A Machine Learning Approach to Prostate Cancer Risk Classification Through Use of RNA Sequencing Data

Matthew Casey[1](✉), Baldwin Chen[1], Jonathan Zhou[2], and Nianjun Zhou[3]

[1] Ardsley High School, Ardsley, NY 10522, USA
mattcasey02@gmail.com, baldwinchen@gmail.com
[2] Horace Greeley High School, Chappaqua, NY 10514, USA
jozhou@students.ccsd.ws
[3] IBM, 1101 Kitchawan Road, Yorktown Heights, NY 10598, USA
jzhou@us.ibm.com

Abstract. Advancements in RNA sequencing technology have made genomic data acquired during sequencing more precise, making models fitted to sequencing data more practical. Previous studies conducted regarding prostate cancer diagnosis have been limited to microarray data, with limited successes. We utilized The Cancer Genome Atlas' (TCGA) prostate cancer sequencing data to test the viability of fitting machine learning models to RNA sequencing data. A major challenge associated with the sequencing data is its high dimensionality. In this research, we addressed two complementary tasks. The first was to identify genes most associated with potential cancer. We started by using the mutual information metric to identify the most significant genes. Furthermore, we applied the Recursive Feature Elimination (RFE) algorithm to reduce the number of genes needed to identify cancer. The second task was to create a classification model to separate potential high-risk patients from the healthy ones. For the second task, we combated the high dimensionality challenge with Principal Component Analysis (PCA). In addition to high dimensionality, another challenge is the imbalanced data set that has a 10:1 class imbalance of cancerous and healthy tissue respectively. To combat this problem, we used the Synthetic Minority Oversampling Technique (SMOTE) to create synthetic observations and equalize the class distribution. We trained and tested a logistic regression model using 5-fold cross-validation. The results were promising, significantly reducing the false negative rate as compared to current diagnostic techniques while still keeping the false positive rate low. The model showed great improvements over previous machine learning attempts to diagnose prostate cancer. Our model could be applied as part of the patient diagnosis pipeline, helping to improve accuracy.

Keywords: Machine learning · RNA sequencing · Prostate cancer diagnosis · Upsampling · Logistic regression · Multi-layer perceptron · Auto-encoding · Mutual information · Recursive feature elimination · The Cancer Genome Atlas

© Springer Nature Switzerland AG 2019
K. Chen et al. (Eds.): BigData 2019, LNCS 11514, pp. 65–79, 2019.
https://doi.org/10.1007/978-3-030-23551-2_5

1 Introduction

Prostate cancer is the second leading cause of cancer death among men in the United States, with an estimated 31,620 deaths predicted in 2019 [1]. The cancer is characterized by a malignant tumor found within the prostate and is mainly found in men 65 and older [1]. Currently, diagnosis begins with a preliminary blood test called a Prostate Specific Antigen (PSA) test. If the patient shows an abnormally high PSA level, a doctor may recommend that they undergo a second PSA test and a Digital Rectal Exam (DRE), in which a doctor palpates the prostate to check for abnormalities implying a tumor. A prostate biopsy may be recommended if a lump is detected or if PSA levels continue to rise[1]. After the biopsy, the cancer is assigned a Gleason score and stage, both of which indicate the severity of the cancer [2].

There is a great need for an accurate early detection method for prostate cancer. Various studies have raised concerns regarding PSA testing and the effects it has on patients. It is estimated that PSA tests have overdiagnosis rates of between 23% and 42% [3]. Often, this can lead to unnecessary anxiety and decreased general health for patients [4]. The results from the PSA tests often prompt patients to get a biopsy done; however, this can lead to more confusion due to the inaccuracy of biopsy-based diagnosis. A study done in 2013 found that the standard 12-core biopsy method of diagnosis is very ineffective [5]. According to the study, Gleason scores were underestimated in 47.8% of patients. In addition, they found that the false negative rate could be 30% or more, meaning that many patients with cancer were diagnosed as healthy. Lastly, they found the detection rate was even lower for patients with lower PSA values. These issues have prompted researchers to seek new methods of diagnosis, one of which is through the use of RNA sequencing data.

RNA sequencing is a gene sequencing technique which gives a more precise view of a cell's transcriptome than previous microarray or Sanger sequencing based methods [6]. DNA microarrays are used to measure gene expression levels; however, they are not as detailed as RNA sequencing is [6]. The data acquired from RNA sequencing is important for cancer classification because certain differences in the transcriptome can indicate the presence of prostate cancer. The sequencing could be performed after the biopsy in addition to slide pathology.

A recent study done by the National Institute of Health compiled RNA sequencing data for 33 different types of cancer. The database, called The Cancer Genome Atlas (TCGA) contains data for both healthy and cancerous samples for each cancer. We selected prostate cancer because of its prevalence in society and current problems with diagnosis. In addition, the prostate cancer dataset has a relatively high number of samples. All of the datasets have many more cancerous samples than healthy samples and have a sample-feature imbalance. This means that the number of features significantly outnumbers the samples, which is common in gene and healthcare research. Thus, picking one of the cancers with a relatively high amount of samples ensures that the results are more robust.

[1] https://www.cancer.gov/types/prostate/psa-fact-sheet#q1.

The purpose of this project was two-fold. The first was to identify genes related to cancer. The second was to produce effective classification models that can outperform standard biopsies and microarray-based models. The costs of genomic sequencing have decreased, making prostate cancer classification using RNA sequencing data a much more practical option to enhance diagnostic techniques [7]. Biopsies do not always result in accurate results, and our model can enhance the results of biopsies, leading to greater predictive accuracy.

The paper is organized as follows. The second section will discuss related works. The third section will discuss problem formulation and data acquisition. The fourth section will address how to identify key gene sequences related to Prostate Cancer using mutual information and recursive feature elimination. In the fifth section, we will develop a binary classification model, utilizing logistic regression to distinguish cancerous and non-cancerous individuals. In the sixth section, we will discuss additional efforts we explored or are currently in progress. In the final section, we will summarize what we have accomplished and future works.

2 Related Work

Various attempts have been made to classify cancer based on tissue samples. These various studies have used microarray, clinical, imaging, and RNA sequencing data. Many recent works have utilized microarray datasets for cancer classification. The most recent study developed a new approach which aimed to improve accuracy when using microarray data for classification [8]. Early studies conducted that attempted to diagnose prostate cancer with machine learning utilized microarray datasets. Various studies were conducted using different methods and were tested on five different microarray datasets [9–14]. These studies aimed to predict whether a cancer would metastasize or not. Although the results of all the microarray based studies are significant, the increased information gain about the transcriptome from RNA sequencing should give way to improved classification accuracy.

A recent study done attempted to diagnose prostate cancer with machine learning through use of clinical data [15]. They trained an Artificial Neural Network (ANN) with data consisting of 22 clinical features. They found that although their model performed well, it needed improvements before being suitable for clinical applications.

The TCGA database has already been used for cancer classification. The data contained in the database goes beyond RNA sequencing data. A recent study used slide images for classification of lung cell cancer through use of a Convolutional Neural Network (CNN) [16]. Their results improved upon those achieved by doctors in manual diagnosis.

Based on our research, no published studies have yet attempted using prostate cancer RNA sequencing data from the TCGA database for cancer classification. However, studies have been published using the breast cancer dataset for classification [17, 18]. The two main challenges associated with all of the TCGA datasets for advanced analytical study are high dimensionality and class imbalance. The high dimensionality poses two problems. Firstly, there are so many features that the models will not be able to accurately separate the data into healthy and sick. Thus, the results of any model trained on the data may be poor. Secondly, the number of features is many times greater than the number of samples. This is known as a feature-sample imbalance, and it causes models trained on the data to be unstable and leads to overfitting. This makes the results of any model trained on the data unreliable. The studies conducted by Danaee et al. [17] and Golcuk et al. [18] using TCGA breast cancer tested many methods to combat these problems.

Danaee et al. tried various methods of dimensionality reduction such as a Stacked Denoising Autoencoder (SDAE), differentially expressed genes, PCA, and KPCA. They tried each of these methods with three different models, an Artificial Neural Network (ANN), a Support-Vector Machine (SVM) with a linear kernel, and an SVM with a radial basis function kernel (SVM-RBF). They calculated five metrics for each model: accuracy, sensitivity, specificity, precision, and f-measure. They found that the highest accuracy was attained using the SDAE for dimensionality reduction followed by the SVM-RBF model. This method also had the highest F-measure. The highest sensitivity was achieved with the SDAE as well, but with the ANN model. The KPCA with SVM-RBF model attained the highest specificity and precision.

Golcuk et al. conducted a study which aimed to improve upon the results achieved by Danaee et al. As a baseline they tried three dimensionality reduction algorithms (PCA, KPCA, and NMF), followed by a SVM. They also tried utilizing a ladder network, which does not require a reduction in dimensionality. They found that the ladder network slightly outperformed both the SDAE and SVM models from the previous study in almost all metrics. The only metric in which it performed worse was specificity, showing a slight decrease as compared with the KPCA and SVM-RBF model.

Only one of these studies however dealt with the class imbalance problem. The first study utilized the same SMOTE technique that we will use in this study to increase the number of samples in the dataset. The second study, however, failed to address the class imbalance. Both their test and validation sets have only 20% of the data, and with such a low number of healthy samples in the dataset already, each of these datasets had very few healthy samples. As a result, the results of the model could have been inflated and makes their results less reliable than those of the first study.

Both of these studies used only the gene expression data and neglected to use the other three datasets obtained from RNA sequencing. The other datasets present a significant source of information that could help to improve the performance of models.

3 Data Acquisition and Preprocessing

In this section, we will discuss how we acquire the genomic datasets from the TCGA database and the preprocessing method to combine the datasets based on shared IDs. Furthermore, we will boost our dataset using the Synthetic Minority Oversampling Technique (SMOTE). The purpose of doing the preprocessing is to prepare the data for use in the gene selection and classification models, and to combat model instability issues.

3.1 Data Acquisition

In order to download the RNA sequencing data from the TCGA database, we used an open-source tool called TCGA-Assembler 2 [19]. All four main RNA sequencing datasets were downloaded as well as the clinical data. Various types of genomic data were included: Exon expression, Exon Junctions, Isoform expression, and Gene expression data. Exons, or sections of genes that provide the code needed to create proteins, as opposed to introns, which are designed to not code for anything. Exon data in the TCGA data set displays the positions of the exons on the individual chromosomes based off of distance from the ends of the chromosomes. It also shows the expression levels of each of the exons. Exon Junctions, or the positions where two exons meet, show mutations occurring when the individual exons are combined to form a single pre-mRNA chain and allow for scientists to observe similarities in specific mutations in a specific area that are common to all cancerous patients [8]. Isoforms, genes that serve almost identical purposes, but are composed of different exons in different orders or completely different bases, show the genes that are similar to each other. Lastly, gene expression data shows the exact levels at which a gene is expressed, allowing researchers to identify genes that are common to cancerous patients and those that are common to healthy individuals.

After downloading the data, the built-in processing functions were used to clean up the raw data. These functions extract the most useful parts of the data for analysis. For instance, for gene expression data, normalized count values are extracted, and for exon expression data, RPKM (Reads Per Kilobase of transcript, per Million mapped reads) values are extracted. These values are selected because they are comparable from sample to sample, unlike the raw data, making them far more useful for analysis. Each dataset was outputted in a tab-delimited text file and was used later in our own preprocessing. At this point, the high dimensionality of the data becomes very clear, with nearly 600,000 total features across the four datasets (Table 1).

Table 1. Number of features for each dataset

	Gene	Exon Junction	Isoform	Exon
Feature count	20531	249566	73598	239321

3.2 Preprocessing of RNA Sequencing Data

The four different data types arising from the RNA sequencing were preprocessed separately due to slight differences in structure. For each data type, the data was first formatted so that the index was the sample ID and the columns were the features. Then, the samples were categorized as either 0 for cancerous or 1 for cancer free, by extracting the 13th and 14th digits of the sample ID. A few samples were originally categorized as 6, or metastatic, and those were reclassified to be 0.

For the gene quantification and isoform quantification datasets, each gene/isoform had two columns, one corresponding to the raw count per transcript and the other corresponding to a scaled value which was independent of transcript length. In order to make more accurate comparisons of gene and isoform expression between samples, the raw count column was dropped for each gene/isoform. In addition, the scaled values were multiplied by one million to convert them into a Transcripts Per Million (TPM) value. Following completion of individual preprocessing, the data were merged into one larger dataset. Since all of the data came from the same sequencing process, each sample had data in each dataset, and the merge was done using the sample ID as the reference.

Table 2. Sample gene data

Patient ID	UNK-100130426	UNK-100133144	UNK-100134869	UNK-10357	UNK-10431
TCGA-2A-A8VL-01	0.0	1.294659	0.788167	11.358769	85.259114
TCGA-2A-A8VO-01	0.0	1.121938	0.593362	8.164714	52.753502

Due to space restrictions, we display only a snapshot of one of the four datasets (Table 2). This dataset contains 20531 genes and 546 patients, however we only show five genes and two patients. The values in the table represent the expression of the gene.

3.3 Upsampling with Synthetic Minority Oversampling Technique

Before any testing could be done using machine learning models, we first upsampled the data in order to deal with model instability resulting from the 10:1 class imbalance between cancerous and normal (minority) cells. The Synthetic Minority Oversampling Technique (SMOTE) is a method in which synthetic data of the minority class is created which closely resembles the original data [20]. In order to do this, it creates a new observation randomly on the imaginary line connecting an existing data point with the data point closest to it. As a result, the data stays in the same general cluster, but the amount of samples is increased. We increased the number of samples from 496 cancerous and 50 healthy to 3000 cancerous and 1500 healthy. By doing this, we increased the robustness of the results of the classifiers since more healthy samples are contained in each test part of the test/train split. Without this method, there was a very low amount of healthy samples in the test set, which may have artificially inflated the accuracy.

4 Key Gene Sequence Identification

In this section, we will first identify a set of genes related to Prostate Cancer using the techniques of mutual information and Recursive Feature Elimination. The mutual information is used to give a preliminary measure of each gene's importance. The RFE is then used to minimize the gene sequences needed to perform diagnostics and predictions.

4.1 Gene Sequence Selection Using Mutual Information

The first step in the feature selection process was the identification of genes that are significant in determining whether a cell is cancerous. We used the mutual information metric to determine the most significant genes. Mutual information is a metric which quantifies the amount of information gained about one variable by observing the other. High mutual information between a gene and the target would mean that knowing the expression of the gene would give the model a good indication of whether the sample is cancerous. Equation 1 defines the mutual information between two discrete variables X and Y and was used to calculate the mutual information between each gene and the target (cancerous or healthy).

$$I(X;Y) = \sum\nolimits_{y \in \mathcal{Y}} \sum\nolimits_{x \in \mathcal{X}} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{1}$$

The results of the mutual information are summarized in Table 3.

Table 3. Mutual information rankings

Gene ID	Mutual information
Gene_POLR2H-5437	0.167398
Gene_GSTM4-2948	0.163962
Gene_APOBEC3C-27350	0.161822
Gene_HPN-3249	0.159367
Gene_ETNK2-55224	0.154814
Gene_ANGPT1-284	0.154056
Gene_GSTP1-2950	0.153113
Gene_LURAP1-541468	0.151971
Gene_TMLHE-55217	0.151118
Gene_MCF2-4168	0.149987
Gene_SLC19A1-6573	0.149467
Gene_NKX2-3-159296	0.148963
Gene_PYCR1-5831	0.148119
Gene_PLP2-5355	0.146773
Gene_HOXC6-3223	0.145437
Gene_EFNB1-1947	0.144863
Gene_NKAPL-222698	0.144755
Gene_MARCKSL1-65108	0.14449
Gene_ASPA-443	0.144207
Gene_NECAB1-64168	0.143234

Table 3 contains the top twenty genes with the highest mutual information values. The meaning of these genes can be found on the National Center for Biotechnology Information (NCBI) database. We examined several genes from Table 3 using the NCBI database and noticed some of those genes are theorized to be related with cancers.

4.2 Feature Selection with Recursive Feature Elimination

In order to improve the performance of a model trained on the selected features, feature selection through Recursive Feature Elimination (RFE) was employed on the gene quantification data. The advantage of this method over purely using the genes with the highest mutual information is that it will pick genes that are not related to each other. Two genes may have very high mutual information values but if they are highly correlated to each other, dependent on the same hidden variable, or are otherwise related, one of them is redundant when doing classification. With RFE, the chosen genes are not related, and therefore the variance of the data is better represented. Using RFE, the highest ranked genes selected will give the highest accuracy when classifying a sample. A further advantage of this method is that because only a few genes need to be used when doing classification, the large sample/feature imbalance no longer exists, and any models created will be more robust.

We trained the model on the genes which had a mutual information value of above 0.05. We selected this threshold to ensure we maximize the number of significant genes used in the selection model. With this threshold, we kept only 2565 genes for use from the original 20532 genes (Fig. 1).

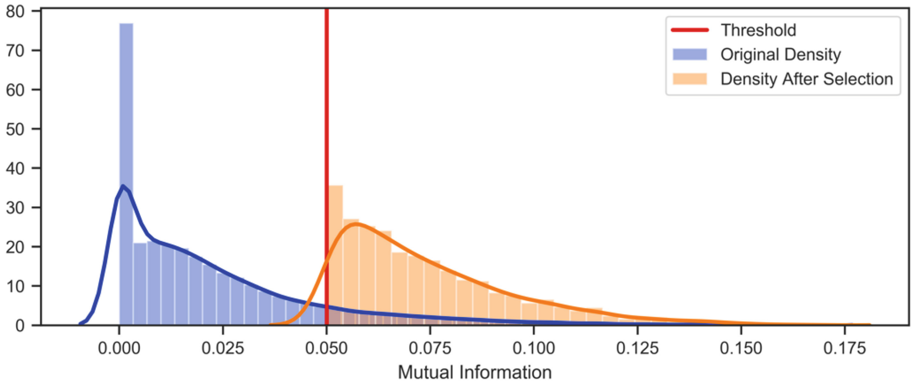

Fig. 1. Density plot of mutual information w/threshold of MI = 0.05

Without such initial filtering, the RFE is infeasible from a computational perspective. During each iteration of training, a logistic regression classifier was trained to classify samples as healthy or cancerous. After training, the model ranked the features in order of importance, and the least important feature was removed. This process was repeated until there was only one feature left, and a ranked list of all of the genes was obtained. The results of the RFE are summarized in Table 4.

Table 4. Recursive feature elimination rankings

Gene ID	Mutual information	RFE rank (Logistic)
Gene_HPN-3249	0.159367224	1
Gene_GSTM1-2944	0.134699717	2
Gene_APOE-348	0.059651147	3
Gene_MRPL41-64975	0.091810873	4
Gene_TRIB1-10221	0.060442517	5
Gene_RPL18A-6142	0.051059319	6
Gene_ISG15-9636	0.052118627	7
Gene_RHOB-388	0.056353286	8
Gene_AMACR-23600	0.107861574	9
Gene_MYLK-4638	0.134553865	10
Gene_FLNA-2316	0.058148017	11
Gene_EEF1G-1937	0.059776923	12
Gene_RPL37-6167	0.074615666	13
Gene_WFDC2-10406	0.099545078	14
Gene_APOC1-341	0.078565194	15
Gene_RPLP0-6175	0.054809254	16
Gene_PCP4-5121	0.082975965	17
Gene_GDF15-9518	0.058465024	18
Gene_RPL28-6158	0.103847129	19
Gene_ATP5MF-9551	0.055716636	20

Many of the genes found in Table 4 are not found in Table 3. This indicates that many of the genes with high mutual information were related to each other and as a result were not found to be as significant by the RFE.

Compared to the model that will be generated in Sect. 5 using PCA, any models using this set of features will have higher interpretability. These features are genes that can be easily related to pathology. In addition, it is economically advantageous to sequence only a few genes as compared with sequencing the whole genome [21]. Besides the topics discussed in Sect. 6, in the future we would like to apply RFE to the combined data, and use the resulting features for classification.

5 Predictive Modeling

In this section, we developed a predictive model using the combined data. The features were condensed using PCA (Principal Component Analysis), allowing us to develop a classification model. In our current research, a logistic regression model was developed to fit the preprocessed datasets from Sect. 3. In Sect. 6, we will discuss how to extend this model to include the prediction of cancer stages using more advanced modeling approaches.

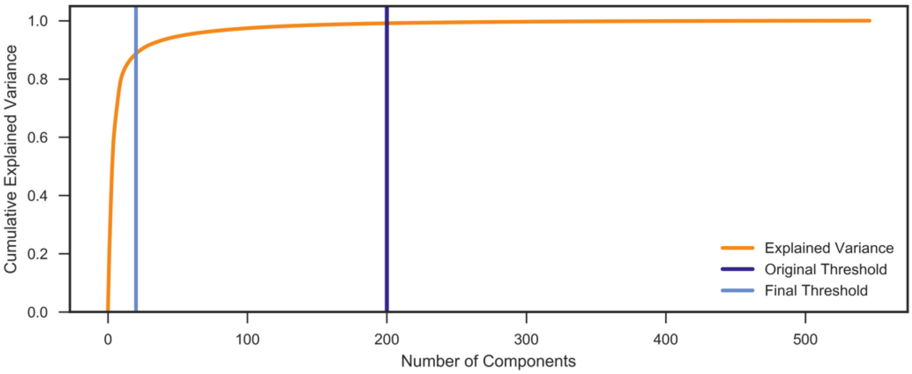

Fig. 2. PCA cumulative explained variance

5.1 Dimensionality Reduction with Principal Component Analysis

High dimensionality prevents models from producing effective results because it is difficult for the model to extract target information from the data. Dimensionality must be reduced in order for models to function. Besides RFE, the other dimensionality reduction technique we used was Principal Component Analysis (PCA). This technique reduces a set of possibly correlated variables into a set of principal components, which are uncorrelated variables. These principal components represent the variance in the data, with each successive principal component representing less variance.

In order to pick how many components to reduce the data to, we found the cumulative explained variance for each number of components (Fig. 2). We originally selected 200 components for the combined data and 50 for each individual dataset. We found however that the logistic regression models trained on this data were heavily overfitting, achieving nearly 100% accuracies. A full summary of the performance of the combined data 200 principal component model can be found in Table 5. We hypothesized that the overfitting was caused by the feature sample imbalance, and that there were still too many features being used. To combat this problem, we decided to reduce the number of principal components for each data type to 5 and the number on the combined data to 20. We chose this number of components because they explained nearly 90% of the variance, a heuristic often utilized to determine how many principal components to use. Each of these five resulting datasets was later used for classification.

5.2 Fitting of the Models

We created six different logistic regression models, one for each of the post-PCA datasets, and one for the top five features from the recursive feature elimination. We chose logistic regression due to the linear nature of the data and the robustness of the logistic regression algorithm. 5-fold cross-validation was performed and repeated twenty times in order to validate the robustness of the classifier. The upsampled data was used, allowing us to do the cross-validation, which would otherwise not have been possible.

Various metrics were calculated for each model trained during the cross-validation, and the final metric represents the average of all of these results. The metrics collected were accuracy, precision, recall, f1 score, ROC AUC, PR AUC, false-negative rate, and false-positive rate. Precision quantifies how many instances that were predicted positive are actually positive. Recall quantifies how many instances that are actually positive were predicted as such. F1 score is the harmonic average of precision and recall and it quantifies the overall performance of the model. ROC (Receiver Operating Characteristics) AUC is the area under the curve with false positive rate on the x-axis and true positive rate on the y-axis. PR (Precision-Recall) AUC is the area under the curve with recall on the x-axis and precision on the y-axis. Accuracy was chosen just to see a benchmark, but due to the imbalanced nature of the dataset, the precision, recall, and f1 score provide a much more meaningful quantification of the model's performance. ROC AUC was calculated since it is a standard metric. However, PR AUC gives a more meaningful representation of the model performance. Figure 3 shows the PR and ROC curves for the combined PCA model.

Table 5. Model metrics

	Dataset	Accuracy	Precision	Recall	F1 score	ROC AUC	PR AUC	FN	FP
PCA	Gene	0.850	0.783	0.736	0.759	0.913	0.841	0.084	0.065
	Exon	0.896	0.859	0.806	0.832	0.952	0.902	0.062	0.042
	Exon Junction	0.919	0.885	0.859	0.872	0.970	0.946	0.045	0.036
	Isoform	0.879	0.832	0.780	0.805	0.933	0.887	0.070	0.050
	Combined	0.972	0.969	0.942	0.955	0.995	0.987	0.019	0.010
	Combined200	0.999	0.995	1.000	0.998	0.999	0.999	0.000	0.001
RFE	Gene	0.961	0.926	0.941	0.912	0.993	0.985	0.014	0.024

Table 5 gives a summary of each model's performance. We created seven models and report eight metrics on each of them. For PCA models, we applied the PCA algorithm to five different prostate cancer sequencing datasets (Gene, Exon, Exon Junction, Isoform, and all the former combined) and trained a logistic regression model on the resulting principal components. These models are referred to as PCA gene, PCA exon, PCA exon junction, PCA isoform, and PCA combined respectively. The PCA Combined200 model uses the top 200 principal components instead of the top 20, as discussed in Sect. 5.1. The RFE gene model uses features selected by the recursive feature elimination algorithm applied to the Gene dataset.

Out of the four individual PCA models, the gene model performed the worst. However, the RFE gene model outperformed all the individual PCA models, implying that RFE is a better way to do feature reduction, as there may be less information loss. This gives another reason for RFE to be applied to the combined data in future work.

The combined PCA model performed the best on all metrics except false negative rate, with a ROC AUC score of 0.99, PR AUC of 0.98, and an F1 score of 0.96. The RFE gene model achieved a false negative rate lower than the combined PCA model. However, it performed more poorly with regards to PR AUC and F1 score. Thus, overall the RFE gene model is a less effective classifier than the combined PCA Model.

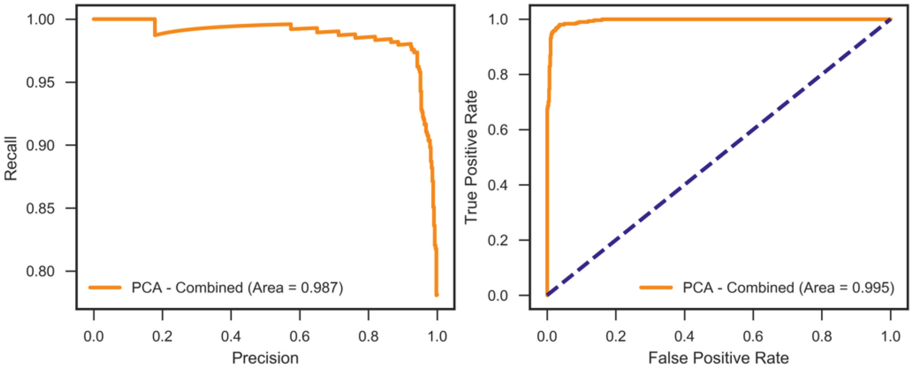

Fig. 3. PR (left) and ROC (right) curves for combined PCA model

6 Discussion

Our current approach, logistic regression, is limited due to its inability to determine the stage of cancer because it is a binary model. In order to predict the patient's cancer stage, we attempted using a multi-classification approach. We utilized an Artificial Neural Network (ANN), a deep learning approach, with a Rectified Linear Unit (ReLU) as our activation function in the middle layers and SoftMax in the final layer. We used the post-PCA combined dataset to train the model.

To analyze the performance of our ANN, we compared the performance (cross-entropy) of the model to that of a zero model. The zero model is a model that represents randomly guessing which stage of cancer (if any) the patient had. We found that our model made no significant improvement in performance as compared with the zero model. The zero model had an accuracy of 0.2501 and our model had an accuracy of 0.2839. We hypothesize that this may be because the stage and size of the cancer are not reflected by the divergence in RNA sequences from healthy cells. Our next step is to incorporate the clinical data to create a more effective classifier.

During our analysis we found that any model generated on our dataset was highly dependent on input data and thus unstable in terms of its performance with regards to the test-train split made. This is because any stable model must have a stable set of statistical properties in the training set and a reliable statistical relationship between the features and target. However, for such a small sample size, as is the case here, any split would significantly disturb the distribution of feature values of the population in training and testing data sets, especially for testing set. This is because slight changes in the number of healthy tissues in the testing set would drastically change the generated model performance. Thus, our challenge becomes whether we can build a stable model with this data set.

A potential area for improvement is in the area of feature reduction. Our current approach with PCA may be limited by its linearity, and this may be improved upon using an autoencoder technique to create and highlight new features of importance. One of our on-going efforts is to use multiple layers of auto-encoding neural networks to do the compression. However, we must be cautious of problems associated with model complexity in this case.

A limitation which was brought up concerned the performance of mutual information with RFE. It was suggested that using a random forest in place of RFE may be more optimal as it already has a built-in information theoretic feature selection criterion and may improve overall performance. The reasoning is that since no features are completely correlated, there is still loss of information when the RFE is used. In other words, if one feature is removed, there may be an inadvertent and undesirable loss of information. After creating a random forest of 300 trees with depths of 2 we found that the most important selected features were comparable to those as selected by mutual information. For example, we found that PLOR2H-5437, which ranked first in mutual information, was ranked seventh in the random forest.

7 Conclusion and Future Work

Our model significantly improves upon previous prostate cancer research done using microarray data, and research done using RNA sequencing breast cancer data. In addition, it improves upon the accuracy of the standard biopsy procedure and has the potential to standardize the process. The current method of identifying cancer through the use of a biopsy is inaccurate and based on the skill of the technician, which varies from location to location [22]. Our model provides a way to improve upon this, because it can match or exceed the accuracy of diagnosis of a highly experienced technician, and removes the disparity that exists between highly experienced and less experienced technicians. In addition, there is an estimated 30% false negative rate associated with the current method of diagnosis [5]. Our method greatly reduces this number to 1.3%, meaning that many fewer cases of prostate cancer are missed, especially the less severe cases. Although our proposed system still requires the use of a biopsy, it has the potential to be a very useful tool which can help doctors accurately diagnose patients and minimize false positives and false negatives.

In this paper, we have presented an approach which has the potential to improve upon the microarray method of prostate cancer identification greatly. Our accuracy also significantly improved upon previous attempts to classify prostate cancer using machine learning. Our method, however, only demonstrates the initial advantages of using machine learning in cancer prediction. This work is limited by the dataset used, which forces us to create false observations to fix the class imbalance. Future work needs to address this class imbalance and attempt other ways of dealing with it which may have less of a potential to artificially inflate the results. Currently, our model does not utilize clinical data when making a classification. In future work, we will also fit the clinical data to see what effect on model effectiveness that data may have. We are testing deep learning approaches to compare this approach to traditional machine learning methods. We would also like to apply the RFE algorithm to the combined data to see how it performs. We hypothesize that it would outperform the results achieved with the PCA.

References

1. American Cancer Society Cancer Facts & Figures 2019. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf. Accessed 27 Jan 2019
2. American Joint Committee on Cancer: AJCC Cancer Staging Manual, 8th edn. Springer, New York (2017). https://doi.org/10.1007/978-1-4757-3656-4
3. Draisma, G., et al.: Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context. J. Natl. Cancer Inst. **101**(6), 374–383 (2009)
4. Albertsen, P.C.: The unintended burden of increased prostate cancer detection associated with prostate cancer screening and diagnosis. Urology **75**(2), 399–405 (2010)
5. Serefoglu, E.C., Altinova, S., Ugras, N.S., Akıncıoğlu, E., Asil, E., Balbay, M.D.: How reliable is 12-core prostate biopsy procedure in the detection of prostate cancer? Can. Urol. Assoc. J. **7**(5–6), E293–E298 (2013)
6. Kukurba, K.R., Montgomery, S.B.: RNA sequencing and analysis. Cold Spring Harb. Protoc. **2015**(11), 951–969 (2015)
7. Deep Learning for genomic data analysis. https://repositorio-aberto.up.pt/bitstream/10216/106492/2/205645.pdf. Accessed 27 Jan 2019
8. Mitra, S., Saha, S., Acharya, S.: Fusion of stability and multi-objective optimization for solving cancer tissue classification problem. Expert Syst. Appl. **113**, 377–396 (2018)
9. Penney, K.L., et al.: mRNA expression signature of Gleason grade predicts lethal prostate cancer. J. Clin. Oncol. **29**(17), 2391–2396 (2011)
10. Cuzick, J., et al.: Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. Lancet Oncol. **12**(3), 245–255 (2011)
11. Erho, N., et al.: Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. PLoS ONE **8**(6), e66855 (2013)
12. Mo, F., et al.: Stromal gene expression is predictive for metastatic primary prostate cancer. Eur. Urol. **73**(4), 524–532 (2018)
13. Tyekucheva, S., et al.: Stromal and epithelial transcriptional map of initiation progression and metastatic potential of human prostate cancer. Nat. Commun. **8**(1), 420 (2017)
14. Sharifi-Noghabi, H., et al.: Deep Genomic Signature for early metastasis prediction in prostate cancer. bioRxiv, 276055 (2018)
15. Takeuchi, T., Hattori-Kato, M., Okuno, Y., Iwai, S., Mikami, K.: Prediction of prostate cancer by deep learning with multilayer artificial neural network. bioRxiv, 291609 (2018)
16. Coudray, N., et al.: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat. Med. **24**(10), 1559 (2018)
17. Danaee, P., Ghaeini, R., Hendrix, D.A.: A deep learning approach for cancer detection and relevant gene identification. In: Pacific Symposium on Biocomputing, vol. 22, pp. 219–229 (2016)
18. Golcuk, G., Tuncel, M.A., Canakoglu, A.: Exploiting ladder networks for gene expression classification. In: Rojas, I., Ortuño, F. (eds.) IWBBIO 2018. LNCS, vol. 10813, pp. 270–278. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-78723-7_23

19. Wei, L., Jin, Z., Yang, S., Xu, Y., Zhu, Y., Ji, Y.: TCGA-Assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. Bioinformatics **34**(9), 1615–1617 (2017)
20. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
21. Advances in the molecular diagnosis of cancer. https://repositorio.unican.es/xmlui/bitstream/handle/10902/14278/Cadrecha Sanchez Natalia.pdf?sequence=1&isAllowed=y. Accessed 27 Jan 2019
22. Cancer Classification using Gene Expression Data with Deep Learning. https://www.politesi.polimi.it/bitstream/10589/138427/7/thesis.pdf. Accessed 27 Jan 2019