Finding most relevant biomarkers for prostate cancer using feature selection and dimensionality reduction

Zarreen Naowal Reza[1] Susha Suresh[2]

Abstract— **Prostate cancer is a major health concern for men being a life-risking disease effecting the gland in the male reproductive system named prostate. It usually grows very slowly, often causing no significant symptoms until it reaches an advanced stage. Thus, prognosis and diagnosis of this disease in early stage is a big challenge. Biomarkers are really useful in overcoming this challenge as it can detect the presence and progression of cancer in the body. In this study, Prostate Adenocarcinoma (TCGA, Provisional) dataset with 494 samples or patients report are accumulated including 71 clinical variables (biomarkers) and 60483 gene expressions from National Cancer Institutes (NCIs) Genomic Data Commons (GDC). By removing low variant features with VarianceThreshold and redundant features with Pearson Correlation technique, 370 most relevant genes are picked out of 60483 gene expressions. After that, Principal Component Analysis (PCA) is applied followed by Linear Discriminant Analysis (LDA) to reduce the dimension of the feature set to only between 3 to 7, based on individual classifiers for individual target variable. These final features are used for the classification of Gleason score and clinical t-stage, prediction of tumor recurrence and disease free survival status using Random Forest classifier and Support Vector Machine. The performance of the models are evaluated on completely separated validation set providing 96 to 100 percent accuracy for each of the cases which is a huge improvement from the past relevant studies both in terms of accuracy and computational speed.**

I. INTRODUCTION

Prostate cancer is a leading cause of death world-wide and the third leading cause of cancer death in Northern American men. According to [1], an estimated 180,890 new cases of prostate cancer (PCa) was diagnosed in 2016 and more than 26,000 died from the disease in United States (US). 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 and recurrence among others.Although, 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 genes. 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 [2], a transcriptome-based genome-wide approach is proposed to identify prognostic prostate cancer biomarkers and selected 23 gene transcripts are combined with Gleason score to obtain the level of tumor aggressiveness. Accuracy obtained from this study based on Area under the receiver-operating characteristic curve (AUC) is 0.83 to 0.88 using Gleason score as one of biomarkers along with selected genes. In our study, classification of Gleason score and clinical t-stage, prediction of tumor recurrence and disease free survival status are performed solely based on the relevant subset of genes with accuracy ranged from 0.94 to 1.0 for respective targets. The feature subset of 370 genes are selected out of 60483 genes using low-variance and collinearity removal techniques. PCA followed by LDA is applied on the gene subset to reduce the dimensionality of features based on maximizing the variance in the dataset as well as maximizing the separation between multiple classes. Finally, Random Forest and Support Vector Machine classifiers are used to fit the model to classify and predict each of the four response variables. Performance of each model for each dataset is evaluated using 10-

[1]Zareen Naowal Reza, 104721252, School of Computer Science, University of Windsor

[2]Susha Suresh, 104868759, School of Computer Science, University of Windsor

fold cross validation. Final erformance of each classifier on each classification and prediction task are measured from the ROC curve and AUC analysis gained from the performance on the held-out validation set. For classifying Gleason score, 0.95 and 0.98, for classifying clinical t-stage, 0.96 and 1.0, for predicting tumor recurrence, 0.96 and 1.0 for predicting disease free survival status, 0.96 and 1.0 accuracy score is obtained for Support Vector Machine and Random Forest classifier respectively. It is evident that Random Forest classifier gives slightly better result than SVM. However, the strength of this study is that the accuracy obtained for each of the response variable with the sole use of gene expressions has surpassed all previous accuracy scores obtained by fellow researchers and the computation speed is also very fast, 4.75 seconds to be exact.

II. Methodology

The entire research is divided into four main stages, namely Understanding and Preparing dataset, Feature Selection, Dimensionality Reduction and Classification. The paper focuses mainly on the best results produced by the study of data, other experimental results and methods are also discussed.

A. Understanding and Preparing Dataset

The datasets used in this paper are collected from cBioPortal for Cancer Genomics under the Prostate Adenocarcinoma (TCGA, Provisional) section. Initially, there were two datasets. One dataset contains 494 patient samples of 71 clinical variables and the other contains 60483 gene expressions for each of the 494 patients. For classification and prediction of the four above-mentioned targets, four separate datasets are created from the initial datasets each of which contains all patient IDs as rows and the gene expressions as columns including each of the target variables. Before deciding which techniques has to be applied on any dataset, a fair understanding of the target and features are necessary. Here, the target variables are Gleason score, clinical t-stage, tumor recurrence and disease free survival status. Gleason scores are the levels of cancerous cells that fall under one of the 5 pattern grades as they change from normal cells to tumor cells. Cancer cells with Gleason score of 6-7 can be considered as weaker whereas 8-10 is considered as poorly differentiated or high grade. These cancers are generally aggressive and likely to grow and spread more quickly.Clinical t-stage consists of 4 categories for describing the local extent of a prostate tumor, ranging from T1 to T4, each having subcategories. According to [3], T1 stage is when the tumor is not visible through ultrasound or other imaging techniques. T2 stage is when the tumor is felt with digital rectal exam (DRE) or seen with imaging such as transrectal ultrasound, but it still appears to be confined to the prostate. 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 the rectum, the bladder, and/or the wall of the pelvis etc. Tumor recurrence is the new tumor event after successful initial treatment, that means it will the answer the question to whether the patient should expect the recurrence of prostate tumor in future followed by he has been given initial treatment. Thus in this paper it has two classes - yes and no. Finally, disease free survival status is the category to predict whether the patient will be able to live a disease free life or the tumor will be recurred or progressed. Like tumor recurrence, it has two classes - disease-free and recurred/progressed. One important point to be noted is that all of the four target variables are related and dependent to each other to some extent. Therefore, the successful classification of one of them, like 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 all the four outcomes if they are obtained through same relevant gene subsets.

Other Methods: When prostate cancer is caught in its earliest stages, initial therapy can lead to high chances for cure, with most men living cancer-free for many years. But prostate cancer can be slow to grow following initial therapy, and especially for men with high risk prostate cancer and the initial treatment the patient chose, recurrence can occur in 30-90% of men. To find the recurrence of prostate cancer, another parallel study done. In this study the recurrence of prostate cancer after

initial treatment is predicted as Yes or No. To predict the recurrence of prostate cancer, a dataset with most relevant gene expression along with this other clinical variables like Primary site, Gleason score, tumor level, history of malignancy, bone scan result, CT Scan result, MRI result and Lymph count are considered.

B. Feature Selection

At first, a low-variance feature selector named VarianceThreshold is used to remove all the features that have the same value in all samples. This feature selection algorithm looks for features that have trainin-set variance close to zero and discard the from the dataset. In this prostate cancer dataset, 3,018 features are removed for having near-zero variance. One important thing to mention is that it is highly recommended to scale or normalize the feature set before applying this kind of low-variance selector methods otherwise the variance estimates can be misleading between higher value features and lower value features. Normalization of data is also mandatory to make the values of each feature in the data to have zero- mean (when subtracting the mean in the numerator) and unit-variance. In this study, Min-Max scaling technique is used to scale the data to a range of 0 to 1. The equation used for this technique is the following, $X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$ where X is the training dataset. After low-variant features are deducted, outliers or anomalies in the dataset are detected through an advanced method called Isolation Forest. The Isolation Forest algorithm isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. The way that the algorithm constructs the separation is by first creating isolation trees or random decision trees. Then, the score is calculated as the path length to isolate the observation. In order to avoid issues due to the randomness of the tree algorithm, the process is repeated several times and the average path length is calculated and normalized [4].Here are the plots of the points in the feature set of applying LDA before and after removing the outliers.

50 outliers are detected and removed from the remaining feature set. Finally, feature redundancy

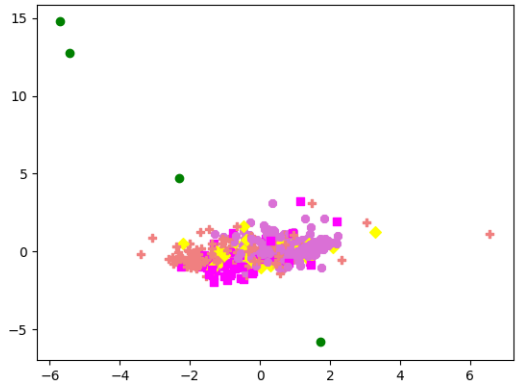

Fig. 1. LDA with outliers on Gleason Score data

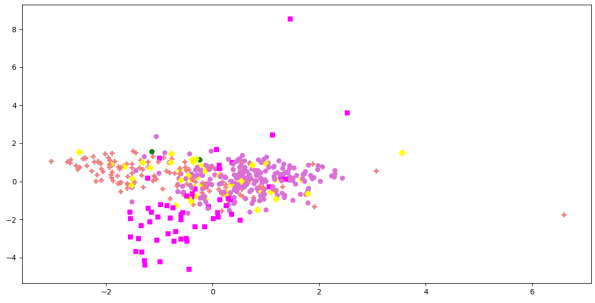

Fig. 2. LDA without outliers on Gleason Score data

is minimized by removing highly correlated features. Pearson correlation method is used to calculate the correlation between the features and one of the feature is removed from a pair having correlation value over pre-defined threshold. Pearson correlation coefficient is a measure of the strength of a linear association between two variables ranging a correlation value from -1 to +1. A value of 0 indicates that there is no association between the two variables whereas a value greater than 0 indicates a positive association and a value less than 0 indicates a negative association. High collinear variables contain duplicate and redundant information which can highly intervene with the model performance. In this gene expression dataset, there are thousands of features most of which are highly correlated with each other. For finding the best subset of relevant gene features, it is required to pick only those genes which are less correlated with each other but highly correlated

with the target variable. Therefore, Pearson correlation method is used to remove all the features that are collinear to other features. After performing this correlation method, only 370 features are left out of the previously obtained 57,465 features. These are the most relevant features for performing the classification tasks.

C. Dimensionality Reduction

Because the mutually correlated features are removed, now it is time to pick the features that are highly correlated with the target variable. In this paper, PCA followed by LDA is applied as techniques of dimensionality reduction. PCA ignores class labels and its goal is to find the directions or principal components that maximize the variance in a dataset. In contrast to PCA, LDA computes the directions or linear discriminants that represents the axes that maximize the separation between multiple classes. Applying only LDA gives around 94 to 96 percent accuracy for Gleason score and clinical t-stage dataset using only 4 to 5 linear discriminants. However, applying PCA before LDA to maximize the data variance increases the accuracy from 94-96 percent to 98-100 percent. On the other hand, using LDA solely gives accuracy score of 0.99 for both tumor recurrence and disease-free survival status. Here are the plots of applying LDA on the Gleason score and clinical t-stage and applying both PCA and LDA on the Gleason score and clinical t-stage dataset.

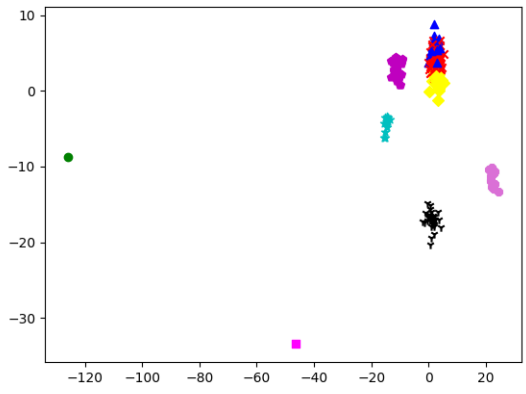

Fig. 3. Plot of T Stages with LDA and PCA

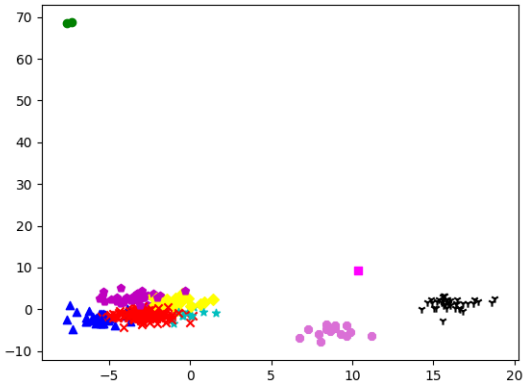

Fig. 4. Plot of T Stages with LDA

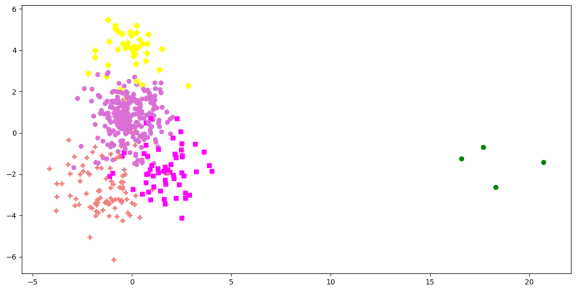

Fig. 5. Plot of Gleason scores with LDA

Other Methods: To reduce the dimensionality of gene expression data pradtcgagenes, PCA is applied on the dataset. The goal of dimensionality reduction is therefore to select features that allow for an accurate description of the disease Recurrence of cancer after initial treatment, and subsequently, reliable classification, diagnosis, and prognosis.

III. Classification

In this part, samples of Gleason score and clinical t-stage datasets are classified by class labels. On the other hand, tumor recurrence and disease-free survival status are predicted from the respective dataset. Both for classification and prediction, Support Vector Machine and Random Forest classifiers are used.

A. Support Vector Machines

Support vector machines (SVMs, also support vector networks) are supervised learning models

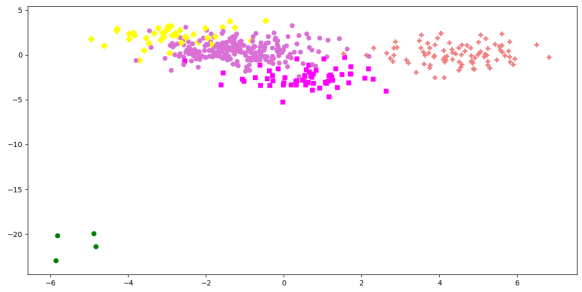

Fig. 6. Plot of Gleason scores with LDA and PCA

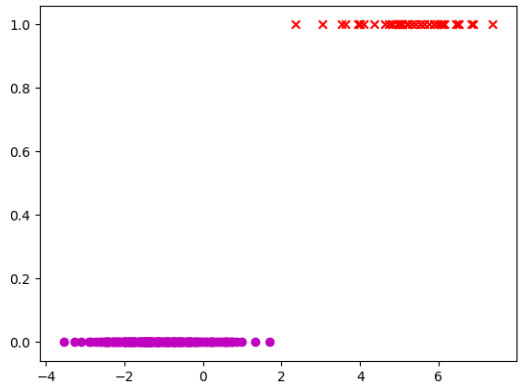

Fig. 7. Plot of LDA for new tumor event after initial treatment

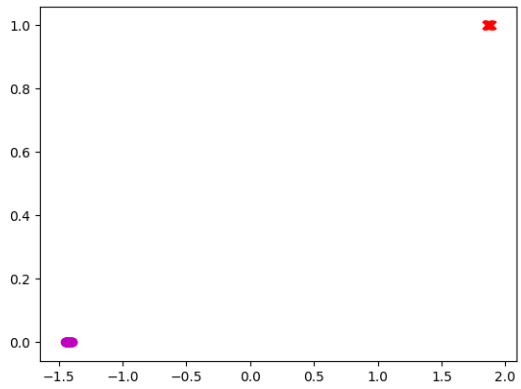

Fig. 8. Plot of LDA for new tumor event after initial treatment

with associated learning algorithms that analyze data used for classification. Given a set of training examples, each marked as belonging to one or the other of two or multiple categories, an SVM training algorithm builds a model that assigns new examples to one or the two of the other multiple categories, making it a non-probabilistic binary linear classifier. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. An SVM learns to discriminate between the members and non-members of a given functional class based on expression data. Having learned the expression features of the class, the SVM could recognize new genes as members or non-members of the class based on their expression data. On the processed dataset, SVM with different Kernels namely:

1) Linear Kernel: The application of a support vector machine with a linear kernel is to perform classification or regression. It will perform best when there is a linear decision boundary or a linear fit to the data. SVM with linear kernel is run on the dataset to predict Gleason scores, DFS Status, Tumor recurrence and T-Stage, because Linear SVM is less prone to overfitting than non-linear. It also works well when the number of features in the given dataset is very large compared to the training samples. The dataset gives the highest prediction test accuracy of 99 percent for DFS Status.

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

Fig. 9. Linear kernel Equation

2) Polynomial Kernel: Second, SVM with polynomial kernel is run on the dataset to predict Gleason scores, DFS Status, Tumor recurrence and T-Stage. Intuitively, the polynomial kernel looks not only at the given features of input samples to determine their similarity, but also combinations of these. SVM with Polynomial kernel gives the best accuracy for each of the datasets among all the SVM kernels.

3) RBF Kernel: Third, SVM with RBF kernel is used to predict the Gleason score, DFS Sta-

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

Fig. 10. Polynomial kernel Equation

tus, Tumor recurrence and T-Stage. Best hyper-parameters are determined using grid search. With RBF kernel, the highest test accuracy of 99 percent for prediction of T-Stage is obtained.

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

Fig. 11. RBF kernel Equation

4) Sigmoid Kernel: Sigmoid kernel is giving very less test accuracy (about 0.75-0.77) for prediction of Gleason scores, DFS Status, Tumor recurrence and T-Stage compared to other kernels.

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh\left(\beta \mathbf{x}_i^t \mathbf{x}_j + \gamma\right)$$

Fig. 12. Sigmoid kernel Equation

B. Random Forest

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. It grows as many decision trees in order to classify each sample separately.Each tree votes a class for each sample and class having the majority of votes is assigned to individual samples by the forest.Each tree is split on features based on their gini importance ranking. As a result, classification is performed using the most important features or attributes. Random decision forests solve the overfitting tendency of decision trees on training set. It is one of simplest and basic algorithm that tends to perform consistently better in computational biology tasks as well. One advantage of Random Forest is that it works really well on highly unbalanced datasets. In this paper, all of the four datasets have unbalanced distribution of samples which can cause other classifiers to be biased towards the dominating classes. However, Random Forest has tackled this problem well and given very high accuracy on each of the four datasets.

IV. Result

Performance of each of the classifier on each of the classification and prediction task is analyzed and compared with different parameters.

Targets	SVM (PCA+LDA)	RF (PCA+LDA)	SVM (LDA)	RF (LDA)
Gleason Score	Kernel: poly No. of PCs: 370 No. of LDs: 3 Mean accuracy on 10-fold CV: 0.974523809524 Accuracy on Validation Set: 0.966292134831	No. of PCs: 370 No. of LDs: 7 Mean accuracy on 10-fold CV: 0.983095238095 Accuracy on Validation Set: 0.988764044944	Kernel: poly No. of PCs: n/a No. of LDs: 3 Mean accuracy on 10-fold CV: 0.946111111111 Accuracy on Validation Set: 0.910112359551	No. of PCs: n/a No. of LDs: 4 Mean accuracy on 10-fold CV: 0.954841269841 Accuracy on Validation Set: 0.977528089888
Clinical t-stage	Kernel: poly No. of PCs: 370 No. of LDs: 5 Mean accuracy on 10-fold CV: 0.993218390805 Accuracy on Validation Set: 1.0	No. of PCs: 370 No. of LDs: 7 Mean accuracy on 10-fold CV: 0.996551724138 Accuracy on Validation Set: 1.0	Kernel: poly No. of PCs: n/a No. of LDs: 5 Mean accuracy on 10-fold CV: mean: 0.979655172414 Accuracy on Validation Set: 0.973333333333	No. of PCs: n/a No. of LDs: 7 Mean accuracy on 10-fold CV: 0.996666666667 Accuracy on Validation Set: 0.986666666667
Tumor Recurrence			Kernel: rbf No. of PCs: 370 No. of LDs:2 Mean accuracy on 10-fold CV: 0.99 Accuracy on Validation Set:	No. of PCs: 370 No. of LDs:2 Mean accuracy on 10-fold CV: 0.99 Accuracy on Validation Set:
Disease-free status			Kernel: poly No. of PCs: n/a No. of LDs: 2 Mean accuracy on 10-fold CV: 0.995 Accuracy on Validation Set: 1.0	No. of PCs: n/a No. of LDs:2 Mean accuracy on 10-fold CV: 0.99 Accuracy on Validation Set: 1.0

According to the result analysis table, for Gleason score, Random Forest classifier with PCA followed by LDA gives the highest accuracy score of 0.99 on validation set. On the other hand, both SVM with polynomial kernel and Random Forest classifier give accuracy score of 1.0 in classifying clinical t-stage. In case of predicting tumor recurrence, Random Forest with LDA gives 99 percent accuracy on validation set. However, both SVM with polynomial kernel and Random Forest give 100 percent validation accuracy in predicting disease-free survival status of patients based on gene expressions. This result analysis table demonstrates that although both SVM and Random Forest classifiers provide almost similar performance, still Random Forest outperforms SVM in most of the cases.

TABLE I

CONFUSION MATRIX FOR SVM ON GLEASON SCORE

True/Predicted	Class 0	Class 1	Class 2	Class 3	Class 4
Class 0	8	0	0	0	0
Class 1	0	46	0	0	0
Class 2	0	1	12	0	0
Class 3	0	1	0	19	0
Class 4	0	1	0	0	1

1) SVM Gleason Score: The Confusion Matrix for the classifier is shown in fig I. Here the classes are Gleason score 6,7,8,9 and 10. The diagonal elements in the confusion matrix defines the number of times each class being classified as true classes. According to the matrix, SVM model does a good job in classifying the different labels of Gleason Score based on gene expressions.

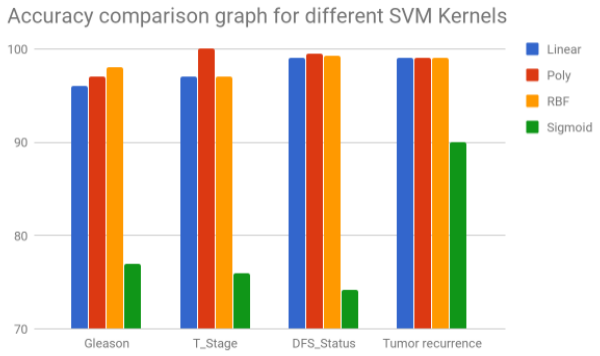

Fig. 13. SVM kernel Accuracies

Fig. 13 is the graph generated from accuracies for different targets using four SVM kernels, Linear, Polynomial, RBF and Sigmoid.

2) RF Gleason Score: The Confusion Matrix for the classifier is shown in fig II. The diagonal elements in the confusion matrix defines the number of times each class being classified as true classes. According to the matrix, RF model does a good job in classifying the different labels of Gleason Score based on gene expressions.

3) SVM Clinical t-stage: The Confusion Matrix for the classifier is shown in fig IV. The diagonal elements in the confusion matrix defines the number of times each class being classified as true classes. According to the matrix, SVM model does a good job in classifying the different labels of t-stage based on gene expressions.

TABLE II

CONFUSION MATRIX FOR RF ON GLEASON SCORE

True/Predicted	Class 0	Class 1	Class 2	Class 3	Class 4
Class 0	5	0	0	0	0
Class 1	0	52	0	0	0
Class 2	0	1	10	0	0
Class 3	0	1	0	19	0
Class 4	0	0	0	0	1

TABLE III

CONFUSION MATRIX FOR SVM ON NEW TUMOR EVENT AFTER INITIAL TREATMENT

True/Predicted	Class 0	Class 1
Class 0	25	0
Class 1	1	2

4) RF Clinical t-stage: The Confusion Matrix for the classifier is shown in fig **??**. The diagonal elements in the confusion matrix defines the number of times each class being classified as true classes. According to the matrix, RF model does a good job in classifying the different labels of t-stage based on gene expressions.

5) SVM Cancer Recurrence: The Confusion Matrix for the classifier is shown in fig III. The diagonal elements in the confusion matrix defines the number of times each class being classified as true classes. According to the matrix, SVM model does a good job in classifying the different labels of New tumor event after initial treatment on gene expressions.

6) SVM DFS Status: The Confusion Matrix for the classifier is shown in fig VI. The diagonal elements in the confusion matrix defines the number of times each class being classified as true classes.

TABLE IV

CONFUSION MATRIX FOR SVM POLYNOMIAL KERNEL ON CLINICAL TSTAGE

True/Pred	0	1	2	3	4	5	6	7
0	1	0	0	0	0	0	0	0
1	0	37	0	0	0	0	0	0
2	0	0	2	0	0	0	0	0
3	0	0	0	7	0	0	0	0
4	0	0	0	0	7	0	0	0
5	0	0	0	0	0	10	0	0
6	0	0	0	0	0	0	9	0
7	0	0	0	0	1	0	0	0

TABLE V

CONFUSION MATRIX FOR RF ON CLINICAL TSTAGE

True/Pred	0	1	2	3	4	5	6	7
0	1	0	0	0	0	0	0	0
1	0	36	0	0	0	0	0	0
2	0	0	3	0	0	0	0	0
3	0	0	0	10	0	0	0	0
4	0	0	0	0	8	0	0	0
5	0	0	0	0	0	10	0	0
6	0	0	0	0	0	0	3	0
7	0	0	0	0	0	0	0	4

TABLE VI

CONFUSION MATRIX FOR SVM ON DFS STATUS

True/Predicted	Class 0	Class 1
Class 0	35	0
Class 1	8	0

According to the matrix, SVM model does a good job in classifying the different labels of New tumor event after initial treatment on gene expressions.

Other Methods: The dataset with 10 most relevant gene expressions along with other label encoded clinical data were run through both SVM and it is concluded that the other approach did not produce expected result. The Accuracy score of each classifier was measured using validation data and SVM produced an accuracy of 82%, Random Forest produced an accuracy of 83%.

V. CONCLUSION

Being a very common yet difficult to diagnose early and properly has made prostate cancer an ideal sector to apply machine learning and pattern recognition techniques. Although numerous quality researches are conducted on prostate cancer, the level of accuracy obtained using only gene expression features was not satisfactory. In this proposed method, Gleason score, clinical t-stage, tumor recurrence and disease-free survival status are classified and predicted using the most relevant features out of thousands of gene features. The accuracy level is very satisfactory and the computational speed and power is also reduced to great extent. In future, this model can be extended to decide accurate treatment for individual patients using only their genes which will allow doctors to diagnose and prognosis prostate cancer at very early stages.

REFERENCES

[1] 1. Siegel RL, Miller KD and Jemal A (2016) Cancer statistics, 2016. CA Cancer J Clin 66, 730.

[2] Rubicz, R., Zhao, S., Wright, J. L., Coleman, I., Grasso, C., Geybels, M. S., Leonardson, A., Kolb, S., April, C., Bibikova, M., Troyer, D., Lance, R., Lin, D. W., Ostrander, E. A., Nelson, P. S., Fan, J.-B., Feng, Z. and Stanford, J. L. (2017), Gene expression panel predicts metastatic-lethal prostate cancer outcomes in men diagnosed with clinically localized prostate cancer. Mol Oncol, 11: 140150. doi:10.1002/1878-0261.12014.

[3] .Prostate Cancer Stages. (n.d.). Retrieved December 13, 2017, from https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/staging.html.

[4] Bahnsen, A. C. (2016, November 11). How to Use Isolation Forests for Anomaly Detection. Retrieved December 13, 2017, from https://insidebigdata.com/2016/11/11/how-to-use-isolation-forests-for-anomaly-detection/.

[5] Finding Transcripts Associated with Prostate Cancer Gleason Stages using Next Generation Sequencing and Machine Learning Techniques