After capturing the features statistically and then with the help of pattern mining, in this chapter, the features are filtered out using the interaction information among them. The gene regulatory network comes into picture. Following is the structure of the chapter.:

1. Motivation behind using structural information
2. Related work
3. Feature Selection
4. Training the Classifiers
5. Predictions

4.1 Motivation behind using structural information

There have been many studies that have put tremendous effort in finding the right set of bio-markers and build more efficient discriminating models. Whereas being inefficient in investigating any information related to the interactions among the genes and their behaviour. Also cancer is known to be a disease related to different pathways, whereas there have not been accurate and abundant exploration on the inter-molecular interactions, the protein interactions or any other physical interactions. There are many reasons for this , one of this being, multiple pathways that do not have a continuous connection but rather a more complex interaction. This all the more increases the need for looking at cancer as a part of a complex

network mechanism and not as identifying single independent markers[16].Therefore, in this chapter we take in account two types of network:

1. External gene regulatory network using BIOGRID[40]
2. A co-expression gene network built using the gene expression matrix

4.1.1 Gene Regulatory networks and Gene co-expression networks

Gene Regulatory networks is a compilation of various gene and gene products interactions like DNA , RNA or their complexes. It helps us understand that genes do work in isolation but rather is a complex network that propel cellular functions. In a gene regulatory network, nodes represent genes and edges represent regulatory relationships.
Gene coexpression network(GCNs) is a set of undirected connections between genes associated with each other using correlations between them. GCNs are a type of gene regulatory network made from a gene expression data by using pair wise correlations between every gene and a significant threshold. Many algorithms also use mutual information between genes to identify the interactions[38].

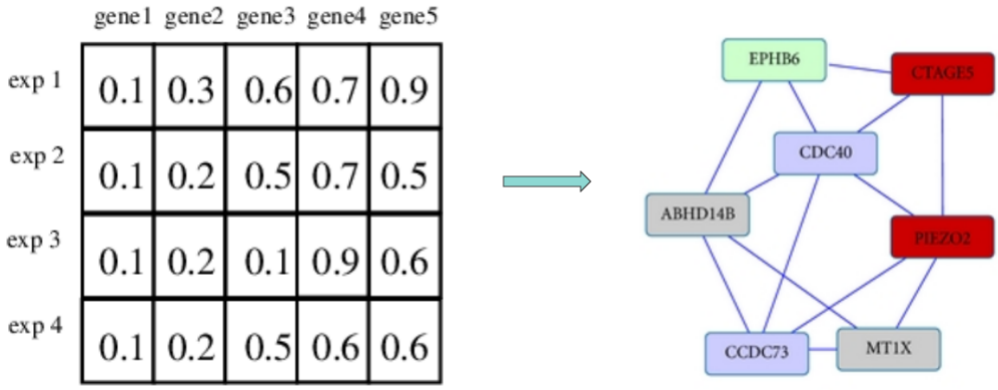


Fig. 4.1 Gene co-expression network

4.2 Related work

There are many research papers that have taken into account GRN's for discovering the complex interaction mechanism between nodes. Yang Yang et al. have used coexpression network to identify the progostic genes in cancer[54]. These networks are also used for gene disease predictions[50]. Gene expression network has been used to identify disease gene biomarkers by identifying gene network base feature set. This has also been further

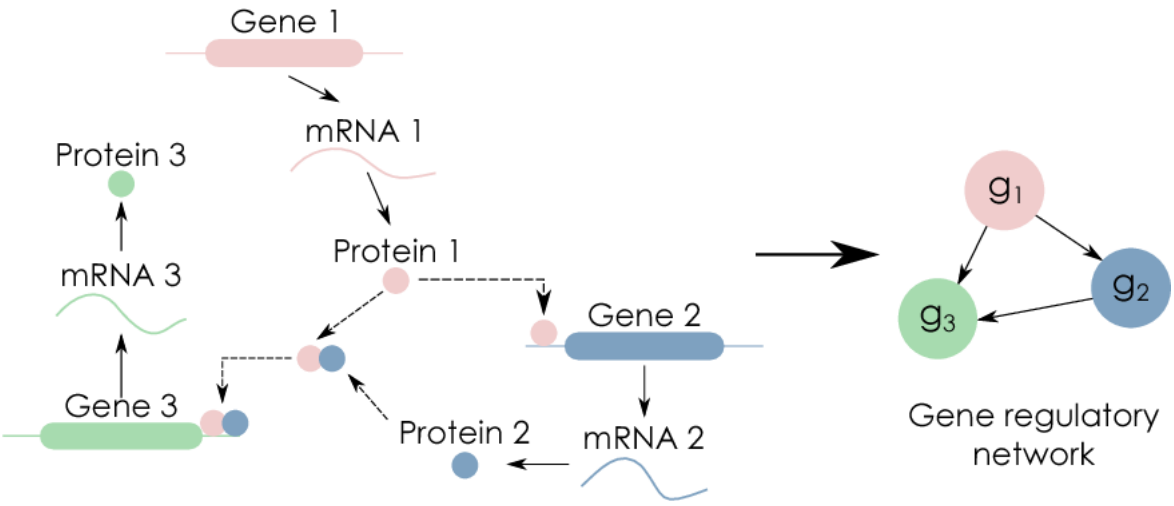


Fig. 4.2 Gene regulatory network

improved by introducing gene subnetwork feature set[13]. Lee et al.[32] and Sootanan et al.[44] proposed methods for identifying genes based on gene-set data to obtain genes which in turn gave high performance for diseases classification.

4.3 Feature Extraction

. The pipeline used here is illustrated in Fig.4.3.

4.3.1 Building a co-expression network

There are many techniques that are used to build co-expression networks, like the most common library used is WGCNA[30] that is built in R. Also there is GeNet[8] that uses adjacency matrix and correlations to build networks.J.S.Dussaut et al. also introduced a tool name GeRNet[14] which incorporate a biclustering tool that finds new relations between the nodes of the network.The problem with most of approaches are that they either suffer from high computational cost, less memory available, error prone or even overfitting. Due to this these techniques are restricted to build networks for small organisms. ARACNE[37] was introduced that overcame these problems, as it introduces a principled and controlled way to introduce the interactions. ARACNE also removes the vast majority of indirect candidate interactions using a well-known information theoretic property, the data processing inequality. This network was a weighted network with the weight values to be the co expression values of the genes.

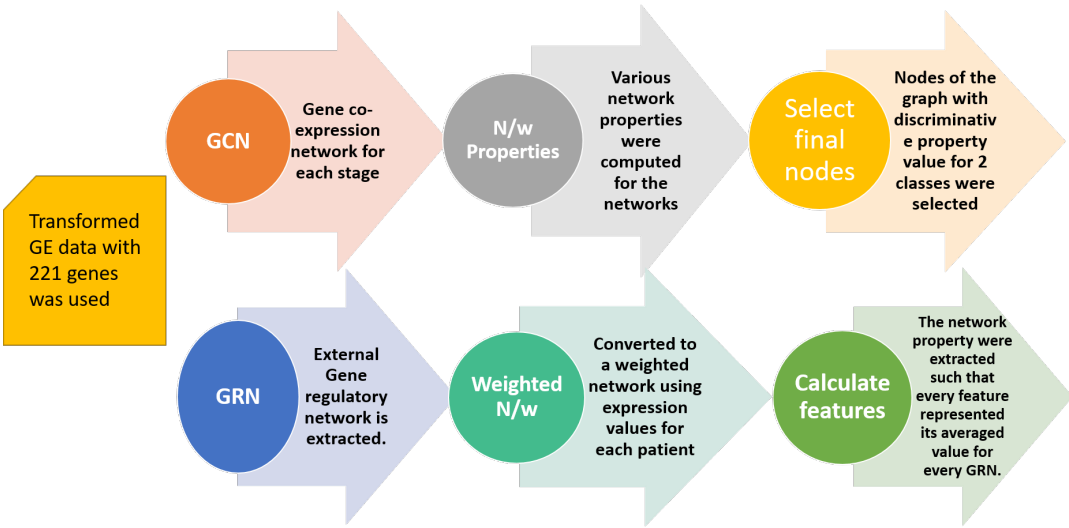


Fig. 4.3 Feature selection pipeline using network features

4.3.2 Extracting an external network

Another network was extracted with the help of the repository of human gene interactions available on BIOGRID. The interactions that contained the genes present among the 19165 genes in RCC were selected. The maximum connected subgraph of the network was extracted from all the components. Since this graph only contained the nodes and the edge connections, gene expressions were used to add weights to the edges. For each patient, a new weighted graph was obtained. These graphs were node weighted graphs, to convert them to edge weighted graphs the following strategy was used:
"For every edge uv, there were now two directed edges, edge u to v having the weight of v and the edge v to u having the weight of u."

Feature selection was done using two methods:

1. Using the co-expression network one for each class
2. Using the external GRN, converted to a weighted network for each patient.

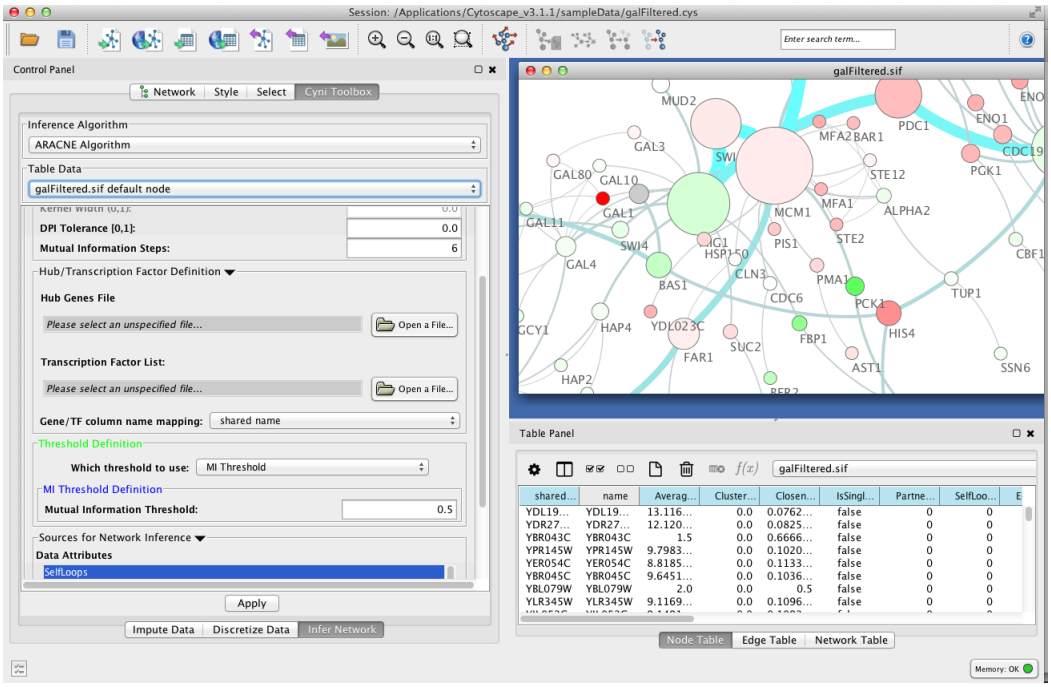


Fig. 4.4 Using ARACNE to make the co-expression network

4.3.3 Using the co-expression network

The features from these expression graphs were extracted as follows:

1. The two networks for early and late stage , were then converted to networkX graphs[9](library of python).

2. For both the graphs their nodes sorted in increasing order of each network property were obtained. The properties that were calculated on these two graphs are explained in the Table 4.1.

3. For each property , the ordered sets of nodes from each class were traversed, and the nodes that were at a specified threshold distance(frequency of nodes in between) apart were selected.

4. Intersection of all the nodes obtained from each property were extracted.

Fig 4.4 shows the distribution of the number of gene with changing value of difference between the two class. It can be observed that with the increasing difference the number of gene decrease.

These made the final set of features obtained using the gene interaction information.

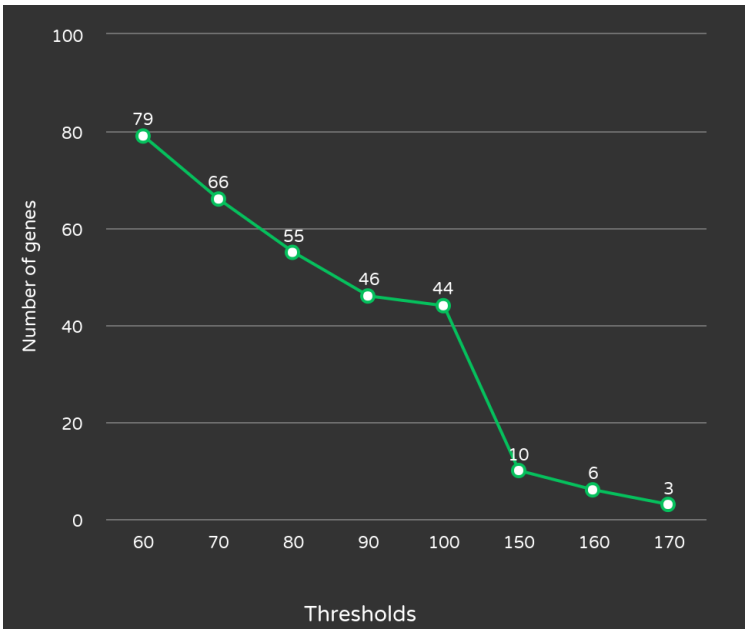


Fig. 4.5 Difference between the two classes vs the number of genes

4.3.4 Using the external GRN

The weighted GRN obtained for each patient was used to obtain network features.The aim was to consider those features which do not have a computational cost , due to to a high number of edges in each network and time constraint. Centrality and clustering based features were used. The list of features that were calculated for each graphs are explained in the Table 4.1. The network features were extracted such that every feature represented the averaged value for every GRN. The new transformed data now formed was a combination of all the features with their average value for each patient. This data was then passed for training a classifier.

4.4 Training the Classifiers

Refer to section 2.4 and 2.5

4.5 Predictions

The trained classifier was then used to make predictions for the unseen test dataset. The predicted labels were then compared against the true labels and accuracy was computed. Accuracy is defined as the percentage of correctly classified samples from the entire test set.

Network Features	Definition	Formula	Method
Weighted Degree Centrality(1 hop)	The fraction of sum of the weights of the nodes it is connected to	$\mathbf{DC} = \frac{\sum \mathbf{w_i}}{\mathbf{n-1}}$,where w_i is the sum of adjacent nodes and n is the total nodes	Both
Closeness Centrality	The reciprocal of the average shortest path distance to a node over all n-1 reachable nodes.	$\mathbf{C(u)} = \frac{\mathbf{n-1}}{\sum_{\mathbf{v=1}}^{\mathbf{n-1}} \mathbf{d(v,u)})}$, where d(v, u) is the shortest-path distance between v and u, and n is the number of nodes that can reach u.	Method 1
Edge Betweeness Centrality	The sum of the fraction of all-pairs shortest paths that pass through that edge.	$\mathbf{c_B(v)} = \sum_{\mathbf{s,t \in V}} \frac{\sigma(\mathbf{s,t|e})}{\sigma(\mathbf{s,t})}$, where V is the set of vertices, $\sigma(s,t)$ is the number of shortest (s,t) paths and $\sigma(s,t|e)$ is the number of those paths passing through e.	Method 1
Current Flow Betweeness Centralities	It uses an electrical current model for information spreading in contrast to betweenness centrality which uses shortest paths.	$\mathbf{C_{CB}(V)} = \frac{\sum \mathbf{I_V^{(st)}}}{\frac{1}{2}\mathbf{n(n-1)})}$, where $\frac{1}{2}n(n-1)$ is a normalizing constant, and Iv(st) is the current flow through node v between source s and sink t.	Method 1
Current Flow Closeness Centralities	It is a variant of closeness centrality based on effective resistance between nodes in a network. This metric is also known as information centrality	$\mathbf{C_{CC}(s)} = \frac{\mathbf{n-1}}{\sum_{\mathbf{s \neq t}} \mathbf{p_{(st)}(s) - p_{(st)}(t)}}$, where $n1$ is a normalizing factor, $p_{(st)}(s)$ is the absolute electrical potential of vertex s based on the electrical current supply from vertex s to vertex t, and $p_{(st)}(s) - p_{(st)}(t)$ corresponds to the effective resistance typically measured as voltage, which can be interpreted as an alternative measure of distance between s and t.	Method 1
Weighted Clustering Coefficient	Defined as the average of the measure of degree to which the nodes of the network tend to cluster together.Also defined as geometric average of the subgraph edge weights	$\mathbf{C_u} = \frac{\mathbf{1}}{\mathbf{deg(u)(deg(u)-1)}} \sum_{\mathbf{vw}} (\widehat{\mathbf{w_{uv}}} \widehat{\mathbf{w_{uw}}} \widehat{\mathbf{w_{vw}}})^{\frac{1}{3}}$, where deg(u) is the degree of u and the edge weights $\widehat{w_{uv}}$ are normalized by the maximum weight in the network $\widehat{w_{uv}} = w_{uv}/max(w)$.	Both
Degree Centrality (2 hops)	Is defined as the degree centrality with the hop radii set to 2, i.e. while computing the centrality, neighbors up to 2 hops are considered for every node.	$\mathbf{DC_2} = \frac{\sum \mathbf{w_i}}{\mathbf{n-1}}$,where w_i is the sum of adjacent nodes with radii 2 and n is the total nodes	Method 1

Table 4.1 Network Based Properties

Chapter 5

Results and Discussion

5.1 Results

After performing all the experiments described in the previous chapters, results obtained with all three methodologies were observed and an ensemble of all the techniques was deduced to observe how each technique built on distinguishing the two stage of cancer using its own set of captured information .

5.1.1 Methodology 1: Feature selection using statistical techniques

In methodology 1, the features under consideration were computed using statistical techniques. Using these a divergent set of features were selected for different values of threshold. SVM classification model was used to select the best set of features that had maximum capability of discriminating the 2 stages by looking at the 8 fold cross validation test accuracy. The scoring measure helped in choosing the most appropriate hyper-parameters leading to train a model that observed best test accuracy without overfitting the model on the train data. The final set of features that were selected consisted of 221 genes. The transformed data was then trained on different training models to observe their performance on these selected genes. The following Table 5.1 and Fig 5.1 show the different observations that were taken:

Metric	Logistic regression	XGBoost	Random forest	SVM linear	SVM rbf
F1- score	0.788	0.765	0.763	0.783	0.768
MCC	0.539	0.476	0.465	0.511	0.476

Table 5.1 F1 score and Matthews correlation coefficient(MCC) values for the 5 classifiers

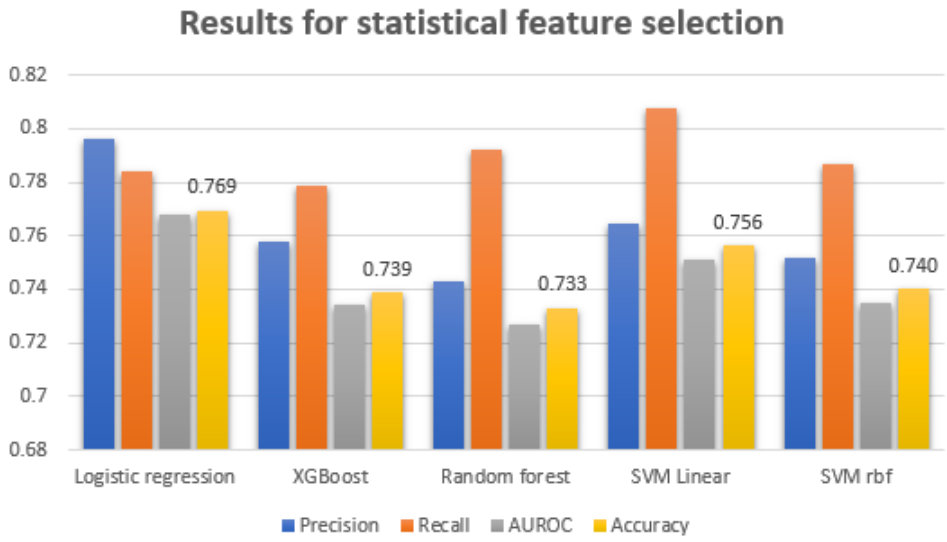


Fig. 5.1 Different metric comparison in among 5 classifiers for methodology 1.

It can be observed that Logistic regression and SVM linear gave maximum accuracy, giving an insight that the data is classified better using linear classifiers. Also further an ensemble of XGboost and Logistic regression was taken so that the final classification could capture all the essential information captured in a tree based classifier and a linear classifier to predict the data better. This gave a new better accuracy of 77.66%.

5.1.2 Biological importance of the genes selected

Functional Enrichment analysis was performed on all the 221 genes that were selected. Fig 5.2, 5.3 and 5.4 show the biological, molecular and cellular comparison for all the annotated genes, where the y axis represents the percentage of genes belonging to a particular process/component. An enrichment p-value is calculated by comparing the observed frequency of an annotation term with the frequency expected by chance; individual terms beyond some cut-off (eg. p-value $<= 0.05$) are deemed enriched[26].

It can be observed that most of the genes were observed to be involved in signal transduction, cell communication and metabolism processes. The molecular functions that most of the genes are involved in are all kinds of transcripitional activities and catalytic activities focusing on the nuclear membrane, cytoplasm and plamsa membrane as the cellular components. The top genes that are deduced from the 221 genes using property of feature importance in random forest, include NR3C2, CTSG that are involved in the ACE inhibitor pathway .CTSG is considered an important gene as it is a part of the apoptotis cycle and the

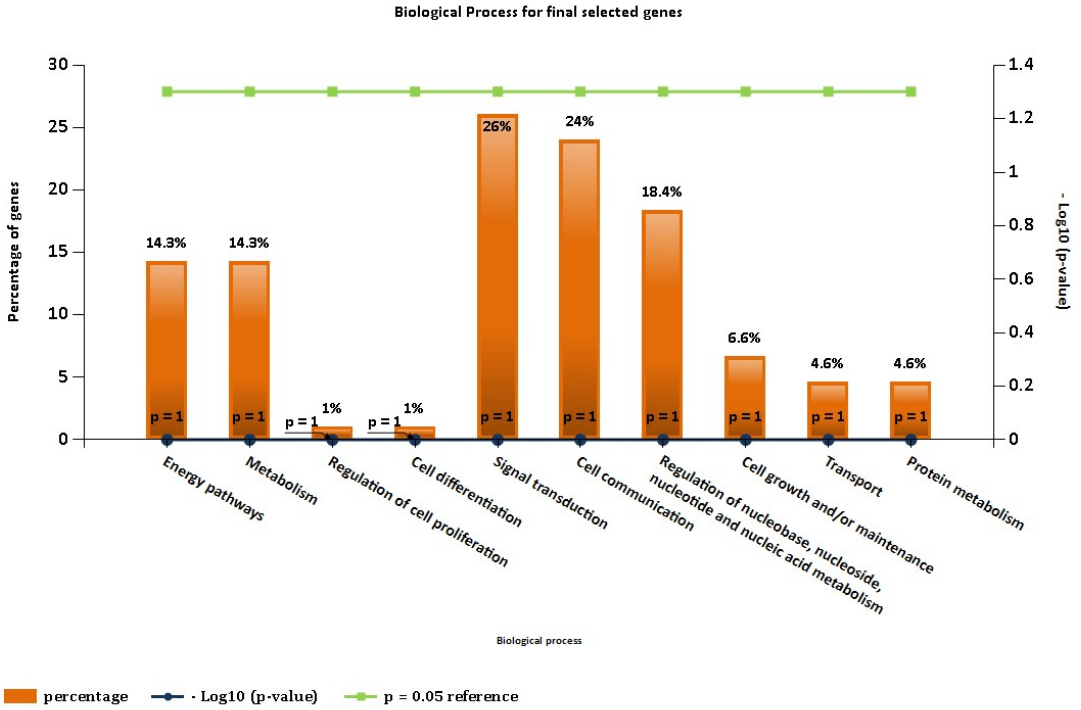


Fig. 5.2 Biological processes

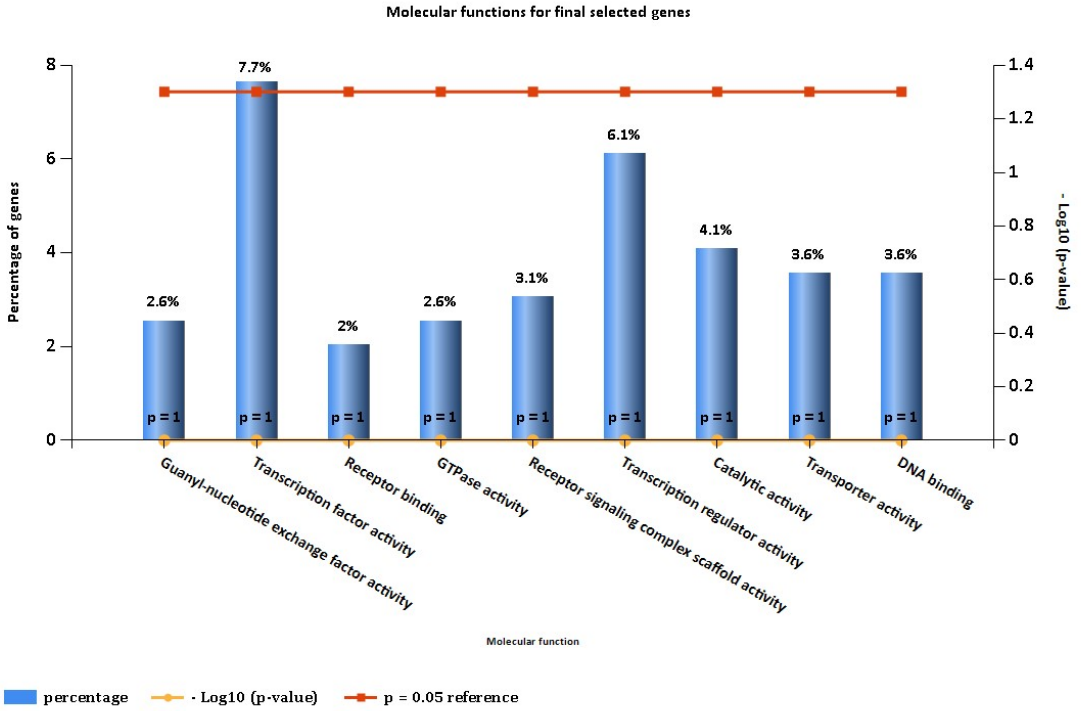


Fig. 5.3 Molecular Functiions

only gene identified, common to discriminate gender in ccRcc[4]. Another gene LIMCH1 was inferred for which RCC tumor patients had missense mutations[41]. DNASE1L3 also among the top genes , is activated during apoptosis for the breakdown of DNA. DNASE1L3, ENAM ,WDR31 were identified to be prognostic markers for renal cancer[49]. Various other genes that were a part of the top genes identified are explained in the Table 5.2; some of these have already been identified to be potential markers and few are our novel findings.

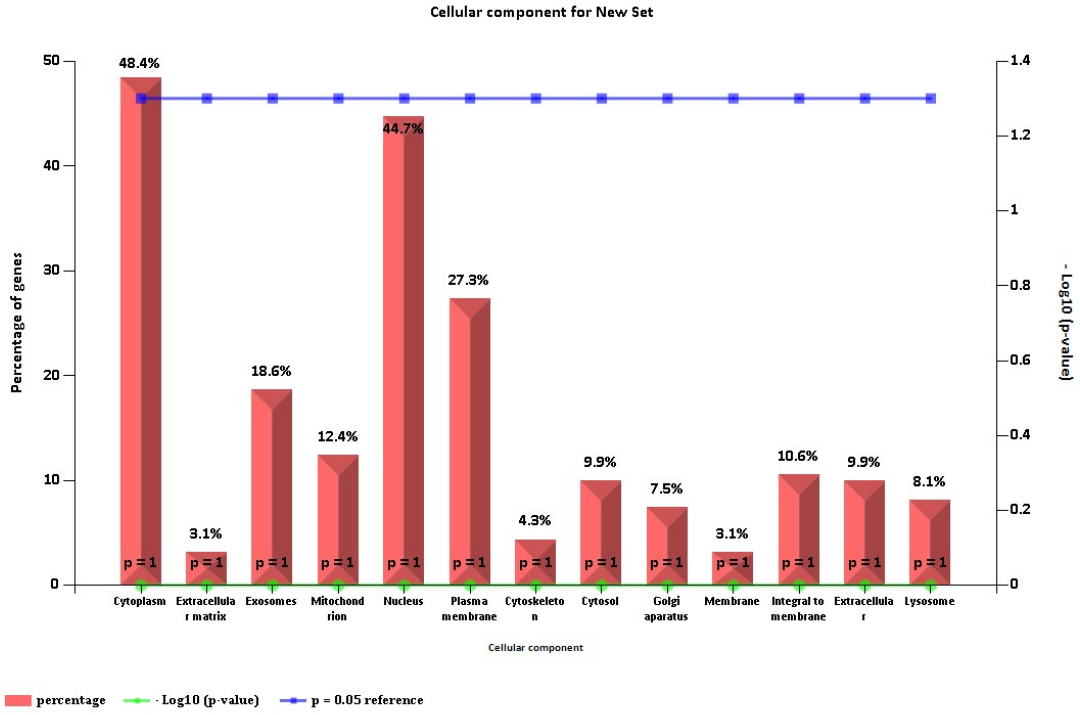


Fig. 5.4 Cellular Components

5.1.3 Methodology 2: Feature Extraction using Item Set Mining

In methodology 2, the feature were now a combination of genes instead of a single marker. Also the gene combinations were such that every gene of the gene set was bounded by a specific range. Cancer being a complex study of pathways, the idea behind this was to capture a set of correlated genes that played an important role in regulating the cancer. The genes were subjected to be bounded by specific ranges because it was observed that gene shows varied differences when expressed from one phenotype to another. Frequent item set mining was used to identify the items(gene bin pairs) that occurred together. These itemsets as features were fed to machine learning classifiers(post hyper parameter tuning), the results

on the unseen data were as per Table 5.3 and Fig 5.5.

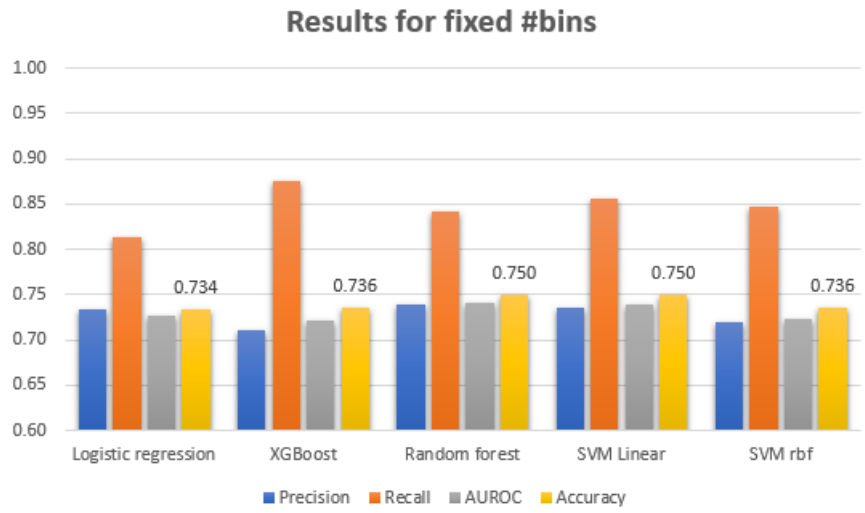


Fig. 5.5 Different metric comparison among 5 classifiers for methodology 2: fixed buckets

Metric	Logistic regression	XGBoost	Random forest	SVM linear	SVM rbf
F1- score	0.769	0.784	0.786	0.789	0.778
MCC	0.465	0.470	0.50	0.501	0.469

Table 5.3 F1 score and Matthews correlation coefficient(MCC) values for the 5 classifiers

It can be observed from the above results that SVM Linear and Random forest gave best results in terms of accuracy. The issue that was observed in this technique is that the recall in substantially higher compared to the precision indicating that the late stage patients have more misclassifications whereas the early stage patients were getting correctly classified. Due to this issue the other technique of clustering the genes was observed where each bucket was the cluster id. The following Table 5.4 and Fig 5.6 shows the results.

Metric	Logistic regression	XGBoost	Random forest	SVM linear	SVM rbf
F1- score	0.757	0.753	0.747	0.757	0.749
MCC	0.468	0.471	0.442	0.454	0.441

Table 5.4 F1 score and Matthews correlation coefficient(MCC) values for the 5 classifiers

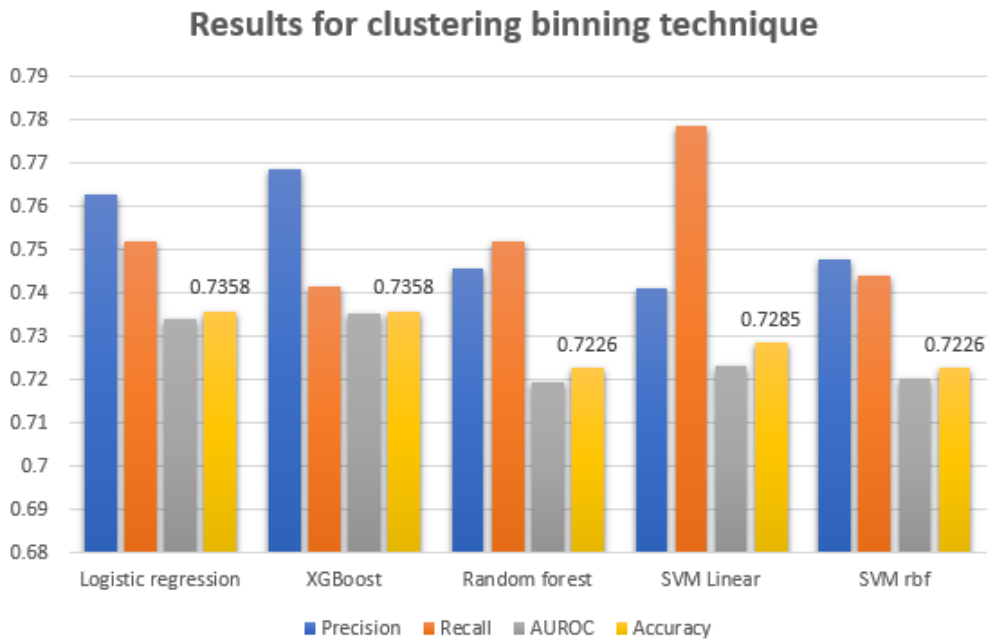


Fig. 5.6 Different metric comparison in among 5 classifiers for methodology 2: clustering

From the above results we can see that Logistic regression and XGboost were performing better compared to the other classifiers. Also it can be seen that although the accuracy is not very high but the difference in precision and recall for this technique does not have huge difference in between them indicating that the information related to early and the late stage patients both were being captured well here.
Looking at this it was decided to take the ensemble of the two bucketing techniques so that the information that was being missed by the earlier technique could now be captured correctly. Results after taking the ensemble are shown in Fig 5.7:

Looking at the observations we can clearly tell that the precision and the recall have improved and there is not much difference between the two, also there is jump in the final accuracy. Therefore, we can now say that both the early and the late stage both are classified appropriately.

5.1.4 Methodology 3:Feature selection using structural information

In the third methodology the purpose was to take into account interaction information among all the genes by building a full fledged network among them. There were two networks that were made: 1.) Gene Co-expression Network: using co-expression among the gene expression values, one for each class. 2) Gene regulatory network: Whole human gene regulatory network was extracted from an external source and was converted to a weighted

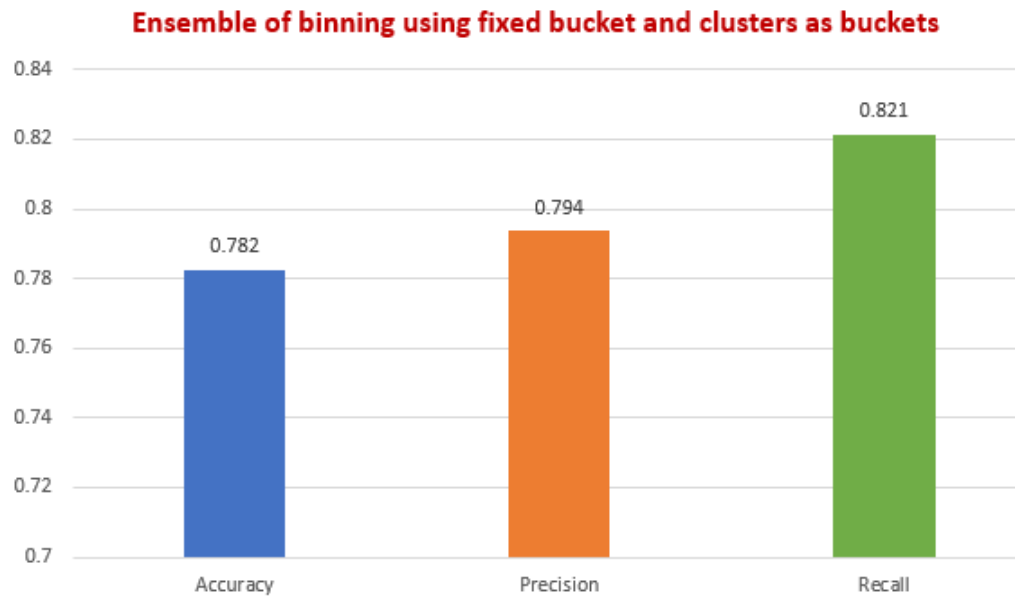


Fig. 5.7 Ensemble of binning technique having fixed 10 bins with Binning technique using clusters as buckets

network for each patient. The network properties that were selected to select the central genes in co-expression network are explained in table4.1

The genes showing distinguishing property values between early and late stage were selected as final set of features and were passed to machine learning classifiers. The results observed on the unseen data is shown in Table 5.5 and Fig 5.8

Metric	Logistic regression	XGBoost	Random forest	SVM linear	SVM rbf
F1- score	0.768	0.745	0.768	0.770	0.775
MCC	0.454	0.450	0.458	0.483	0.491

Table 5.5 F1 score and Matthews correlation coefficient(MCC) values for the 5 classifiers

For the other gene regulatory network there was one weighted graph for each patient. For every graph for each node the property was calculated and then a single averaged value for that network was captured. This formed a matrix containing a value of the network property for every patient. Bins were created within the range 0 - 1 and the average value for nodes having its normalized feature value within those bins was computed. These binned feature vectors were treated as dimensions. The results of Random Forest on unseen data was 68.45%. These networks were huge and the computation of different properties of each network required high computation. Therefore due to lack of time and enough computation, we have left this portion of research for our future work.

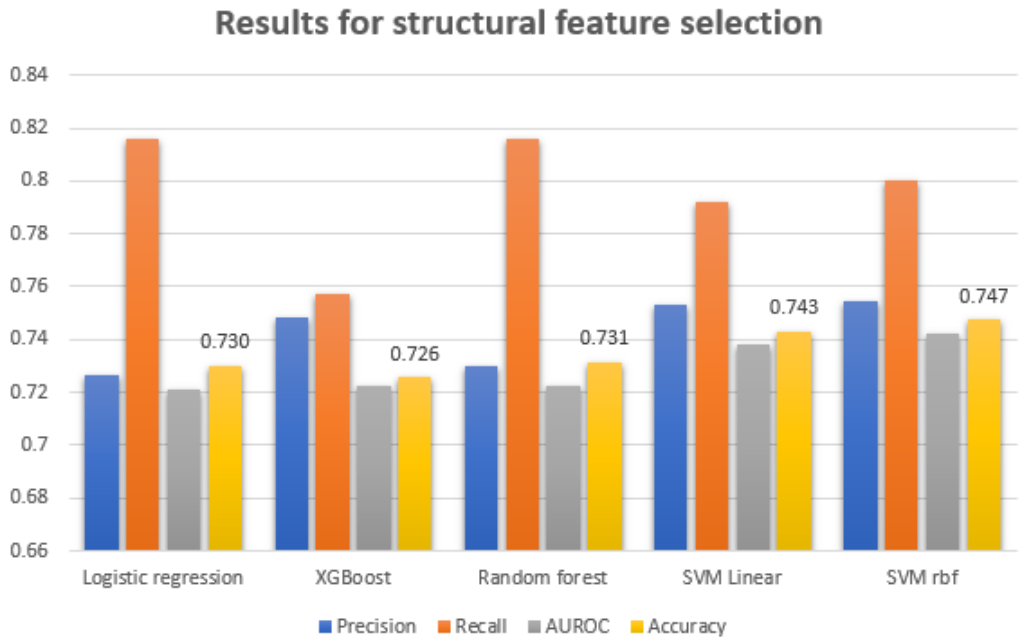


Fig. 5.8 Different metric comparison in among 5 classifiers for methodology 3: Network features for gene co-expression network

5.2 Biological Significance of the genes selected through co-expression network

The purpose of using network properties was to identify those genes that play a distinguishing role in the interactions during the prognosis of cancer. Not only the network properties provide an improved differentiation between the two stages rather also helps in bringing into play the role of the node that is involved in the interactions. Amid all the nodes in the network, there a few nodes that play a central figure in regulating one interaction to the other. These central nodes reflect either structurally or functionally important interactions that further cultivates the tumor growth or any other prognostic property of cancer. There are different centrality measures that have been used to identify genes that are distinguishing in terms of their interactions while moving from one stage of cancer to the other. For example the degree centrality and the closeness centrality majorly looks at how the nodes are connected to each other, whereas the betweeness centrality that looks at the shortest path between a pair of nodes, helping to look at the nodes that act as bridges between various regions. The strategy employed above was to identify nodes that had the maximum capability to distinguish the two stages by looking at a combined set of network properties. A set of 55 genes were selected.

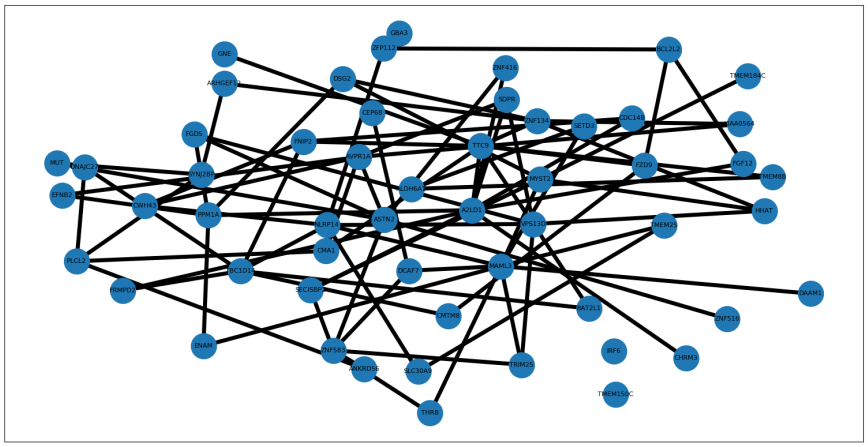


Fig. 5.9 Late stage graph of the final 55 genes

Fig 5.9 and 5.10 shows the two early stage and late stage graph of the final selected 55 genes. Looking at the graph it can be observed that the genes have evidently different connections.

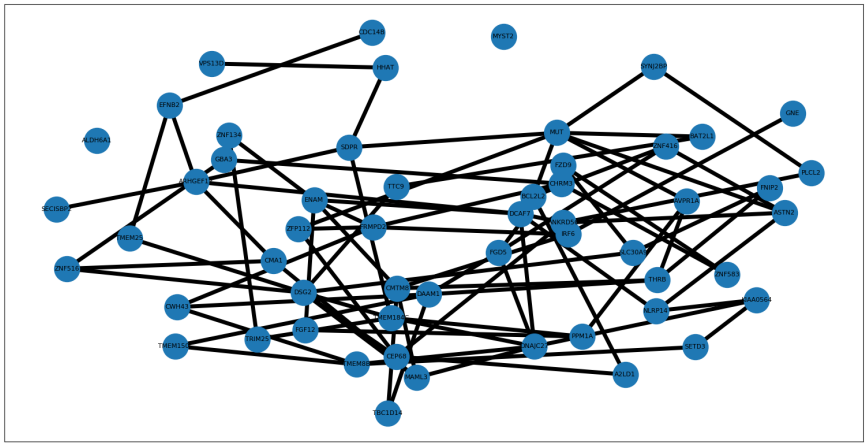


Fig. 5.10 Early stage graph of the final 55 genes

A few top highly discriminating genes were analyzed where TMEM8B gene positively regulating cell adhesion pathway , which plays a very important role in progression of ccRCC[Gong et al.]. DSG2, has been identified having missense mutation in renal cancer

and is involved in biological process of cell adhesion[36]. PLCL2, was also identified as an important gene in cancer development. Biological processes related to PLCL2 were B cell proliferation involved in immune response and B cell differentiation. Another set of genes like KIAA0564, VPS13D was identified , which has been reported as prognostic markers in renal cancer. Various other genes that were a part of the top genes identified are explained in the table5.2, that include a few genes that have already been identified to be potential markers and a few novel ones.

5.3 Ensemble of all the three feature selection techniques

After taking in to account all kinds of information that could be captured related to ccRCC. It was decided to take an ensemble of all the techniques. So that it could be seen that how incorporating different aspects of the cancer study can help predicting early and late stage in a better manner.

Ensemble learning has been a method to improve machine learning results by combining several models. They are meta-algorithms that combine various machine learning results into one predictive result that decrease the variance, bias and give better predictions.

5.3.1 Ensemble of method 1 with method 2 final results

Taking the combination of predictions of the two methods of statistical feature selection and identifying important patterns using frequent item set mining improved prediction results on the unseen data as shown below in Fig 5.11:

5.3.2 Ensemble of method 1, method 2 and method 3 final results

Another combination of the previous ensemble with the network features was taken, so that with the information captured using item set sets also the interaction information was captured. The battery of techniques were now clubbed together to observe the difference in prediction quality. The results observed are shown below in Fig 5.12:

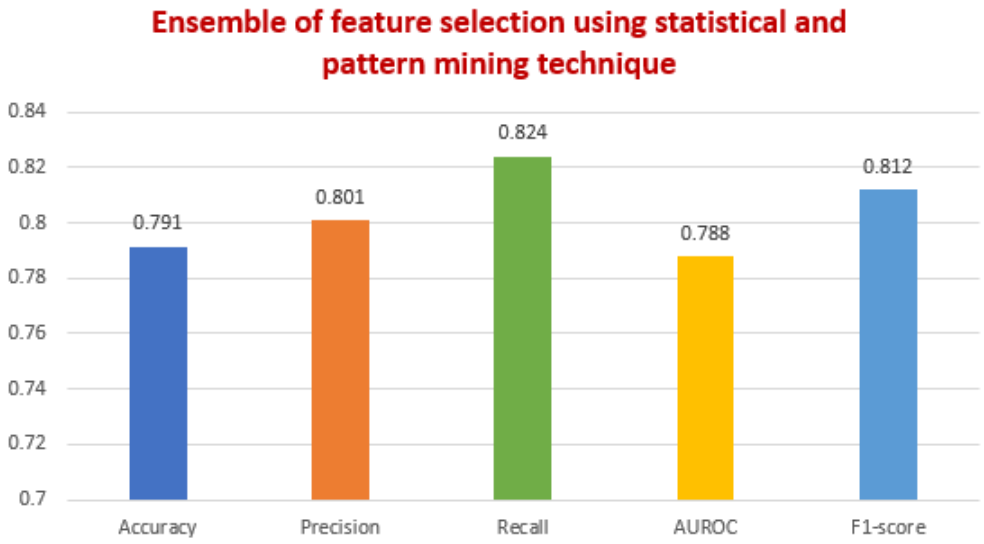


Fig. 5.11 Ensemble of statistical feature selection technique and patterns identified using frequent item set mining

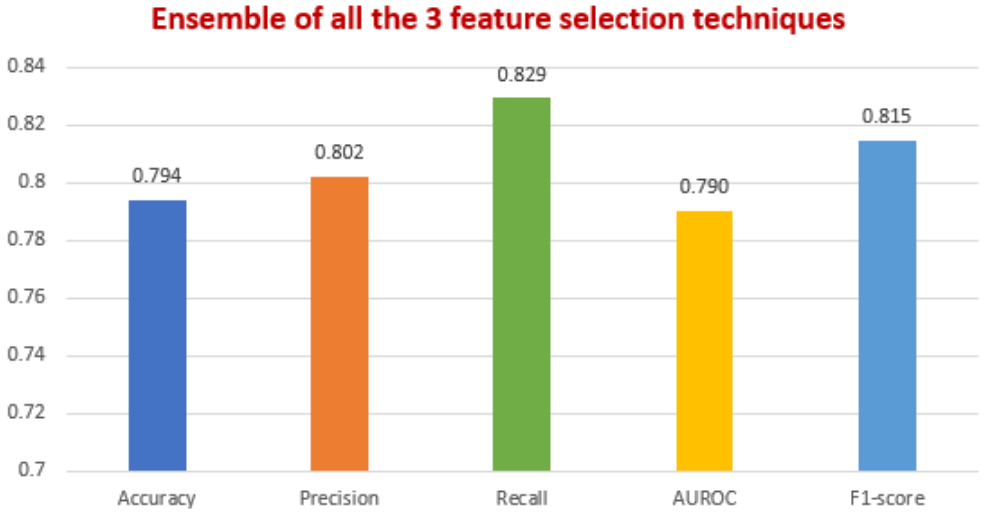


Fig. 5.12 Ensemble of statistical feature selection technique and patterns identified using frequent item set mining with feature selection technique using network properties

Gene name	Molecular function	Already identified or Novel
FAM160A1	Renal cancer showed moderate to high cytoplasmic positivity[49]	[4]
LIMCH1	Negatively affect cell migration and has been identified as a candidte gene for smoking patients of ccRcc	[15]
ACADSB	It was present in the pathway of Fatty acid acid metabolism which was upregulated in ccRCC. It also is a prognostic marker of renal cancer	[43]
NR3C2	It is a tumor suppressing gene whose role in renal cancer has been identified.	[4]
TBC1D14	It has been a part of the GTPase activity which can further lead to cancerous activities	Novel
DNASE1L3	It helps in breakdown of DNA in apoptosis and has been identified as an identifier in the biological pathways of ccRCC phenotype [3]	[4]
C1orf69	is involved in Transferase activity, poly(A) RNA-bindings , identified in renal cancer	[4]
CTSG	involved in the ACE inhibitor pathways and identified as an important gene to discriminate ccRCC in terms of gender	[4]
FGF12	It belongs to fibroblast growth factor(FGF) family and enhances the invasive potential of cancer cells by upregulating secreted proteases. Some FGF members are also considered as potential targets for designing therapies against RCC[48]	Novel
NUDT6	The nudix hydroxylase (NUDT) family of genes have relevant roles in cancer growth and metastasis and NUDT6 has been identified as prognostic markers for patients of ccRCC	[52]
FAM122A	It is being overexpressed in ccRCC and it has been deduced that overexpression of FAM122A enhances cell growth and colony-forming ability[17]	Novel
CMA1	It promotes angiogenesis making it a potential marker in many types of cancers[19]	Novel
GABRB2	Has been identified as a putative target in ccRCC[21] , also has been identified as an important in thyroid cancer but its role in cancer is unclear[29]	Novel
WDR31	It is a member of WD repeat member family and members of this family are involved in a variety of cellular processes, including cell cycle progression, signal transduction, apoptosis, and gene regulation.It has been identified as one of the molecular biomarker for ccRCC.	[10]
DAAM1	It is involved in Wnt signalling pathway that can mediate tumor metasis and has been reported to be highly expressed in ccRCC.	[35]
PLCL2	identified as an important gene in cancer development, it is involved in various processes like cell proliferation	[12]
CMA1	It promotes angiogenesis making it a potential marker in many types of cancers[19]	Novel
ZNF416	It belongs to the zinc finger protein family. It is a part of processes like regulation of transcription. It has not yet been reported in ccRCC	Novel
CMTM8	It is associated with epidermal growth factor which when activated can lead to degradation in tumor cell.[53]	Novel

Table 5.2 Biological importance of the genes selected

Chapter 6

Conclusion and Future Work

6.1 Discussion

Starting with the methodologies, the most obvious difference lies in the level of details at which features have been extracted. Methodology 1 deals with the features at an individual level where it extracts the features that have maximum distinguishing quality between the two classes. Methodology 2 looks at the feature vector as group of features that have a tendency to occur together for certain support value passed by the user. Methodology 3 moves to an even intricate level where it tries to capture all interaction information among the genes. When the genes are coarse grained, then their regulation information and co-occurence information is not is not well captured. This would lead to poorer classification accuracy. The classifier is not fed with refined feature vector set and hence the classification is noisy.

Therefore method 2 and method 3, were able to capture the features at a another level deeper. It can be observe that the feature selection techiques done separately do not give a very good classification accuracy whereas if all the techniques brought together, help gather all the level information giving a jump in the prediction accuracy, also making sure the precision and recall are at par.

Also in all the methodology where the classifiers are trained, it has been taken care from the data preparation to the final prediction that the models are robust. In most of the biological research papers that have used machine learning to classify gene expression, there has not been any mention of the robustness of the classifier[4]. The parameter optimization techniques have not taken random search into account, the problem of overfitting has not been look out for[33]. Also during prediction threshold variations for final prediction is not observed and most of the metrics are not reported. Whereas in terms of biological data where the data itself has been extracted using experiments which can have their own inherent errors , these problems of overfitting and robustness cannot be left out[25]. Therefore, in

our machine learning models we have taken specific care of all these issues and have made generic classifying models that would work on any kind of data.

Using method 1 there were 221 features identified , a few of which have been explained in chapter 2. It was observed that out of the features that were selected many genes had been identified as potential markers in ccRCC, some had been explored in various other cancers revealing the technique used to be reliable. Also a few novel markers were identified.
Also the features selected using the network properties had evidence to have identified in ccRCC .

6.2 Conclusion

We would like to conclude by stating that feature selection strategy greatly affects the classification process and hence intricacy of the features should also be looked at. In terms cancer classification strategies many researchers have restricted themselves to certain strategies for certain datasets. Whereas it has been observed that differently selected features reflect different aspects of the complex gene expression dataset. The microarray dataset being extremely large,it becomes difficult to find the intrinsic characteristics using traditional methods. Therefore, the proposed strategy evidently works well by combining different aspects of feature selection as demonstrated in this thesis. Our final result that is 79.5% accuracy is the maximum accuracy that has been reported till now on the current dataset[28].

6.3 Future Work

The future work that is planned to be done for this thesis include the below points:

1. Look for more seed genes specific to renal cell carcinoma and use their biological knowledge to move towards selecting features.
2. Lasso technique was used to identify important genes. The accuracy obtained was significantly high using these gene. Being a blackbox technique, biological relevance of the technique and the selected genes needs to be looked at.
3. More computationally intensive network features will be computed for the GRNs of each patient.
4. Association rule mining : extension of frequent item set mining will be explored to capture discriminative patterns among the genes

References

[1] Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

[2] Alves, R., Rodriguez-Baena, D. S., and Aguilar-Ruiz, J. S. (2009). Gene association analysis: a survey of frequent pattern mining from gene expression data. *Briefings in Bioinformatics*, 11(2):210–224.

[3] Battaglia, C., Mangano, E., Bicciato, S., Frascati, F., Nuzzo, S., Tinaglia, V., Bianchi, C., Perego, R. A., and Cifola, I. (2012). Molecular portrait of clear cell renal cell carcinoma: An integrative analysis of gene expression and genomic copy number profiling. In *Emerging Research and Treatments in Renal Cell Carcinoma*. IntechOpen.

[4] Bhalla, S., Chaudhary, K., Kumar, R., Sehgal, M., Kaur, H., Sharma, S., and Raghava, G. P. (2017). Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. *Scientific reports*, 7:44997.

[5] Chandra, B. and Gupta, M. (2011). An efficient statistical feature selection approach for classification of gene expression data. *Journal of biomedical informatics*, 44(4):529–535.

[6] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

[7] Cong, G., Tung, A. K., Xu, X., Pan, F., and Yang, J. (2004). Farmer: Finding interesting rule groups in microarray datasets. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 143–154. ACM.

[8] Desai, A. P., Razeghin, M., Meruvia-Pastor, O., and Peña-Castillo, L. (2017). Genet: a web application to explore and share gene co-expression network analysis data. *PeerJ*, 5:e3678.

[9] Developers, N. (2010). Networkx. *networkx. lanl. gov*.

[10] Dimitrieva, S., Schlapbach, R., and Rehrauer, H. (2016). Prognostic value of cross-omics screening for kidney clear cell renal cancer survival. *Biology direct*, 11(1):68.

[11] Ding, C. and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205.

[12] Dmitriev, A. A., Rudenko, E. E., Kudryavtseva, A. V., Krasnov, G. S., Gordiyuk, V. V., Melnikova, N. V., Stakhovsky, E. O., Kononenko, O. A., Pavlova, L. S., Kondratieva, T. T., et al. (2014). Epigenetic alterations of chromosome 3 revealed by noti-microarrays in clear cell renal cell carcinoma. *BioMed research international*, 2014.

[13] Doungpan, N., Engchuan, W., Chan, J. H., and Meechai, A. (2016). Gsnfs: Gene subnetwork biomarker identification of lung cancer expression data. *BMC medical genomics*, 9(3):70.

[14] Dussaut, J. S., Gallo, C. A., Cravero, F., Martínez, M. J., Carballido, J. A., and Ponzoni, I. (2017). Gernet: a gene regulatory network tool. *Biosystems*, 162:1–11.

[15] Eckel-Passow, J. E., Serie, D. J., Bot, B. M., Joseph, R. W., Cheville, J. C., and Parker, A. S. (2014). Anks1b is a smoking-related molecular alteration in clear cell renal cell carcinoma. *BMC urology*, 14(1):14.

[16] Emmert-Streib, F., de Matos Simoes, R., Mullan, P., Haibe-Kains, B., and Dehmer, M. (2014). The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. *Frontiers in genetics*, 5:15.

[17] Fan, L., Liu, M.-H., Guo, M., Hu, C.-X., Yan, Z.-W., Chen, J., Chen, G.-Q., and Huang, Y. (2016). Fam122a, a new endogenous inhibitor of protein phosphatase 2a. *Oncotarget*, 7(39):63887.

[18] Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.-W., and Tseng, V. S. (2014). Spmf: a java open-source pattern mining library. *The Journal of Machine Learning Research*, 15(1):3389–3393.

[19] Gao, L., Nieters, A., and Brenner, H. (2008). Meta-analysis: tumour invasion-related genetic polymorphisms and gastric cancer susceptibility. *Alimentary pharmacology & therapeutics*, 28(5):565–573.

[Gong et al.] Gong, X., Zhao, H., Saar, M., Peehl, D. M., Brooks, M., and James, D. mir-22 regulates invasion, gene expression and predicts overall survival in patients with clear cell renal cell carcinoma. *Kidney Cancer*, (Preprint):1–14.

[21] Gulati, S. (2016). *clear cell Renal Cell Carcinoma: Biomarkers and Networks*. PhD thesis, UCL (University College London).

[22] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

[23] Gyenesei, A., Wagner, U., Barkow-Oesterreicher, S., Stolte, E., and Schlapbach, R. (2007). Mining co-regulated gene profiles for the detection of functional associations in gene expression data. *Bioinformatics*, 23(15):1927–1935.

[24] Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[25] Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12.

[26] Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13.

[27] Ilayaraja, M. and Meyyappan, T. (2013). Mining medical data to identify frequent diseases using apriori algorithm. In *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, pages 194–199. IEEE.

[28] Jagga, Z. and Gupta, D. (2014). Classification models for clear cell renal carcinoma stage progression, based on tumor rnaseq expression trained supervised machine learning algorithms. In *BMC proceedings*, volume 8, page S2. BioMed Central.

[29] Jin, Y., Jin, W., Zheng, Z., Chen, E., Wang, Q., Wang, Y., Wang, O., and Zhang, X. (2017). Gabrb2 plays an important role in the lymph node metastasis of papillary thyroid cancer. *Biochemical and biophysical research communications*, 492(3):323–330.

[30] Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559.

[31] Latkowski, T. and Osowski, S. (2015). Data mining for feature selection in gene expression autism data. *Expert Systems with Applications*, 42(2):864–872.

[32] Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., and Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS computational biology*, 4(11):e1000217.

[33] Lever, J., Krzywinski, M., and Altman, N. (2016). Points of significance: model selection and overfitting.

[34] Li, B. and Dewey, C. N. (2011). Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323.

[35] Li, X., Meng, X., Wei, C., Zhou, Y., Chen, H., Huang, H., and Chen, M. (2018). Dissecting lncrna roles in renal cell carcinoma metastasis and characterizing genomic heterogeneity by single-cell rna-seq. *Molecular Cancer Research*, 16(12):1879–1888.

[36] Liu, K., Ren, Y., Pang, L., Qi, Y., Jia, W., Tao, L., Hu, Z., Zhao, J., Zhang, H., Li, L., et al. (2015). Papillary renal cell carcinoma: a clinicopathological and whole-genome exon sequencing study. *International journal of clinical and experimental pathology*, 8(7):8311.

[37] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, page S7. BioMed Central.

[38] McCall, M. N. (2013). Estimation of gene regulatory networks. *Postdoc journal: a journal of postdoctoral research and postdoctoral affairs*, 1(1):60.

[39] Naulaerts, S., Meysman, P., Bittremieux, W., Vu, T. N., Vanden Berghe, W., Goethals, B., and Laukens, K. (2013). A primer to frequent itemset mining for bioinformatics. *Briefings in bioinformatics*, 16(2):216–231.

[40] Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., et al. (2018). The biogrid interaction database: 2019 update. *Nucleic acids research*, 47(D1):D529–D541.

[41] Rosenberg, J. E., Bambury, R. M., Van Allen, E. M., Drabkin, H. A., Lara, P. N., Harzstark, A. L., Wagle, N., Figlin, R. A., Smith, G. W., Garraway, L. A., et al. (2014). A phase ii trial of as1411 (a novel nucleolin-targeted dna aptamer) in metastatic renal cell carcinoma. *Investigational new drugs*, 32(1):178–187.

[42] Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517.

[43] Schuetz, A. N., Yin-Goen, Q., Amin, M. B., Moreno, C. S., Cohen, C., Hornsby, C. D., Yang, W. L., Petros, J. A., Issa, M. M., Pattaras, J. G., et al. (2005). Molecular classification of renal tumors by gene expression profiling. *The Journal of Molecular Diagnostics*, 7(2):206–218.

[44] Sootanan, P., Prom-on, S., Meechai, A., and Chan, J. H. (2012). Pathway-based microarray analysis for robust disease classification. *Neural Computing and Applications*, 21(4):649–660.

[45] Su, Q., Wang, Y., Jiang, X., Chen, F., and Lu, W.-c. (2017). A cancer gene selection algorithm based on the ks test and cfs. *BioMed research international*, 2017.

[46] Takahashi, M., Rhodes, D. R., Furge, K. A., Kanayama, H.-o., Kagawa, S., Haab, B. B., and Teh, B. T. (2001). Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification. *Proceedings of the National Academy of Sciences*, 98(17):9754–9759.

[47] Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68.

[48] Tsimafeyeu, I. and Bratslavsky, G. (2015). Fibroblast growth factor receptor 1 as a target for the therapy of renal cell carcinoma. *Oncology*, 88(6):321–331.

[49] Uhlén, M., Björling, E., Agaton, C., Szigyarto, C. A.-K., Amini, B., Andersen, E., Andersson, A.-C., Angelidou, P., Asplund, A., Asplund, C., et al. (2005). A human protein atlas for normal and cancer tissues based on antibody proteomics. *Molecular & cellular proteomics*, 4(12):1920–1932.

[50] van Dam, S., Vosa, U., van der Graaf, A., Franke, L., and de Magalhaes, J. P. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, 19(4):575–592.

[51] Wang, J., Zhou, S., Yi, Y., and Kong, J. (2014). An improved feature selection based on effective range for classification. *The Scientific World Journal*, 2014.

[52] Wang, Y., Wan, F., Chang, K., Lu, X., Dai, B., and Ye, D. (2017). Nudt expression is predictive of prognosis in patients with clear cell renal cell carcinoma. *Oncology letters*, 14(5):6121–6128.

[53] Xie, J., Yuan, Y., Liu, Z., Xiao, Y., Zhang, X., Qin, C., Sheng, Z., Xu, T., and Wang, X. (2014). Cmtm3 is frequently reduced in clear cell renal cell carcinoma and exhibits tumor suppressor activities. *Clinical and Translational Oncology*, 16(4):402–409.

[54] Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., and Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature communications*, 5:3231.