Prostate Cancer Dataset Analysis for finding relevant Biomarkers using Machine Learning Techniques

Krishna Brahmbhatt
Computer Science
University of Windsor
Windsor, ON, Canada
brahmbhk@uwindsor.ca

***Abstract*—Prostate Cancer is one of the most common type of cancer found in prostate gland of male reproductive system which is a very slow-growing cancer, often causing no symptoms until it is in an advanced stage. But once prostate cancer begins to grow quickly or spreads outside the prostate, it is very dangerous. And thus, it is very important to detect this disease in an early stage. In this project, I present an advanced approach for finding meaningful biomarkers for Prostate Cancer by using Machine Learning Techniques. Biomarkers are essentially chemicals that indicate the normal and abnormal processes in the body.**

In this project, we have Prostate Adenocarcinoma clinical data which has 72 biomarkers/clinical variables of 499 patients/samples and Prostate Adenocarcinoma gene data with 60493 gene expressions and 494 samples. I combined data from TCGA-clinical-data and TCGA-genes datasets and removed low variance features by applying Variance threshold technique along with data filtering wherein I remove gene expressions which has zero, Nan or N/A values. This reduced features from 60k to almost 20-25k. Thereafter I applied feature selection using Random Forest (Information Gain) to reduce it to 600 features. Lastly, I used mRMR technique to select top 20 features followed with Linear Discriminant Analysis (LDA) in predicting Gleason score and used Chi-Squared feature selection with LDA for predicting Clinic t-stage.

***Index Terms*—Machine Learning, Feature Selection, Linear Discriminant Analysis, Random Forest Classifier, Repeated Stratified KFold, Variance Threshold,Chi-Squared, mRMR technique, Prostate Cancer Dataset**

I. Introduction

Prostate cancer is the most common cancer among American and Canadian men. It is third leading cause of death from cancer in Canadian men and second leading in American men. It is deadly and very dangerous if it is not identified in early stage. The American Cancer Society's estimates for prostate cancer in the United States for 2021 are about 248,530 new cases of prostate cancer and about 34,130 deaths[3]. However, the death rate dropped by around half from the mid-1990s to the mid-2010s as a result of advances in screening and treatment. There has been a lot of research on cancer diagnosis by using machine learning techniques. In this report, I propose a model for finding meaningful and relevant biomarkers using Machine Learning Techniques such as Feature Selection, dimensionality reduction, Classification, etc. Traditionally, prostate cancer studies focus primarily on discovering biomarkers for differentiation between benign and malignant tumors. However, recent studies have been considering some other aspects of the tumors including progression, metastasis, Gleason score, clinic t-stage and recurrence among others. Gene expressions are very helpful to diagnose and prognosis any form of cancer it performs decently only if the most relevant gene subsets are taken into account out of hundreds of thousands of genes[8]. Therefore, picking the best feature subset is the most challenging task in the prediction of different aspects of prostate cancer which is receiving immense attention from researchers all around the world. In this report, I am proposing a machine learning model which selects the best sunset of features to predict Gleason score and clinic t-stage. Selected top 20 features for Gleason score and clinic t-stage using various features selection methods, dimensionality reduction and classification. Further details are discuss in the following report.

II. Methodology

Machine Learning is modernized learning with for all intents and purposes zero human intervention. It incorporates model which give us useful information on some of the most important genes playing major role in prostate cancer. This section focuses on the machine learning techniques implemented in order to find meaningful biomarkers for prostate cancer. This model is mainly divided into four stages, Dataset and Preprocessing Data, Feature Selection, Dimensionality Reduction and lastly Classification. Filtered data is normalized and scaled so that highly variable genes are extracted as part of the feature selection step, and for dimensionality reduction, I used LDA (Linear Dimensionality Reduction) which transforms data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data. For classification, I used Random Forest, Support Vector Machine and K Nearest Neighbours to produce best results by the study of data. Other experimental results and methods are also discussed in this report.

A. Dataset

The dataset used in this paper collected from cBioPortal for Cancer Genomics under the Prostate Adenocarcinoma (TCGA, Provisional). There were two datasets: TCGA clinical data which has 499 samples and 72 clinical variables from which I have deducted 5 samples which were not present in other. The second dataset is TCGA gene data which has 494 samples and 60493 genes. In this research I have analyzed both the datasets and targeted two clinical variables, Gleason score and clinical t-stage. The Gleason classification system is most often used to grade prostate cancer. It is used for adenocarcinoma, which is the most common type of prostate cancer. The grade is a description of how the cancer cells look and act compared to normal cells[8]. Knowing the grade gives your healthcare team an idea of how quickly the cancer may be growing and how likely it is to spread. This helps them plan cancer treatment. Gleason 6 or lower: the cells look similar to healthy cells, which is called well differentiated. Gleason 7: the cells look somewhat similar to healthy cells, which is called moderately differentiated. Gleason 8, 9, or 10: the cells look very different from healthy cells, which is called poorly differentiated or undifferentiated[5]. Clinical stages are also used to describe prostate cancer. It is based on the results of DRE (digital rectal exam), PSA testing, and Gleason score. Clinical t-stage consists of 4 categories for describing the local extent of a prostate tumor, ranging from T1 to T4, each having subcategories. T1 stage is when the tumor is not visible through ultrasound or other imaging techniques. T2 stage is when the tumor is felt with DRE or seen with imaging such as transrectal ultrasound, but it still appears to be confined to the prostate[8]. On the other hand, T3 stage defines that the tumor has grown outside prostate and may have grown into the seminal vesicles. Lastly, T4 tumors are those which has grown into tissues next to prostate such as the rectum, the bladder, and/or the wall of the pelvis etc[5]. Gleason score will lead to successful prediction of the latter targets like clinical t-stage or tumor recurrence. As a result, similar level of accuracy score is expected for both outcomes if they are obtained through same relevant gene subsets.

B. Data Preprocessing

This step includes filtering out genes and cells based on quality metrics, normalization and scaling, feature selection, and quality control. Firstly, I removed all the NAN values, [not available], [discrepancies] and some columns with mostly or all the zero values, 10% to 13% of data was removed by keeping some general filters. After removing some common values prepossessing of data takes place. Data preprocessing is a crucial step which helps enhance the quality of data to promote the extraction of meaningful insights from the data. It refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a machine learning model. In simple words, data preprocessing in Machine Learning is a data mining technique that transforms raw data into an understandable and readable format. For this model, I have used Min-Max scaling technique which is used to scale the data to a range of 0 to 1. All features will be transformed into the range [0,1] meaning that the minimum and maximum value of a feature/variable is going to be 0 and 1, respectively. Variables that are measured at different scales do not contribute equally to the model fitting, model learned function and might end up creating a bias. Thus, to deal with this potential problem I used normalization such as Min-Max Scaling prior to model fitting. Below is the equation for Min-Max Scaler.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

C. Feature Selection

Feature selection methods use statistical measures to score the correlation or dependence between input variables that can be filtered to choose the most relevant features. For this study, I have used Low-variance followed by feature selection using random forest, mRMR feature selection and chi-squared. Details are explained in following report.

1) Low Variance: It is a simple baseline approach to feature selection which removes all features whose variance doesn't meet threshold. By default, it removes all zero-variance features, i.e., features that have the same value in all samples. If the variance is low or close to zero, then a feature is approximately constant and will not improve the performance of the model. In that case, it should be removed. By applying low variance features were reduced to 20k to 25k from 60k. I have applied this approach for predicting Gleason score as well as Clinic t-stage. Below is the equation for variance:

$$\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{n}$$

2) FS using Random forest: Feature selection using Random forest comes under the category of Embedded methods. Embedded methods combine the qualities of filter and wrapper methods. To get better information, I have selected features that are at the top of the trees, those features in general are more important than features that are selected at the end nodes of the trees, as generally the top splits lead to bigger information gains[4]. The main idea is to identify features which contain the most information regarding the target feature and then split the dataset along the values of these features such that the target feature values at the resulting nodes are as pure as possible. This informativeness is given by a measure called 'information gain'. For predicting Gleason score and Clinic t-stage, I am selecting top 600 features from 20k to 25k. For this we are calculating information gain which is simply the expected reduction in entropy caused by partitioning the data set according to its attribute. The information gain (Gain(S,A) of an attribute A relative to a collection of data set S, is defined as-

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$E = -\sum_{i}^{C} p_i \log_2 p_i$$

3) Minimum redundancy maximum relevance (mRMR): mRMR tends to select features with a high correlation with the class (output) and a low correlation between themselves. Features are selected one by one by applying a greedy search to maximize the objective function, which is a function of relevance and redundancy[10]. Two commonly used types of the objective function are MID (Mutual Information Difference criterion) and MIQ (Mutual Information Quotient criterion) representing the difference or the quotient of relevance and redundancy, respectively. I have used MIQ technique to process feature selection, the basic concept of minimum redundancy is to select the genes such that they are mutually maximally dissimilar to other genes. Let's denote the subset of genes that we are seeking. The average minimum redundancy is given as:

$$\text{Minimum}W = \frac{1}{|s|^2} \sum_{i,j \in s} I(i, j),$$

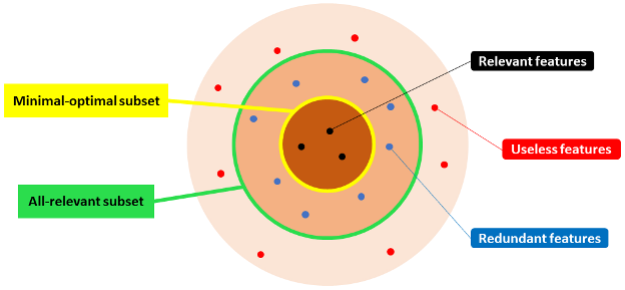

Fig. 1. mRMR Feature Selection.

In order to handle multivariate temporal data without previous data flattening, mRMR (MIQ) uses modified evaluation procedure for relevance and redundancy. Below are some of the other related methods of mRMR[11].

For this project, I have used mRMR (MIQ) to select top 20 features from 600 features for better prediction of Gleason score.

4) Chi-Squared: This feature selection method measures the degree of independence of a feature to the classes. A set of features is passed in this method based on which the features will be ranked based on the index and we can select K number of features from total n features. When two features are independent, the observed count is close to the expected count, thus we will have smaller Chi-Square value. So high Chi-Square value indicates that the hypothesis of independence is incorrect. In simple words, higher the Chi-Square value the feature is more dependent on the response and it can be

Type	Acronym	Full Name	Formula
Discrete	MID	Mutual information difference	$\max_{i \in \Omega_S}[I(i,h) - \frac{1}{\lvert S \rvert} \sum_{j \in S} I(i,j)]$
	MIQ	Mutual information quotient	$\max_{i \in \Omega_S}\{I(i,h)/[\frac{1}{\lvert S \rvert} \sum_{j \in S} I(i,j)]\}$
Continuous	FCD	*F*-test correlation difference	$\max_{i \in \Omega_S}[F(i,h) - \frac{1}{\lvert S \rvert} \sum_{j \in S} \lvert c(i,j) \rvert]$
	FCQ	*F*-test correlation quotient	$\max_{i \in \Omega_S}\{F(i,h)/[\frac{1}{\lvert S \rvert} \sum_{j \in S} \lvert c(i,j) \rvert]\}$
	FDM	*F*-test distance multiplicative	$\max_{i \in \Omega_S}[F(i,h) \cdot \frac{1}{\lvert S \rvert} \sum_{j \in S} d(i,j)]$
	FSQ	*F*-test similarity quotient	$\max_{i \in \Omega_S}\{F(i,h)/[\frac{1}{\lvert S \rvert} \sum_{j \in S} \frac{1}{d(i,j)}]\}$

Fig. 2. mRMR methods.

selected for model training. It is used in statistics to test the independence of two events. Given the data of two variables, we can get observed count O and expected count E. Chi-Square measures how expected count E and observed count O deviates each other. I have used top 20 features for Clinic t-stage's better prediction. Mathematical equation is as shown below.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

D. Dimensionality Reduction

Dimensionality reduction is the process of reducing the dimensionality of the feature space with consideration by obtaining a set of principal features. More input features often make a predictive modeling task more challenging to model, more generally referred to as the curse of dimensionality. High-dimensionality statistics and dimensionality reduction techniques are often used for data visualization. Dimensionality reduction techniques can be used in applied machine learning to simplify a classification or regression dataset in order to better fit a predictive model. This technique is usually performed on data prior to modeling and can be performed after data cleaning and data scaling and before training a predictive model. When we keep adding features without increasing the number of training samples as well, the dimensionality of the feature space grows and becomes sparser and sparser. Due to this sparsity, it becomes much easier to find a "perfect" solution for the machine learning model which highly likely leads to over-fitting. An over-fitted model would work too well on the training dataset so that it fails on future data and makes the prediction unreliable.

Linear Dimensionality reduction (LDA): The linear discriminant analysis (LDA) is one of the most traditional linear dimensionality reduction methods. This technique for multi-class classification can be used to automatically perform dimensionality reduction. Linear projections for

dimensionality reduction, computed using linear discriminant analysis are commonly based on optimization of certain separability criteria in the output space. LDA computes the directions or linear discriminants that represents the axes that maximize the separation between multiple classes. For this study, I have used LDA for both Gleason score as well as Clinic t-stage. Without applying LDA gives around 92 to 94 percent accuracy for Gleason score and clinical t-stage dataset, by using LDA for n-components from 3 to 6 increases the accuracy to 98 to 100 percent.

III. Classification

Classification is a process of categorizing a given set of data into classes; it can be performed on both structured and unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories. The model predicts or draws a conclusion to the input data given for training, it will predict the class or category for the data. The most common classification problems are – speech recognition, face detection, handwriting recognition, document classification, etc[7]. It can be either a binary classification problem or a multi-class problem. In this project, I have used Random forest, Support Vector Machines and K-nearest neighbour to predict Gleason score and Clinic t-stage.

K-Fold Cross Validation

Cross-validation is a statistical method used to estimate the skill of machine learning models on unseen data. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. This procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation[12]. It shuffles the dataset randomly and split the dataset into k groups. For each unique group, it takes the group as a hold out or test data set then the remaining groups as a training data set. Fit's a model on the training set and evaluate it on the test set. It retains the evaluation score and discard the model. Lastly, summarize the skill of the model using the sample of model evaluation scores. Each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times. For this model, I have used 10-Fold cross Validation[12].

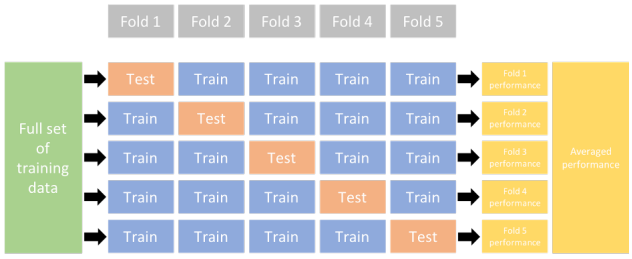

Fig. 3. K-fold cross validation

A. Random Forest

Random forest is an ensemble learning method for classification, regression, etc. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction(regression) of the individual trees. A random forest is a meta-estimator that fits a number of trees on various sub-samples of data sets and then uses an average to improve the accuracy in the model's predictive nature[1]. The sub-sample size is always the same as that of the original input size, but the samples are often drawn with replacements. Random forests are basically a bag containing n Decision Trees having a different set of hyper-parameters and trained on different subsets of data. Let's say, I have 100 decision trees in my Random forest bag!! As I just said, these decision trees have a different set of hyper-parameters and a different subset of training data, so the decision or the prediction given by these trees can vary a lot. Let's consider that I have somehow trained all these 100 trees with their respective subset of data. Now I will ask all the hundred trees in my bag that what is their prediction on my test data. Now we need to take only one decision on one example or one test data, we do it by taking a simple vote. We go with what the majority of the trees have predicted for that example. In this paper, all two datasets have unbalanced distribution of samples which can cause other classifiers to be biased towards the dominating classes[7]. However, Random Forest has tackled this problem well and has given higher accuracy for both Gleason score and Clinic t-stage.

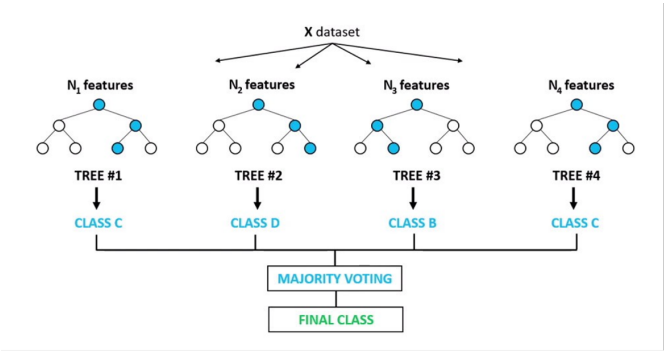

Fig. 4. Random Forest Classifier

B. SVM

Support Vector Machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N - the number of features) that distinctly

classifies the data points[7]. This algorithm takes data from the lower dimensional Y space and derives a linear function in a higher dimensional space Y'. Classification is performed on higher dimensional space. This method makes use of support vectors to derive the classifier. Its aim is to maximize the distance between the nearest point of either classes or the hyper-plane. This distance is called as margin[12].

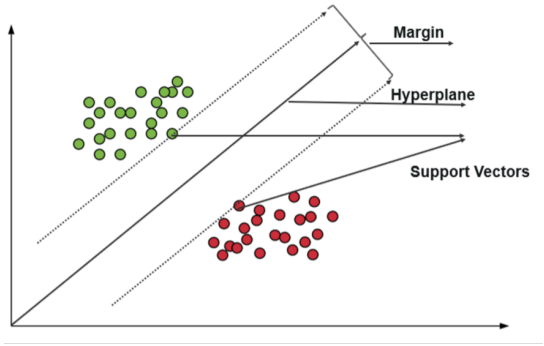

Fig. 5. Support Vector Machine Classifier

SVM represents the training data as points in space separated into categories by a gap as wide as possible. New points are then added to space by predicting which category they fall into and which space they will belong to[12]. SVM could recognize new genes as members or non-members of the class based on their expression data. SVM has 3 kernels:

1) Linear kernel: Linear kernel is used when the data is Linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a Large number of Features in a particular Dataset. SVM with linear kernel is run on the dataset to predict Gleason scores and Clinic t-stage, because Linear SVM is less prone to overfitting than non-linear. The accuracy I got for Gleason score using linear kernel is 99% and for Clinic t-stage accuracy was 95% Below is the formula for linear kernel.

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^t \mathbf{x}_j$$

2) Polynomial kernel: The polynomial kernel is a kernel function commonly used with support vector machines, it represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models. Poly kernel looks not only at the given features of input samples to determine their similarity, but also combinations of these. The accuracy I got for Gleason score using linear kernel is 100% and for Clinic t-stage accuracy was 97.5% Below is the formula for linear kernel.

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^t \mathbf{x}_j + 1)^q$$

3) RBF kernel: The RBF kernel defines similarity to be the Euclidean distance between the two inputs. If the two inputs are right on top of each-other, they get the maximal similarity of 1. If they are "too far" away from each other, the RBF kernel just says that they aren't similar (returning a value near 0). This function for two points X1 and X2 computes the similarity or how close they are to each other. The accuracy I got for Gleason score using RBF kernel is 100% and for Clinic t-stage accuracy was 98% This kernel can be mathematically represented as follows:

$$K(X_1, X_2) = exp(-\frac{||X_1 - X_2||^2}{2\sigma^2})$$

C. K-nearest neighbor

It is a lazy learning algorithm that stores all instances corresponding to training data in n-dimensional space and it does not focus on constructing a general internal model, instead, it works on storing instances of training data[7]. To label a new point, it looks at the labeled points closest to that new point also known as its nearest neighbours. It has those neighbours vote, so whichever label most of the neighbours have is the label for the new point. The "k" is the number of neighbours it checks. For this project I have used k=7, and the accuracy I got for Gleason score is 100% and for Clinic t-stage accuracy was 98%

IV. Result

TABLE I
GLEASON SCORE CONFUSION MATRIX USING RANDOM FOREST

True/Predicted	Class 0	Class 1	Class 2	Class 3
Class 0	9	0	0	0
Class 1	0	54	0	0
Class 2	0	0	10	0
Class 3	0	0	0	26

Fig. 6. Gleason Score data distribution.

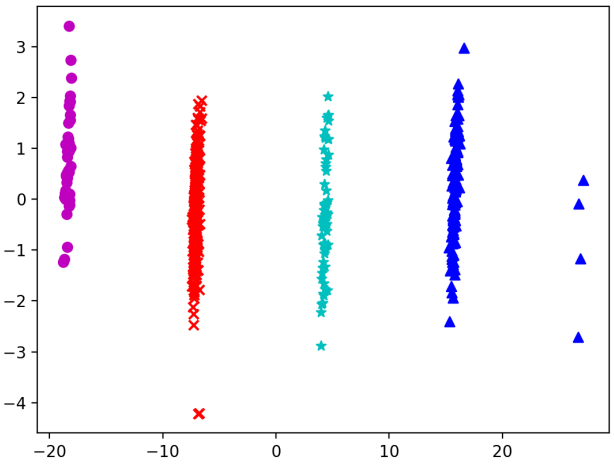

TABLE II
GLEASON SCORE ACCURACY COMPARISON

MRMR Features	RF	SVM-RBF	KNN
10	0.6566	0.617	0.610
15	0.639	0.639	0.628
20	1.000	1.000	1.000
25	1.000	1.000	1.000
30	0.999	1.000	1.000

TABLE III
GLEASON SCORE SVM - ACCURACY COMPARISON

MRMR Features	SVM-Linear	SVM-Poly	SVM-RBF
10	0.659	0.617	0.645
15	0.6768	0.627	0.652
20	1.000	1.000	1.000
25	1.000	1.000	1.000
30	1.000	1.000	1.000

Above given Table I is a confusion matrix for Gleason Score, there are 5 classes – 6, 7, 8, 9, 10. In dataset 5th class (score 10) samples were very less which was creating a discrepancy in predicting correct results so I have merged 9 and 10 score for more clear results. Table II and III are accuracy comparisons between different classifiers. Table II shows comparisons between random forest, SVM- RBF and K-NN. Table III shows comparisons between 3 different kernel functions of SVM.

TABLE IV
CLINIC T-STAGE CONFUSION MATRIX USING RANDOM FOREST

True/Predicted	0	1	2	3	4	5	6	7	8
0	1	0	0	0	0	0	0	0	0
1	0	37	0	0	0	0	0	0	0
2	0	0	1	0	0	0	0	0	0
3	0	0	0	13	0	0	0	0	0
4	0	0	0	0	11	0	0	0	0
5	0	0	0	0	0	8	0	0	0
6	0	0	0	0	0	0	8	0	0
7	0	0	0	0	0	0	0	1	0
8	0	0	0	0	0	0	0	0	1

TABLE V
CLINIC-T STAGE ACCURACY COMPARISON

Chi-Squared Features	RF	SVM-RBF	KNN
20	0.999	0.980	0.989
40	0.953	0.950	0.903
60	0.922	0.900	0.972
80	1.000	0.998	0.960
100	1.000	0.997	1.000

Similarly, above given Table IV is a confusion matrix for clinic t-stage score, there are in total 9 classes. Table V and VI are accuracy comparisons between different classifiers. Table V shows comparisons between RF, SVM- RBF and K-NN. Figure VI shows comparisons between 3 different kernel functions of SVM.

Fig 6 shows the Gleason score data after the feature selection process and Fig 7 shows the Clinic score data.

Fig. 7. Clinic Score data distribution.

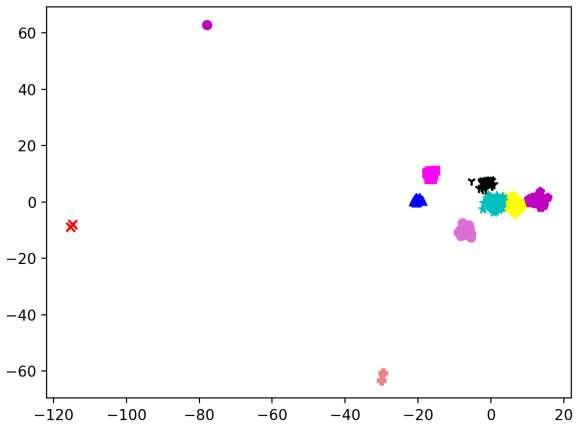

TABLE VI
CLINIC-T STAGE SVM - ACCURACY COMPARISON

Chi-Squared Features	SVM - Linear	SVM - Poly	SVM - RBF
20	0.951	0.975	0.980
40	0.953	0.955	0.950
60	0.972	0.944	0.900
80	0.960	0.987	0.998
100	0.930	0.997	0.997

V. DISCUSSION

Algorithm Analysis

According to the results, mRMR feature selection followed by LDA and Random forest classifier gives the highest accuracy score of 0.99- 1.00 on validation set. On the other hand, both Support Vector Machine and K-nearest neighbor classifier give accuracy score of 0.98 to 1.00 which is a good accuracy score. From all the above results, I think once we got relevant biomarkers, SVM -RBF and K-NN also works for getting good accuracy. The reason why Random forest is working best is because it can handle large data sets due to its capability to work with many variables running to thousands. Also, random forest is a combination of decision trees that can be modeled for prediction and behavior analysis.

mRMR feature selection for Gleason score provided relevant biomarkers which increased accuracy drastically. mRMR performed accuracy and stability measurements when it is used on different data sets. I found out that, the two feature selection methods within mRMR, MID and MIQ result in features with similar accuracy. On the other hand, MIQ results in more stable feature sets than MID and therefore should be preferred over MID, especially for prostate cancer dataset.

TABLE VII
BEFORE AND AFTER LDA COMPARISON

Prediction	Gleason Score	Clinic Score
Before LDA	0.955	0.919
After LDA	0.999	0.997

In Table VII, there was a good difference between the accuracy when I added LDA to classification. Before LDA accuracy for both Gleason score and Clinic-T stage was 92 to 96 and after I added LDA it increased to 98 to 100 percent.

Chi-Squared method is definitely computationally lighter in comparison with other feature selection methods, but here we are not aware of how many number of features exactly we want for classification. This becomes a classic two variable optimization problem where accuracy and number of features are inversely proportional[6]. When chi-square is used with SVM gave lower results compared to other classifiers and worked best with K-nearest and Random forest when used to predict clinic t-stage.

Considering a practical scenario, biologists want few 30-50 gene expressions or features based on which they can work efficiently. This constraint is satisfied by Random Forest algorithm, along with that it will never overfit the dataset which makes it better choice for classification. Based on this fact and accuracy reports, we present best 30 gene expressions which gave best accuracy for prostate cancer dataset when used with the model I propose in this report.

Based on the results and discussion we conclude that for this problem, we can use Random Forest for classification, mRMR and chi-square as feature selection technique followed by LDA for better accuracy as well as visualization of data. There are some cases where we can enhance our feature selection by applying wrapper-based feature selection method on actual obtained and gain more detail knowledge to diagnose and prognosis prostate cancer at very early stage.

VI. Conclusion

In this project we have huge data along with many classes and hence we need to come up with a strategy/technique that finds the best gene expression features and predicts the result. Prostate cancer has become a perfect zone to apply machine learning techniques as it is very common problem but very difficult to diagnose in early stage. Since it is a very common problem, we have huge data available and thus we need to first preprocess the data and then apply techniques like feature selection, dimensionality reduction and classification. We started with 60k and reduced to 20 most relevant gene expressions to predict Gleason score and Clinic t-stage. We also discussed on various observations and methods on how to get the best accuracy possible. After taking the genes information of a patient, this model can be used to decide on accurate treatments. This way doctors with the use of computational power will be able to diagnose prostate cancer at a very early stage. Based on the results and discussion we conclude that for this problem, we can use Random forest for classification, mRMR/chi-squared for feature selection There are some cases where we can make feature selection more efficient by applying wrapper based feature selection method on actual obtained features from Random Forest to cut down features to more convenient number. The accuracy level is very satisfactory and the computational speed and power is also reduced to great extent.

TABLE VIII
Top 30 Features

Name of Feature	Importance Score
ENSG00000019582.13	1
ENSG00000269890.1	1.231
ENSG00000212766.8	1.825
ENSG00000189164.13	2.412
GLEASON_PATTERN_PRIMARY	2.932
ENSG00000108958.4	3.48
ENSG00000230513.1	4.031
ENSG00000144152.11	4.565
ENSG00000251192.6	5.048
ENSG00000010539.10	5.593
ENSG00000254285.3	6.08
ENSG00000271335.4	6.537
ENSG00000132622.9	6.945
ENSG00000183765.19	7.346
ENSG00000164366.3	7.781
ENSG00000223658.6	8.264
ENSG00000267681.1	8.669
ENSG00000200550.1	9.058
GLEASON_PATTERN_SECONDARY	9.46
ENSG00000128203.6	9.858
ENSG00000251611.1	10.26
ENSG00000242375.1	10.617
ENSG00000050327.13	10.917
ENSG00000261643.1	11.101
ENSG00000228158.2	11.317
ENSG00000260711.2	11.612
ENSG00000265458.1	11.948
ENSG00000272143.1	12.156
ENSG00000237991.3	12.418

References

Bibliography

[1] A. Dubey, "Feature Selection Using Random forest," Medium, 15-Dec-2018. [Online]. Available: https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f. [Accessed: 18-Apr-2021].

[2] Alihantabak, "Prostate Cancer Predictions with ML and DL Methods," Kaggle, 06-Jan-2019. [Online]. Available: https://www.kaggle.com/alihantabak/prostate-cancer-predictions-with-ml-and-dl-methods. [Accessed: 18-Apr-2021].

[3] The American Cancer Society medical and editorial content team, "Key Statistics for Prostate Cancer: Prostate Cancer Facts," American Cancer Society, 12-Jan-2021. [Online]. Available: https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html. [Accessed: 18-Apr-2021].

[4] E. Lewinson, "Explaining Feature Importance by example of a Random Forest," Medium, 17-Apr-2020. [Online]. Available: https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e. [Accessed: 18-Apr-2021].

[5] "Grading prostate cancer - Canadian Cancer Society," www.cancer.ca. [Online]. Available: https://www.cancer.ca/en/cancer-information/cancer-type/prostate/grading/?region=on. [Accessed: 18-Apr-2021].

[6] H. Sulistiani and A. Tjahyanto, "Comparative Analysis of Feature Selection Method to Predict Customer Loyalty." Center for Scientific Publication, 2017.

[7] M. Waseem, “Classification In Machine Learning: Classification Algorithms,” Edureka, 21-Jul-2020. [Online]. Available: https://www.edureka.co/blog/classification-in-machine-learning/. [Accessed: 18-Apr-2021].

[8] “Prostate Cancer - Stages and Grades,” Cancer.Net, 27-Jan-2021. [Online]. Available: https://www.cancer.net/cancer-types/prostate-cancer/stages-and-grades. [Accessed: 18-Apr-2021].

[9] “Prostate Cancer Canada,” Canadian Cancer Society. [Online]. Available: https://www.prostatecancer.ca/. [Accessed: 18-Apr-2021].

[10] S. Mazzanti, “‘MRMR’ Explained Exactly How You Wished Someone Explained to You,” Medium, 12-Feb-2021. [Online]. Available: https://towardsdatascience.com/mrmr-explained-exactly-how-you-wished-someone-explained-to-you-9cf4ed27458b. [Accessed: 18-Apr-2021].

[11] Z. Zhao, R. Anand, and M. Wang, Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform, Aug. 2019.

[12] J. Brownlee, “A Gentle Introduction to k-fold Cross-Validation,” Machine Learning Mastery, 02-Aug-2020. [Online]. Available: https://machinelearningmastery.com/k-fold-cross-validation/. [Accessed: 18-Apr-2021].

[13] R. Gandhi, “Support Vector Machine - Introduction to Machine Learning Algorithms,” Medium, 05-Jul-2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47. [Accessed: 18-Apr-2021].