Prediction of Gleason Score, T-Stage and Recurrence Indicator based on RNA Sequence dataset

H. Sodiwala†, H. Sonetta†, S. Sondhi†
†School of Computer Science, University of Windsor, ON, Canada

***Abstract*—Prostate cancer, a carcinogenic disease affecting the prostate gland in the male reproductive system, is the leading cause of cancer death in men. The cancer begins when healthy cells in the prostate gland change and start to grow out of control, forming a tumour. Prostate cancer is somewhat unusual when compared with other types of cancer. This is because the cancer grows very slowly and may not cause symptoms or problems for years or ever. Thus, prognosis and diagnosis of this disease in early stage is a big challenge. To overcome this challenge, we use biomarkers to detect the presence and progression of cancer in the body. In this work, we utilize the cBioPortal Prostate Adenocarcinoma (TCGA, Provisional) dataset from the National Cancer Institutes (NCIs) Genomic Data Commons (GDC) [1]. This dataset consists of 496 patient reports having 94 clinical features and over 20,533 gene expressions. Before applying various classification algorithms on the dataset, we reduced the dimension of the dataset by removing the low variant features using VarianceThreshold technique. Feature selection is the most important aspect of data preparation in the machine learning pipeline, for which we used a tree-based feature selection technique. The data had a huge class imbalance problem which was tackled using oversampling and undersampling techniques. Later, applying PCA and LDA on the selected features helped us to visualize the dataset in 3 dimensions. The selected features were used for classification of T-stage, Gleason score and Tumor recurrence using K-NN, RandomForest Classifier, Naive Bayes and Support Vector Machine (SVM). Classification algorithms are applied on the dataset using cross validation methodology. The performance of these algorithms is evaluated on the basis of accuracy, providing a best accuracy of around 95 to 98 percent which is a huge improvement in terms of performance and the computational speed.**

***Index Terms*—**

I. Introduction

Prostate cancer is the second leading cause of cancer death in men, second only to skin cancer. In 2019, ASCO estimated 174,650 new cases of invasive prostate cancer will be diagnosed in men[2]. Looking at such an alarming number, we developed a methodology using various machine learning and pattern recognition techniques to detect the presence and progression of the disease. This was the motivation behind our study. While doing our literature review over this subject we observed that the traditional research and work done was on detection of benign and malignant tumours. However, recent studies are focused on development of novel computational tools for stratification, grading, and prognostication of patients. Doing this accurately, requires gene expression data of individual patients. To give an idea of the complexity of such a dataset, it consists of thousands of gene expressions which have a sparse representation. Although gene expressions are really useful as biomarkers to detect the presence of such carcinogenic diseases, they only have significance if a relevant subset of genes are taken into consideration for modelling the machine learning algorithms. In our study, we have focused on classification and prediction of the most significant target features which give insights on the characteristics of the tumour in the patient. The target features include classification of Gleason score, pathological T-stage and Biometric Recurrence Indicator. Gleason Score is the grading system used to determine the aggressiveness of prostate cancer[3]. This grading system can be used to

choose appropriate treatment options. The Gleason Score ranges from 6-10 (10 being the most aggressive). Pathological T-stage describes how much of the prostate contains cancer and if doctors can feel the cancer or not. It also indicates whether the tumour has grown outside of the prostate to the surrounding tissues. It is usually given as a number from 1 to 4 - a higher number means that the tumour takes up more of the prostate, or that the tumour has grown outside of the prostate into nearby tissues[4]. Biometric Recurrence Indicator tells us if the cancer is found after treatment, and after a period of time when the cancer could not be detected. Our study includes classification of target features purely based on the subset of gene expressions. The subsets are curated using various preprocessing and feature selection techniques. The accuracy of our model ranges from 94% to 98% for respective features. The feature subset of 70 genes are selected out of 20,500 features using tree-based method. Feature selection was done after eliminating redundant or least informative features by using the variance thresholding technique which removes the features having close to zero or no variance. Later, only those features were retained which had a high correlation with the target feature. To tackle the curse of dimensionality, we used various dimensionality reduction techniques like PCA and LDA which helped to reduce the dimensions of the data while retaining the relevant information, this helped in the visualization of our data or more specifically the distribution of classes within each target feature. Dimensionality reduction techniques were leveraged to minimize the intra-class separation and to maximize the between class separation. After selecting the most relevant features from the gene expression dataset we performed classification using Support Vector Machine, Random Forest classifier, Naive Bayes classifier and K-NN algorithm. The performance of each of the classification algorithms is evaluated using 10-fold cross validation technique.

II. METHODOLOGY

The entire research is divided into five main stages, namely Understanding the dataset, Preparing the data, 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 The Dataset

We have utilized the cBioPortal Prostate Adenocarcinoma (TCGA, Provisional) dataset from the National Cancer Institutes (NCIs) Genomic Data Commons (GDC) for this study. The dataset consists of two main types of files - one pertains to the clinical variables of 496 patients while the other has the gene expressions for each patient in the range of 16,000 to 24,000 features. There were several gene expression files and each gives unique information for prediction or classification of a certain clinical feature. It was necessary to understand which clinical features hold strong relation with corresponding gene expression files. So, a comparative study was done by finding the strongest correlation between the gene expression files and the target features. For example, to classify the T- stage feature we found correlation between the target feature and the gene expressions from data_methylation_hm450 and data_RNA_Seq_v2_expression_median file. Also, we got an overview of the dataset by plotting various clinical features. The following plot shows the diagnosis age against the count of patients on the y-axis.

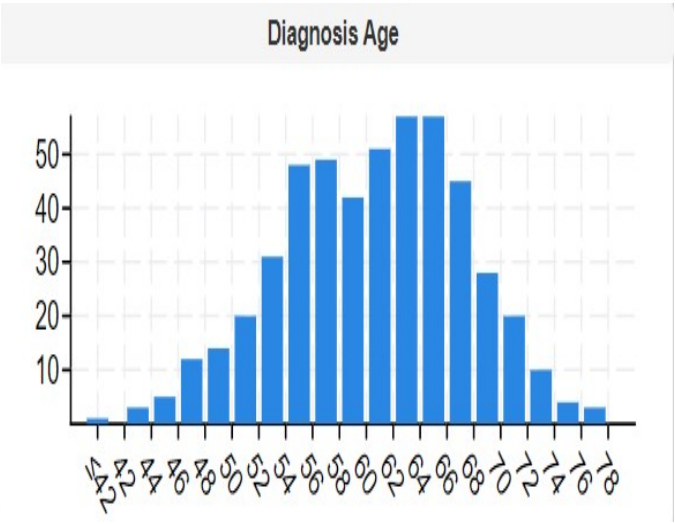

Fig. 1. Distribution of the diagnosis age

The average age of diagnosis for men is around 60. But it is likely that men are at higher risk of

developing prostate cancer by the age of 45. To get a good view of the distribution of classes within our target feature we created visualisations for each of our target features. The following pie chart shows the distribution of classes for T-stage feature

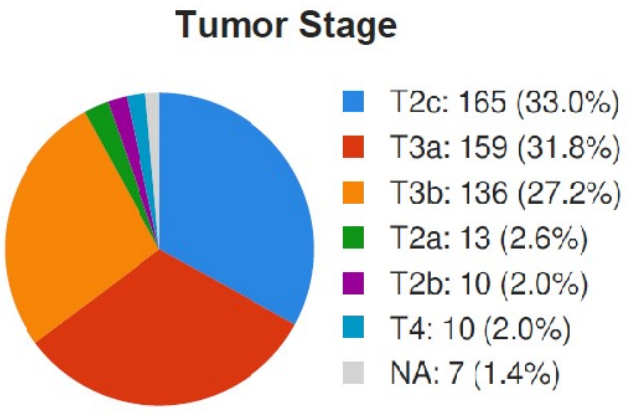

Fig. 2. Distribution of the tumor stage

T-stage consists of 4 categories for describing the local extent of a prostate tumor, ranging from T1 to T4, each having subcategories. The samples having 'NA' values were discarded as it accounted for a very small portion of the data - close to only a percent. Gleason score consists of five classes ranging from 6 to 10. The class distribution is dominated by category 7, which is found in 250 samples out of the 496.

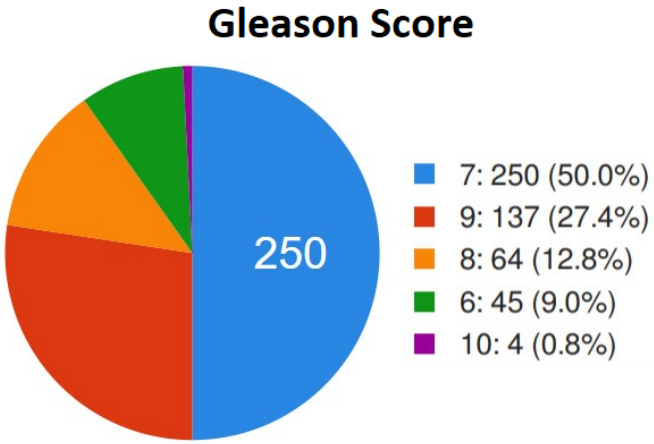

Fig. 3. Distribution of Gleason Score

Figure 4 shows the distribution of classes for Recurrence Indicator target feature: This target feature consists of two values, the class distribution is dominated by class 'YES' which is found in 373 samples. This feature also had NA values (13%), which we discarded as no useful information could be gained from them. Thus for each target feature we created a separate data file having the relevant gene expression as our independent features and target columns as our dependent features.

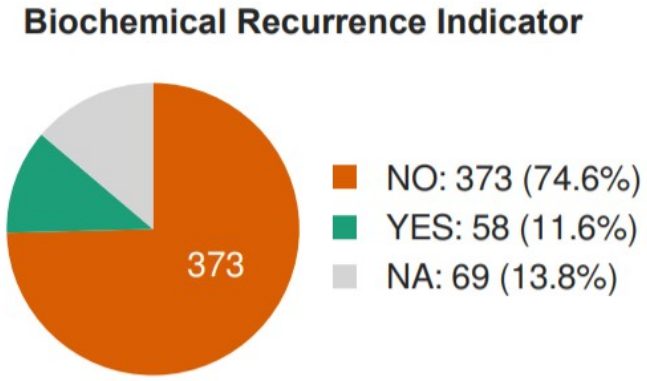

Fig. 4. Distribution of Gleason Score

B. Preparing The Dataset

Removing samples with missing target

For each target variable there are a few samples in the dataset where the target variables are missing. These records will not be helpful in the training or prediction testing. Hence, they are removed before taking any further processing step.

Imputing missing values

Amongst the filtered samples, there exist some, where the values of some features are missing. These values are imputed. As all the features have numerical values, the missing values are replaced by the average value of the respective feature.

Scaling the dataset

The dataset contains several features whose values lies in varying ranges. This nature of the dataset might have misleading effects on the classification stage of the pipeline, which may result in unfair weight assignment to the features and eventually

bad classification.
To mitigate this problem, the data is scaled. That is, the feature vectors are modified such that the values of each feature are on the same scale and at the same time retain the original data.

Removing outliers
This is the step to remove the anomalies or outliers from the dataset. They are the samples that are significantly different from the mass of samples having the same label. They can be safely considered as exceptions and hence can be removed from the dataset in order to prevent them from unfairly affecting the training of the model. This is achieved using a method called *IsolationForest*. It is an ensemble method which explicitly identifies the outliers rather than profiling normal points in the dataset.
This technique is applied for Gleason Score classification model as it has them in a notable amount. Figure 5 and 6 shows the principle component plot of the dataset before and after outlier removal ,respectively.

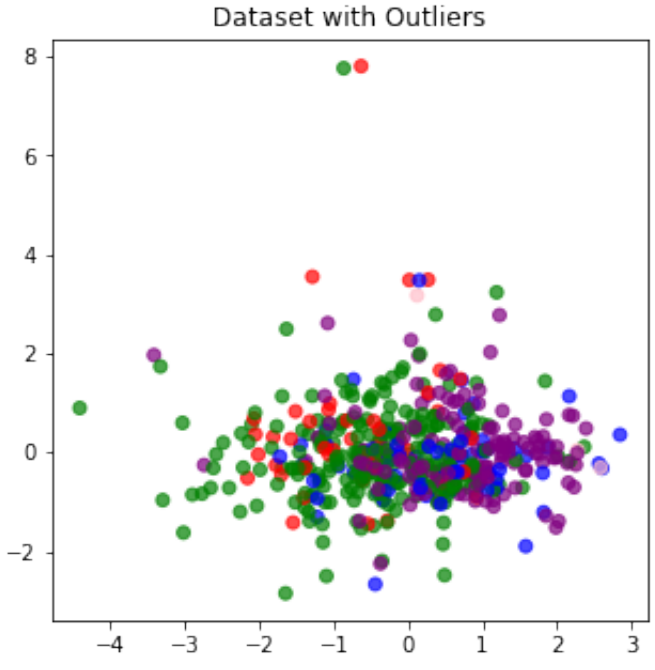

Fig. 5. Gleason Score plot with outliers present

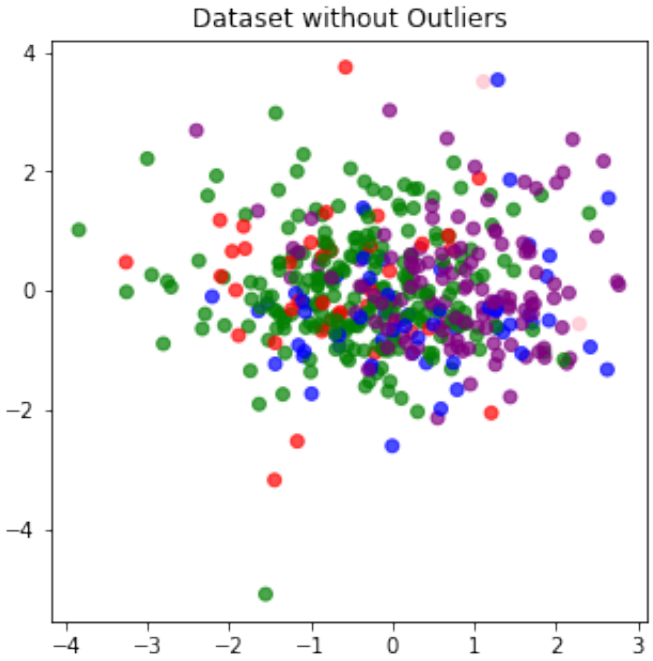

Fig. 6. Gleason Score plot after removal of outliers

C. Feature Selection

Select K Best This is the first step towards feature selection. Here the features are filtered based on the Chi-Square test score of each feature. It is a univariate feature selection technique. The Chi-square test measures the likelihood of a given distribution and selects K best features that are more likely to predict the target class. Hence, this is based on how highly a feature is related to the target variable.

Variance Threshold In this step the variance of each feature is calculated along the samples, which shows how much information (by diversity) is contained in the feature. After calculating the variance, hence the information metric, the features with variance above a certain threshold are selected. This ensures that we do not pass those features to the model for training that would not contribute towards prediction of the target variable.

Selection using Classification Model In this feature selection method, an Ensemble Tree based model is used to find the importance of a feature and thereafter select the features with highest importance. In this project Random Forest Classifier is used as the Ensemble Tree model to calculate this importance value.

D. Dimensionality Reduction

After removing the features that are less useful for training and classification, the next step is to reduce the dimensionality of the remaining dataset to further narrow it down to a space with lower number of dimensions, in which the data is fairly separable. This is to reduce the complexity of the training model in terms of dimensions and thus minimize the computational challenge in training the Machine Learning model In this project the following methods are used to perform dimensionality reduction

Principle Component Analysis This is the first step taken for the dimensionality reduction process. PCA is an unsupervised technique to reduce the dimensions of the data. It does the reduction by mapping the given data points to a lower dimensional space in such a way that they are fairly separable. Hence, it finds what are called the Principal Components, that is, the components that best represent the given data in the required number of dimensions.

Linear Discriminant Analysis After applying PCA a supervised technique of dimensionality reduction is used, which is LDA (Linear Discriminant Analysis). Using the class labels of the given data samples, LDA plots them into a space with given number of dimensions, in such a way that the classes have maximum inter-class separation and minimum intra-class variance. This means that LDA finds a Linear Discriminant Function that separates the classes as much as possible so that the training model can be easily tuned to classify. Figure 7, figure 8 and figure 9 show the 3D (reduced using PCA) plot of the dataset after dimsnesinality reduction for T-Stage, Recurrence Indicator and Gleason Score respectively.

E. Balancing Data

For Gleason score and Recurrence Indicator, the samples are highly imbalanced. This means that the number of samples of one or more classes are dominant and hence causes the model to fit in such a way that it can only classify the test samples of the dominant class. To overcome this problem, the

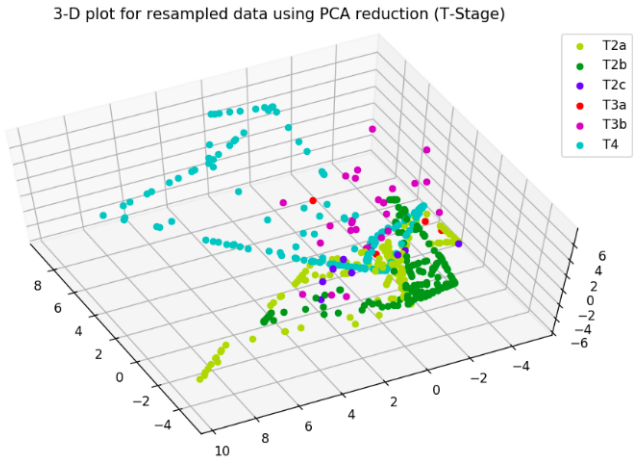

Fig. 7. T-Stage plot after dimensionslity reduction

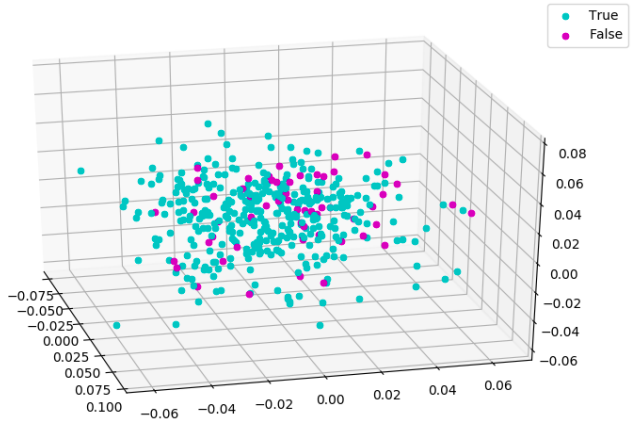

Fig. 8. Recurrence Indicator plot after dimensionslity reduction

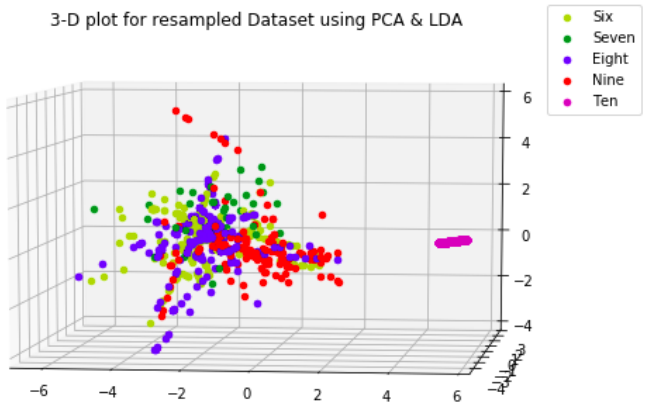

Fig. 9. Gleason Score plot after dimensionslity reduction

technique of data balancing is used which is done by reducing the sample of the dominant class and generate synthetic sample of the minority class In this project, a pipeline of OverSampling model and DownSampling is used. The SMOTEEN library of python is leveraged to balance the dataset.

F. Classification

After preparing the data, selecting relevant features and following a systematic dimensionality reduction procedure, the next step is to train various machine learning models and make a comparative study of how each model performs for each target feature.

In this project the following Machine Learning models are implemented:

1) Support Vector Machines
2) Random Forest Classifier
3) Naive Bayes Classifier
4) K Nearest Neighbours

Support Vector Machines

SVM is a classifier that works on the principle of finding a hyperplane performing a binary classification. It is a non-probabilistic classification model. The major advantage of an SVM model

is that it is very suited to carrying out non-linear classification. This is possible with the use of kernels. A kernel function is used to map the given data points into a higher dimension where they are linearly separable and then form the hyperplane in that space.

The following are some of the popular choices for kernel function with Support Vector Machines

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

Fig. 10. RBF Kernel

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

Fig. 11. Polynomial Kernel

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

Fig. 12. Sigmoid Kernel

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

Fig. 13. Linear Kernel

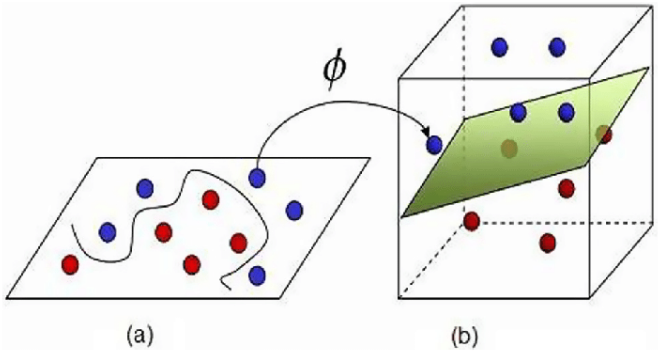

Fig. 14. SVM using Kernel Function

In this project, RBF kernel is used for all the target variables as it is the most suitable kernel, giving highest accuracy.

Random Forest

Random forest is an ensemble model that uses multiple decision trees to make a decision and selects a weighted average as its final prediction. This way it mitigates the problem of overfitting a significant amount. This is achieved by generating several decision trees that use different split criteria based on Gini indexes of different features. This way feature selection indirectly becomes an integral part of the classification process of Random Forests. It is a very simple, but extremely powerful model that works well with both categorical as well as numerical values.

Naive Bayes

Unlike other classifiers used in this project, Naive Bayes is a probabilistic classifier. It works on the principle of Bayes Theorem. It accounts for the prior information we have about the target variable, and calculates the posterior probability distribution to perform the classification as follows:

.

$$\hat{y} = argmax\, p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

K Nearest Neighbours

KNN is a supervised classification model. It does classification by comparing the given data points with its K nearest neighbours and assigns the labels that are in a majority amongst these neighbours. It is one of the simplest classification models. Simply put, it assumes that a sample can be classified based upon the company it keeps.

III. Results

Each classifier was run on the reduced data to make predictions on the three target features that we selected. The results of classification accuracy is shown in the following tables

TABLE I
GLEASON SCORE

SVM	98.75%
Random Forest	69.38%
Naive Bayes	44.89%
KNN	57.14%

TABLE II
T-STAGE

SVM	97.942%
Random Forest	92.34%
Naive Bayes	94.96%
KNN	97.94%

TABLE III
RECURRENCE INDICATOR

SVM	73%
Random Forest	94%
Naive Bayes	83%
KNN	83%

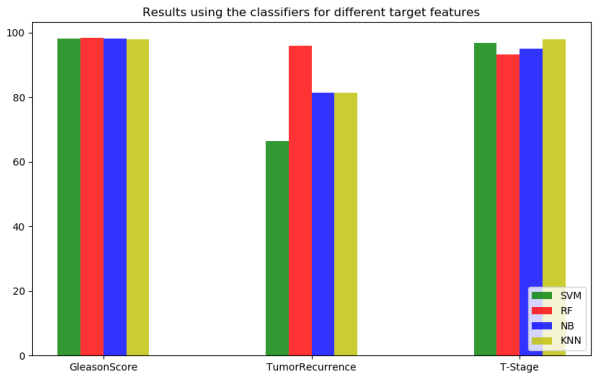

Fig. 15. Comparing the results

Most of our models gave a good accuracy on all target features. In general, the Random Forest classifier performed the best in predicting Gleason score, tumor recurrence and the pathological t-stage. The K-nearest neighbours and Naive Bayes' classifiers also gave acceptable accuracy on all tasks. The support vector machine classifier however gave mixed results. It performs exceptionally well while predicting the Gleason score and the t-stage but drops considerably in accuracy while making predictions on tumor recurrence.

Note as well, that the tumor recurrence was the most challenging task for us to predict, owing to the huge class imbalance in the data and the binary nature of the feature. Further, the presence of Nan values as well, reduced the amount of data we had to work with. However, making certain adjustments in the feature selection process - like using the random forest classifier during feature selection and using more than 2 trees at this stage - affected the eventual performance of every classifier on this feature. To make up for the class imbalance we used the SMOTE technique to resample our data. Since the Gleason score had

imbalanced classes as well, we applied the same technique on the data before making predictions for this as well.

For the KNN classifier, we plotted the accuracy of the model with varying values for the number of neighbours. The plot in Fig. 16 shows the changes in accuracy as k was tweaked from low to high. This helped us decide the optimum value for k while making predictions on all features.

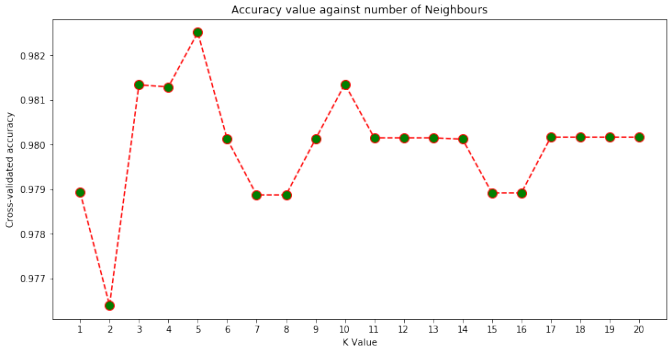

Fig. 16. Effect of varying neighbors on accuracy

The plot in Fig.11, shows the effect of changing the number of neighbors while making predictions on the Gleason score feature.

As a control experiment, we also ran our four classifiers on the original data, without applying any dimensionality reduction or feature selection methods. This did not give us respectable results with a maximum accuracy of 43% amongst all models. The results on all targets were mostly comparable. We also ran our classifiers on the data extracted after applying basic feature selection methods like only variance thresholding. While this did not significantly improve the performance, it gave mixed results with some models performing significantly worse than the others. Still, our most performant model in this case was the Random Forest classifier which gave an accuracy of 64%. We deemed these results unacceptable and hence conclude that our more advanced feature selection methods, along with dimensionality reduction techniques considerably boosted performance.

IV. Team Member Roles

Exploratory data analysis has been done as a combined task by all the three team members. All the members analysed the data, it's features and potential target variables. Finally, a systematic approach is designed to carry out for classification of each target variable. The members worked on

one target variable each which includes data cleaning, feature selection and classification of each target variable. Implementation of target

variables are done by each of the members as follows

1) Gleason Score - Hardik Sonetta
2) Recurrence Indicator - Harsh Sodiwala
3) T-Stage - Shiv Sondhi

V. Conclusion

As discussed previously, prostrate cancer is one of the leading causes of cancer death amongst males. With the huge advancements in machine intelligence this past decade, it is only natural to want to take advantage of the huge potential of such systems and apply them where it matters. Using such machine learning techniques as ours, in the real world, will considerably reduce the pressure on doctors and nurses. Moreover it barely takes any time for such a model to make predictions once it has been trained and may prove to be more cost-efficient than existing systems. In light of these matters, we would like to give our views on a few societal and ethical impacts that models such as ours can have on the real-world. There are a few concerns with using machines to make decisions in such delicate matters. A few of the considerations are listed out below.

- Who is to blame for incorrect predictions?
- What is the best way to introduce the model into the real world of cancer detection.
- Who is going to make sense of the models' results - a doctor or a statistician?
- What legal measures can be taken in the event of miss-classification? Against whom, and when?

These are just a few of the things we must think about before we let our intelligent machines out into the real world to make predictions. In cases, especially such as ours, where human lives are in the balance, we must be more careful than ever. We believe that a good way forward is to create

partnerships between medical professionals as well as data scientists. This way, the responsibilities can be shared and it can allow for better understanding of models and their predictions. The model could first be introduced as a shadow system to the current one. This would ease machine learning and artificial intelligence into the workplace and allow enough time for the people involved - like doctors and nurses - to get used to having these systems around. This can also act as a ”test drive” for the models and prevent them from making decisions that have any direct impact on human lives. Once again, it may be necessary to take extra-measures like hiding results from the operating doctor to prevent biases from developing on a case-to-case basis. Finally, the legal ramifications of these models must be graded according to severity. For example, in our case, false negatives must not be tolerated. False positives on the other hand, are not great either since they can lead to extra costs and possibly trauma or other health-related issues. But the legal punishments for these two outcomes must be differentiated.

Systems such as ours are already in use in certain places. However, most medical institutions have been reluctant to open their doors to intelligent machines - especially when they directly affect human lives. We hope that this can change in the near future and we can find a way to make machines an acceptable norm in hospitals; because these machine models hold the potential to save lives.

REFERENCES

1 cBioPortal, “Prostrateadenocarcinomacancerdatasummary,” 2018. [Online]. Available: https://www.cbioportal.org/study/summary?id=prad_tcga

2 ASCO, “Asco 2019: Reducing rt treatment costs for advanced rectal and prostate cancer,” 2019. [Online]. Available: https://www.appliedradiationoncology.com/articles/asco-2019-reducing-rt-treatment-costs-for-advanced-rectal-and-prostate-cancer

3 Prostrate-Conditions-Education-Council, “What is gleason score?” [Online]. Available: https://www.prostateconditions.org/about-prostate-conditions/prostate-cancer/newly-diagnosed/gleason-score

4 Canadian-Cancer-Society, “Stages of prostrate cancer.” [Online]. Available: https://www.cancer.ca/en/cancer-information/cancer-type/prostate/staging/?region=on