 doi:10.20944/preprints201911.0298.v1

Article

A Hierarchical Machine Learning Model to Discover Gleason Grade Group-specific Biomarkers in Prostate Cancer

Osama Hamzeh [1,‡], Abedalrhman Alkhateeb [1,*,‡], Julia Zheng [1], Srinath Kandalam [1], Crystal Leung [2],and Luis Rueda [1,*]

[1] School of Computer Science, University of Windsor 401 Sunset Ave, Windsor, ON, Canada, N9B 3P4; hamzeho, alkhate, fzheng12z, kandala1, lrueda@uwindsor.ca

[2] Schulich School of Medicine and Dentistry, Western University 1151 Richmond St, London, ON, Canada N6A 5C1; cleung2021@meds.uwo.ca

* Correspondence: lrueda@uwindsor.ca; Tel.: +1-519-253-0000 ext. 3002

‡ These authors contributed equally to this work.

Abstract: 1) Background: One of the most common cancer that affects men worldwide is prostate cancer. This disease motivates parts of the cells in the prostate to lose control of their growth and division, causing the tumor to grow in an uncontrolled way. Gleason scoring is a schematic pathological grading system that is used to examine the potential aggressiveness of the disease in the prostate tissue. In this regard, the advancement in computing and next-generation sequencing technology has allowed us to study pateints' genomic and transcriptomic profiles with different Gleason scores, providing a much more focused insight than higher-resolution microscopy. 2) Methods: We are proposing a machine learning method used to analyze gene expressions of prostate tumors with different Gleason scores, and to identify potential genetic biomarkers for each group. A publicly-available RNA-Seq dataset of a cohort of 104 prostate cancer patients have been retrieved from the National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO) repository. We categorize patients by their Gleason scores into different groups to create a hierarchy of disease progression. A hierarchical model with standard classifiers in different Gleason groups (hereinafter called *nodes*) to identify and predict nodes based on their mRNA or gene expressions. At each node, patient samples are analyzed via class imbalance and hybrid feature selection techniques to build the prediction model. The outcome of each node is a set of genes that can separate the Gleason group from the remaining groups. To validate the proposed method, the set of identified genes are used to classify a second dataset of 499 prostate cancer patients that have been collected from cBioportal [1]. 3) Results: The overall accuracy of applying the method on the first dataset is 93.3%, while it is 87% on the second (validation) dataset. Two genes have been found to be potential biomarkers of specific Gleason groups; PIAS3 has been identifed for Gleason score 4+3=7, while UBE2 could be a potential biomarker for Gleason score 6. Other proposed genes that were not found in the literature might be potential biomarkers. 4) Insigh: The latest literature supports that the genes predicted by the proposed method are strongly correlated with prostate cancer progression and tumour development processes. Furthermore, pathway analysis shows that both PIAS3 and UBE2 share the same protein interaction pathway, the JAK/STAT signaling process.

Keywords: Supervised learning; next generation sequencing; classification; transcriptomics; Gleason score detection; prostate cancer.

 doi:10.20944/preprints201911.0298.v1

Gleason group	Score
1	**6**
2	**3+4=7**
3	**4+3=7**
4	**8**
5	**9 and 10**

Table 1. Gleason groups considered in this study.

1. Introduction

Cancer is among the main causes of death worldwide. Approximately ten million people around the world died because of it in 2018. In the same year, approximately eighteen million new cases were diagnosed. Among males, prostate cancer is the most incident cancer type; 1.276 million new cases were diagnosed in 2019 [2]. To date, most cancer studies have concentrated on finding biomarkers that enable differentiating malignant tumors from benign ones. More recent studies, though, have focused on specific clinical aspects of tumors, such as recurrence, progression, survivability and metastasis, among others.

In the 1950s, Pierre Denoix suggested a system that categorizes solid tumors into different stages [3]. The classification (TNM) of cancer progression is done by utilizing (T) the extension and the size of the main tumor, (N) the lymphatic involvement, and (M) the metastasis levels [4]. By utilizing samples from prostate cancer tumors, a pathologist uses microscopic patterns of the cells to assign a score to the tumor, aka *Gleason score*. That score is calculated by adding two numbers: the most common pattern of the tumor cells is used as the first number, while the second number corresponds to the next most common pattern. Each individual score varies from 3 to 5, depending on the aggressiveness of the tumor, where the highest score means the most aggressive form of cancer [5]. Epstein et al., however, indicated that Scores 2-5 are no longer assigned to the tissue and these multiple scores can be categorized together as group 6, yielding five categories as depicted in Table 1. As such, we have used it as the main scheme for prostate cancer score categorization in our method to detect transcriptomic biomarkers that can accurately classify specific Gleason scores. This categorization strategy has proven to be a clear indication of cancer recurrence, and improved the prognostic role of the Gleason score, especially, for lower scores [6].

The majority of recent research works focus on studying gene expressions for prostate cancer progression. However, differential transcription of the same gene may yield to a different degradation rate for the resulting protein. Gleason scores, which measure the prostate cancer aggressiveness, have been found to be correlated with Protein rates of degradation [7]. The focus of this study though, is on the transcription level to find genes that can identify a specific Gleason group from the others.

On the other hand, advances in next-generation sequencing (NGS) technology has made genomic data analysis widely available. The output of NGS sequencers requires pre-processing algorithms such as aligning the reads to a reference human genome and assembling them into transcripts. Many genomic tools that align the RNA-Seq reads to the Human genome have been proposed, especially BLAST is one of the first tools developed to align reads [8]. TopHat2 is a widely-used, open-source tool that incorporates Bowtie sequence alignment to align reads [9]. STAR is the fastest RNA-Seq sequence alignment algorithm to date, although, it requires huge computational resources to perform efficiently [36].

In addition, machine learning applications in genomic analysis has become a solid approach to analyze RNA-Seq data for studying a multitude of diseases. Alkhateeb et al. proposed a supervised method to discover biomarkers that can predict the likelihood that a prostate cancer tumor will progress to the next stage [11]. Arvaniti et al. proposed a deep learning approach to predict Gleason scores [12]. Their model was trained using tissue microarray (TMA) images of 641 patients with varying Gleason scores, and validated using 245 patient samples with Gleason scores that were reviewed by

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0298.v1

pathologists. Although the study by Arvaniti et al. reported a decent performance measurements (average accuracy 85.72%, and recall 0.57%), it did not report the panel of biomarker genes that were used by the trained convolutional neural network (CNN) to predict Gleason scores. Citak-Er et al. proposed a machine learning approach for predicting Gleason scores [10]. Their method uses a support vector machine (SVM) on prostate images to learn the visual attributes of the disease and to predict the disease outcome. That study was conducted on a limited cohort of prostate cancer patients, and the results showed a higher sensitivity over the specificity in the prediction model (Accuracy = 76.83%, Sensitivity = 83.38%, Specificity = 68.36%).

In this work, we are extending our previously proposed prediction model, which is based on analyzing the RNA-Seq data from patients with different Gleason scores [14]. The method can track transcripts associated with specific genes, in addition to their corresponding expression values. The results of the initial trial show a great potential to build a simple system to diagnose Gleason scores based on NGS data.

2. Results

The first dataset is a collection of 104 samples and their TPM values. Stated as a classification problem, this study designates five classes obtained from joint Gleason groups, the distribution of each group is shown in Figure 1. The dataset was mapped against the human genome version HG19 with 88% to 99% uniquely aligned reads. Throughout a 10-fold cross-validation model, we obtained a total of seven samples that were misclassified and another 97 samples that were classified correctly, with the total number of samples being 104. The accuracy of the model is calculated from the total number of correctly classified samples divided by the total number of samples, resulting in an accuracy of 93.3%.

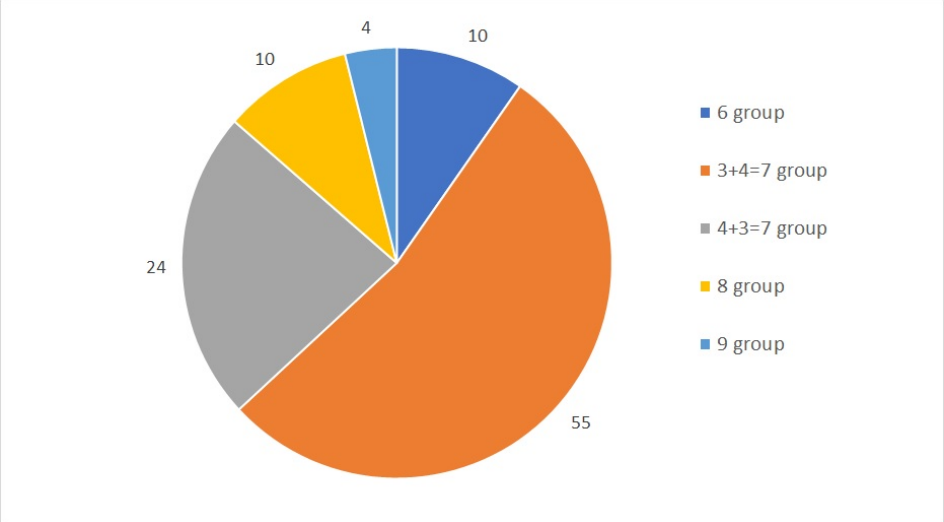

Figure 1. Gleason groups and their distributions.

Six transcripts have been identified, which are differentially expressed in five different Gleason groups. Of these, the corresponding genes shown in Tables 3, 4, 5 and 6 have shown to be the most relevant that can identify prostate cancer with the Gleason groups by the proposed hierarchical method shown in Figure 2. In the hierarchy, different classification methods for each stage are shown in Table 2. The first node of the hierarchy yields 94% accuracy in identifying Gleason group 3+4=7 versus the other groups. The samples are then passed through node 2, in which Gleason group 4+3=7 is identified from the rest with a prediction accuracy of 98%. The other samples are then passed through node 3, where Gleason group 6 is identified with 100% accuracy. The remaining samples are finally processed in the last node, where the Gleason group 8 is identified from the Gleason group 9 with 100% accuracy.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0298.v1

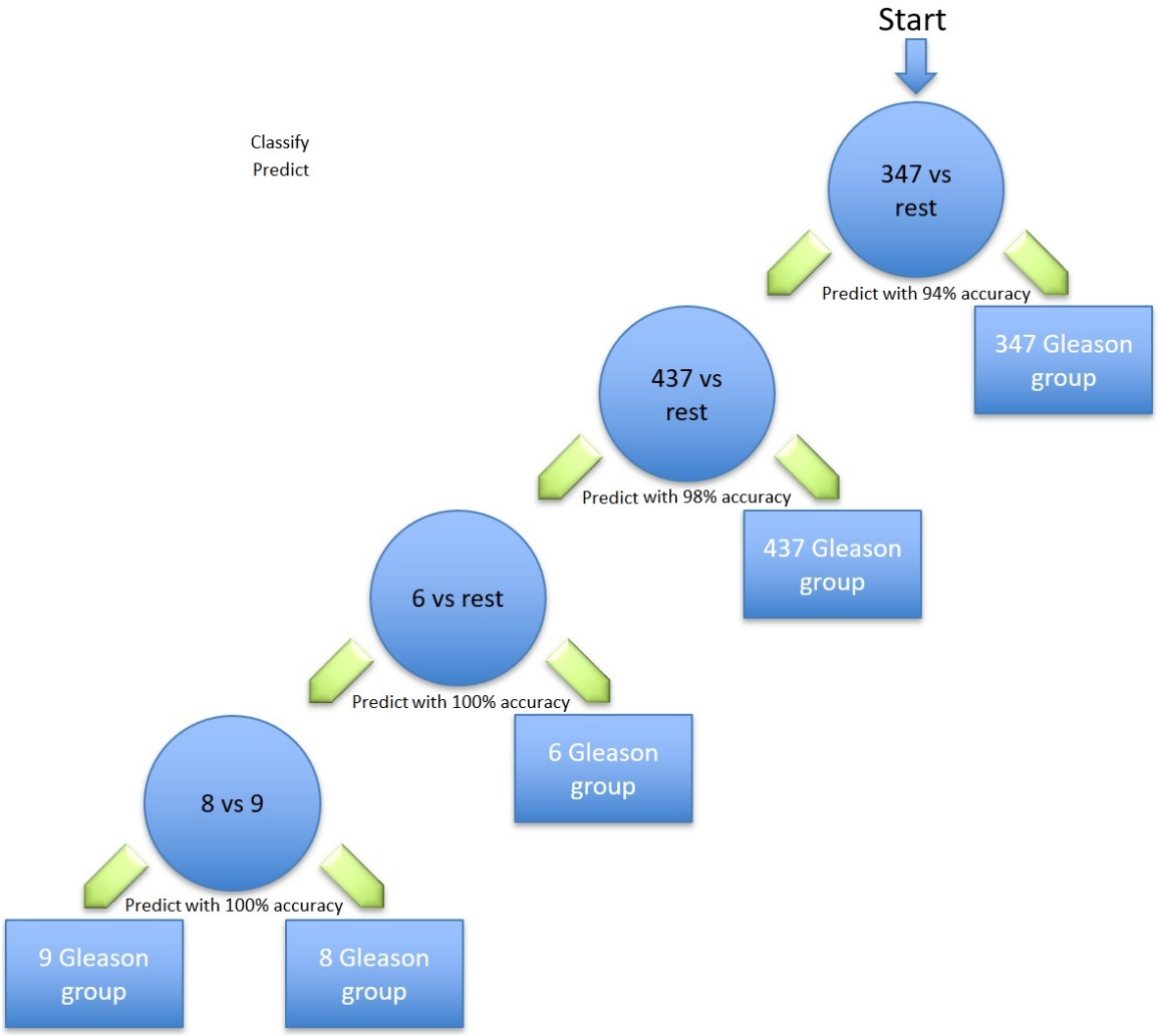

Figure 2. Hierarchical tree of classifications of Gleason groups against the rest, along with the corresponding classification accuracies.

Due to the similarity in the aggressiveness of the tumor and the low number of samples, all the other Gleason scores were merged in the last node.

Table 2. Classification performance for each step in the hierarchy.

Gleason group	Accuracy	Sensitivity	Specificity	F-Measure	MCC	ROC Area
3+4=7 vs Res	94	95	94	0.94	0.88	95
4+3=7 vs Rest	98	100	96	0.98	0.96	99
6 vs Rest	100	100	100	1.00	1.00	100
8 vs 9	100	100	100	1.00	1.00	100

Figure 3 shows the classifiers that have been utilized to identify the set of transcripts that differentiate specific Gleason groups against the rest. The classifiers are represented in the x-axis, while the classification performance are represented in the y-axis.

Naïve Bayes outperformed the other classifiers as it distinguished the first Gleason group from the rest by 94%, the second group by a higher accuracy of 98% and the last two Gleason groups by 100% as shown in Figure 3.

In order to conduct further tests of the model, the method was applied on an another publicly-available dataset [15] obtained from the National Center for Biotechnology Information (NCBI) portal [16], which contains gene expressions for 498 patient samples. The proposed model

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0298.v1

Table 3. Set of resulting transcripts in Gleason group 3+4=7.

Transcript	Gene	Description
NM_001170880	GPR137	G protein-coupled receptor 137 (GPR137), transcript variant 2
NM_001198827	C8orf58	chromosome 8 open reading frame 58 (C8orf58), transcript variant 3
NM_004629	9p13.3	Fanconi anemia complementation group G (FANCG)
NM_001098268	LIG4S	DNA ligase 4 (LIG4), transcript variant 3
NM_016641	GDE1	glycerophosphodiester phosphodiesterase 1 (GDE1), transcript variant 1
NM_002445	MSR1	macrophage scavenger receptor 1 (MSR1), transcript variant SR-AII
NM_001126337	TUFT1	tuftelin 1 (TUFT1), transcript variant 2
NM_033071	SYNE1	spectrin repeat containing nuclear envelope protein 1(SYNE1), transcript variant 2
NM_052906	ELFN2	extracellular leucine rich repeat and fibronectin typeIII domain containing 2 (ELFN2), transcript variant 1
NM_000714	TSPO	translocator protein (TSPO), transcript variant PBR
NM_004374	COX6C	cytochrome c oxidase subunit 6C (COX6C)
NM_001007544	C1orf186	chromosome 1 open reading frame 186 (C1orf186)
NM_001276438	KCNJ15	potassium voltage-gated channel subfamily J member 15 (KCNJ15), transcript variant 7
NM_001252021	TOR2A	torsin family 2 member A (TOR2A), transcript variant 7
NM_152612	CCDC116	coiled-coil domain containing 116 (CCDC116), transcript variant 1

Table 4. Set of resulting transcripts in Gleason group 4+3=7.

Transcript	Gene	Description
NM_001136224	RCOR3	REST corepressor 3 (RCOR3), transcript variant 2
NM_001017967	MARVELD3	MARVEL domain containing 3 (MARVELD3), transcript variant 1
NM_006099	PIAS3	protein inhibitor of activated STAT 3 (PIAS3)
NM_152395	NUDT16	nudix hydrolase 16 (NUDT16), transcript variant 2
NM_006473	TAF6L	TATA-box binding protein associated factor 6 like (TAF6L)
NM_001145541	TCP11L1	t-complex 11 like 1 (TCP11L1), transcript variant 2
NM_182501	MTERF4	mitochondrial transcription termination factor 4 (MTERF4)

showed excellent prediction accuracy on the 498 patients gene expressions. The prediction accuracies for all the Gleason groups were above 90% except for the 437 Gleason group vs rest. 4.

3. Discussion

Most of the genes which the identified transcripts belong to, have been found in the literature to play a significant role in prostate cancer and cancer in general. PIAS3, which contains transcript NM_006099 is known to enhance the transcriptional activity of androgen receptors in prostate

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0298.v1

Table 5. Set of resulting transcripts in Gleason group 6.

Transcript	Gene	Description
NM_003350	UBE2V2	ubiquitin conjugating enzyme E2 V2 (UBE2V2)
NM_153051	MTMR3	myotubularin related protein 3 (MTMR3), transcript variant 2
NM_207445	C15orf54	chromosome 15 open reading frame 54 (C15orf54),

Table 6. Set of resulting transcripts in Gleason group 8.

Transcript	Gene	Description
NM_001258330	EPB41L1	erythrocyte membrane protein band 4.1 like 1 (EPB41L1), transcript variant 4

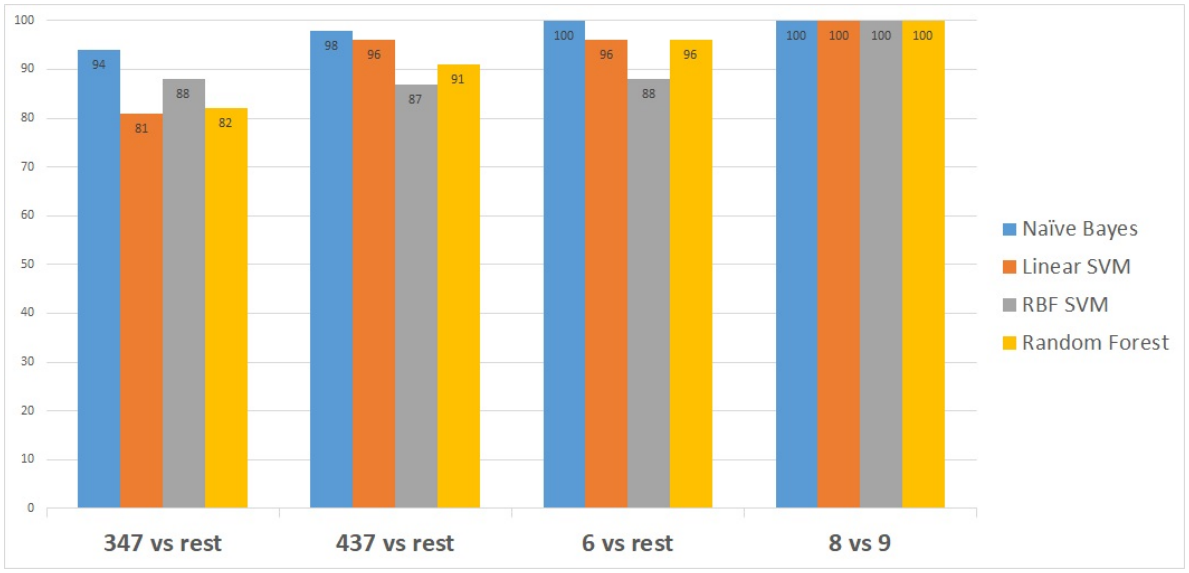

Figure 3. Accuracy obtained by each classifier for classifying one versus the rest for all five Gleason groups.

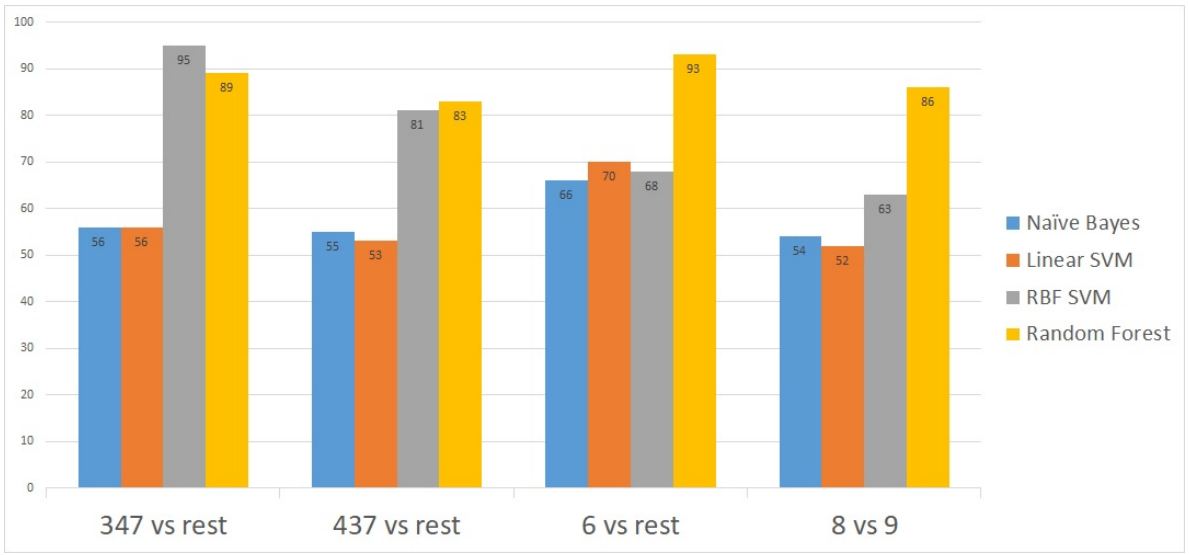

Figure 4. Classification accuracies obtained after applying the model on the second dataset.

 doi:10.20944/preprints201911.0298.v1

cancer cells. This means that the gene plays a significant role in the growth of the disease [17]. Reduced expression of PIAS3 has been directly spotted in a cellular mechanism that promotes epithelial-to-mesenchymal transition and cell motility, which are key factors in prostate cancer aggressiveness of the disease, especially at later stages [18]. This reduction has also been found to be involved in other types of cancer, including lymphoma [19] and lung cancer [20] cells.

Another relevant gene to study further is UBE2V2, which is a DNA repair gene. That gene's transcript, which has been selected in the third node of the hierarchical model, has also been previously identified to play a significant role prostate cancer. It has been reported that an UBE2V2 variant is associated with familial prostate cancer [21]. In another study, differential expression of UBE2V2 has also been associated with poor prognosis in breast cancer [22]. In this study, different quantification of UBE2V2 transcript has been able to predict Gleason score 3 + 3 = 6 applying our method on the first data set. UBE2V2's protein product regulates the activity of protein Ube2N. Ube2N complexed with Ube2V2 can then catalyze polyubiquitin chain synthesis. These polyubiquitin chains play a very important role in *in vivo* DNA repair. Also, differential expression of UBE2V2 may lead to a loss of appropriate DNA repair processes, leading to cancer cell development as reported in [23].

GPR137 was als found to be differentially expressed in Gleason score 3+4=7 tumors. Ren et al. previously showed that GPR137 is upregulated in prostate cancer tissue, and knockdown studies resulted in decreased cell colony formation. It is suggested that GPR137 plays an important role in unregulated cell proliferation, which is a key cancer development process [24].

Differential expression of RCOR3 has been associated with tumors of Gleason Score 4+3=7. This biomarker is discussed here due to the relevance of its protein product to both cancer development and undifferentiated cells. Rest Corepressor 3 (Rcor3) acts in opposition to Rcor1/2 to inhibit the H3K4 demethylation activity of LSD1. This loss of demethylation results in disordered gene expression. The result of this process is that cell differentiation is antagonized [25]. Tumors of Gleason score 4+3=7 contain tissues with poorly structured glands, which is consistent with a loss of cellular differentiation [26].

In the first dataset, differential expression of EPB41L1 (t-test with p-value < 0.05) has been found to be associated with tumors of Gleason group 8. High level of expression in EPB41L1 has been observed to decrease in prostate cancer compared to normal cells. Earlier studies have found proteins encoded by EPB41L1 are associated with the organization of the cell cytoskeleton. Disruption of normal protein expression may play an important role in disorganized cell and tissue structures associated with higher grade prostate cancer [27]. Furthermore, reduced expression of EPB41L1 plays an important role in earlier biochemical recurrence. Its significant gene expresion reduction has also been associated with highly metastatic lung cancer and metastatic forms of breast cancer [28]. In addition, EPB41L1 has been found to be differentially expressed in gastric cancer [29]. EPB41L1 plays an important role in negative regulation of cell metastasis, migration and invasion, which makes it a good candidate to play a significant role in prostate cancer progression and prognosis.

In this study, we found that the proposed machine learning model has helped identify key genes which are differentially expressed between tumour tissues of different Gleason score. These genes are identified as potential biomarkers of cancer grade categories for prognostic use. Specific genes from each category have been shown to play important roles in cancer development and cell cycle regulation such as cyclin D1/D2 and c-Myc, causing the promotion of cell proliferation and inhibition of apoptosis [30]. However, full validation of potential biomarkers would require further clinical studies to determine the predictive value of each gene. It must be noted this forms a barrier to wider implementation of the proposed machine learning model into clinical practice. Certainly, the machine learning model has been able to identify starting points for study of disease prognosis and potential treatment.

Additional literature studies revealed two of these genes, PIAS3 and UBE2, which were identified to be differentially expressed in groups 4+3=7 and 3+3=6, respectively. These genes share a common protein interaction through the Janus Kinase (JAK)/ Signal Transducers and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0298.v1

Activators of Transcription (STAT) signalling pathway. A particular study showed blocking Jak-STAT signaling pathway inhibits tumor initiation in prostate cancer providing a novel therapeutic target [31]. The JAK/STAT signaling pathway has been a target of interest in many cancer studies in recent years. In prostate cancer, it has been shown that the expression levels of JAK/STAT have an impact on the progression of the disease [32,33]. With regards to PIAS3 and UBE2, two identified possible biomarkers, both participate in the same pathway which has been previously linked to disease progression. These genes are of particular interest for further studies. Figure 5 depicts the protein-protein interaction among genes in 4 + 3 = 7 and 6 groups as extracted from ProteomicsDB (*https : //www.proteomicsdb.org/proteomicsdb/#human/proteinDetails/*86810*/interactions*) based on experimental and literature evidence. The figure shows that both PIAS3 and UBE2 share the same protein interaction network.

Application of the machine learning model to other related datasets to identify biomarkers for important prognostic factors, such as response to pharmaceutical treatment, is also an interesting avenue to follow. This model may act as an additional tool for identification of clinically important genetic differences in tumours, which can affect treatment and disease outcome.

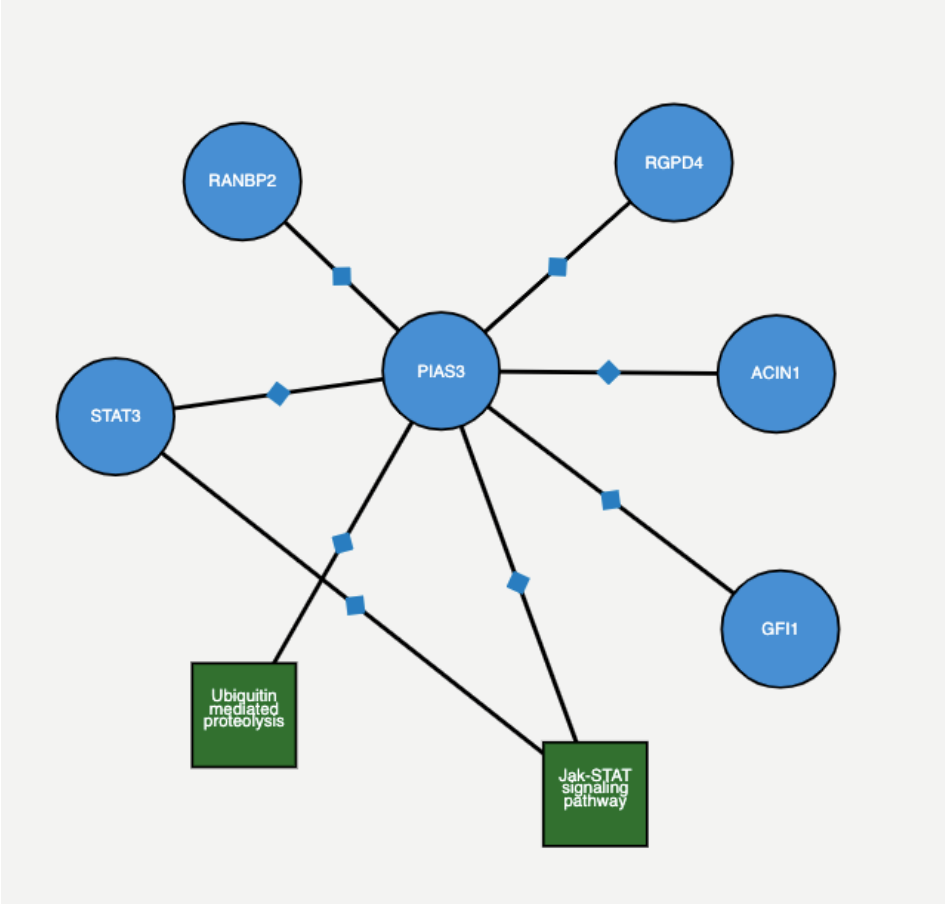

Figure 5. An interactive figure taken from proteomics database STRING [ref], where it shows neighbouring protein binding and pathway interactions for a given gene using STRING and KEGG pathway analysis. Here, the gene of interest is PIAS3, an identified possible biomarker in the 4 + 3 = 7 group. The figure shows the interaction between other proteins and pathways associated with it.

4. Materials and Methods

The primary dataset used in this study has been retrieved from the National Center for Biotechnology Information (NCBI), and is referenced with Gene Expression Omnibus (GEO) number

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0298.v1

Gleason score.	Number of samples
6	10
3+4=7	55
4+3=7	24
8	10
9	4

Table 7. Number of samples in different Gleason groups.

GSE54460 [34]. This RNAseq prostatectomy dataset was extracted from 106 prostate cancer tissue samples and validated on an independent dataset with 140 patients. Several health sciences centers provided data samples as well. The Moffitt Cancer Center (MCC) contributed ten samples from patients who underwent radical prostatectomy between the years 1987 and 2003. The Sunnybrook Health Sciences Centre at the University of Toronto provided 35 samples from patients treated for prostate cancer between the years 1998 and 2006. The Atlanta Veterans Administration Medical Center (AVAMC) donated 61 tissue samples from patients who underwent radical prostatectomy between the years 1990 and 2000. Table 7 shows the number of samples grouped by their Gleason group. Based on Epstein's model, there are five Gleason groups: 4+3=7, 3+4=7, 6, 8, and above 8 (9 and 10).

This dataset was generated by using the Illumina HiSeq 2000 NGS on paired-end sequences of length 51 bp each. The pre-processing pipeline model starts by obtaining the RNA-Seq samples and pre-processing it using SRAtools [35], as depicted in Figure 6. The process continues by incorporating the STAR aligner [36] to align the samples reads into the Human genome (hg19). Then, the process assembles the transcripts and quantify the reads into the assembled transcripts using RSEM [37]. RSEM uses transcripts per million of reads (TPM) to compute the quantification of each read into a transcript.

NGS technology allows us to read the patient's genome and generate a huge amount of raw data in a snapshot. However, this process comes with artifacts and pre-processing must be done before the downstream analysis. These artifacts include duplication and bias reads [38]. Counting the reads that are assembled to Human genome is the physical indicator of a transcript expression. Since the samples are pair-ended reads, TPM is selected to measure the read quantification rather than reads per kilobase per million of reads (RPKM) [39]. The reason for choosing TPM instead of fragments per kilobase per million (FPKM) [41] is that TPM normalizes the reads to the length of the gene first, which makes it easier to compare the quantified reads among samples.

4.1. Class Imbalance

Some classes have markedly lower number of samples than the others, which causes the classifiers to become biased towards the majority class. To solve this problem, multiple resampling methods were deployed and tested to identify a method that would yield the best solution for a specific data set. Oversampling provides a fast solution for minority classes. This method duplicates samples from the minority class and adds them to yield a similar number of samples for each class. Applying oversampling directly did resolve the class imbalance problem and yielded high classification accuracy. However, after taking a closer look at the samples used in these classifiers, a significant overfitting problem was noticed. After applying multiple oversampling and under-sampling methods, the best option was found to be synthetic minority oversampling technique (SMOTE) [42] for oversampling the minority class, while neighborhood cleaning rule (NCL) [43] was used for undersampling the majority class.

NCL works by removing any sample whose class is different from the class of at least two of its three nearest neighbors. SMOTE, instead, introduces a new way of creating new samples, by utilizing the feature vector connecting each sample and introducing a new synthetic sample along the line that connects the two underlying samples. The exact location of the new sample on the line itself is calculated by measuring the Euclidean distance between the two samples and multiplying that value

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0298.v1

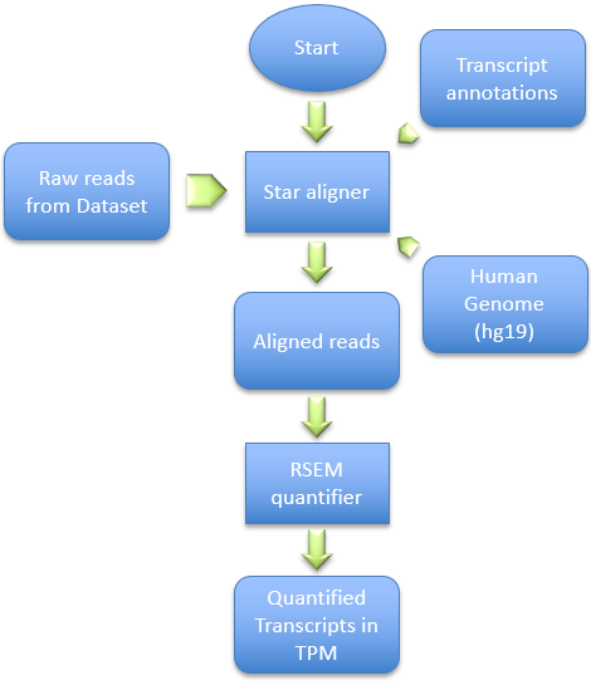

Figure 6. Pre-processing steps used in this study.

by a random number between 0 and 1. Figure 7 shows the mechanism followed by SMOTE by adding new synthetic samples randomly along the line that connects each of the two original samples in the minority class.

4.2. Feature Selection

As output of the pre-processing pipline, the method retrieved 41,971 transcripts with their corresponding quantification measured by TPM. Such a large number of transcripts leads to a complicated classification model, due to the curse of dimensionality [40]. Thus, feature selection was applied to reduce the dimensionality of the problem. The first step of the feature selection step is to filter the transcripts based on their information gain values by selecting the ones with the highest score. The filter method called attribute evaluator is the procedure by which each attribute (transcript) in the dataset is assessed with regards to the class. This procedure produces a list of attributes (transcript) with a score for each attribute showing its effect on the actual class. thenm the attributes with the highest scores are selected, discarding those with lower scores. In this work, information gain (IG) is used as an attribute evaluator to rank each attribute vector [44]. IG of attribute vector X in relation to class vector A is defined as follows:

$$IG(A,X) = H(A) - H(A|X) \tag{1}$$

where

$$H(A) = -\sum_{a\in A} p\,(a)\,log_2(p\,(y)). \tag{2}$$

and

$$H(A|X) = -\sum_{x\in X} p\,(x) \sum_{a\in Y} p(a|x)log_2(p\,(ax)). \tag{3}$$

Here, $H(A)$ is the entropy of the class vector A and $H(A|X)$ is the conditional entropy of A given X.

After filtering the transcripts based on their IG score, a wrapper classification method with minimum redundancy maximum relevance (mRMR) method is used to narrow down the decisive

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0298.v1

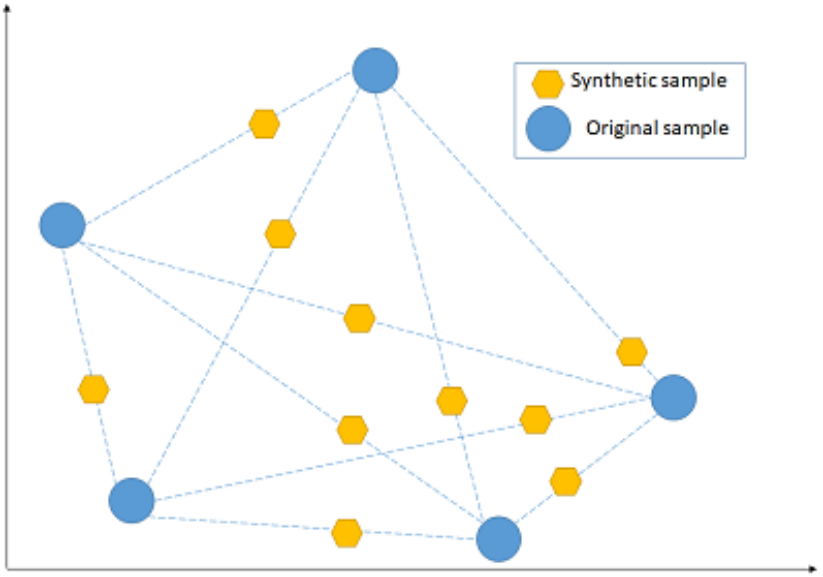

Figure 7. Hypothetical example that shows how synthetic minority oversampling technique (SMOTE) works.

transcripts to a few per group; mRMR has the capability of incorporating any machine learning classifier to select features (transcripts) that minimize the redundancy while increasing the correlation to the class vector at the same time [45]. The wrapper method adds up the features that minimize redundancy (W_i), and maximize the relevance (V_i), with the best possible accuracy of an SVM-linear classifier as per the following equations:

$$W_i = \frac{1}{|S|^2} \sum_{i,j \in S} I(i,j)\,, \tag{4}$$

and

$$V_i = \frac{1}{|S|} \sum_{i \in S} I(h,i)\,. \tag{5}$$

where S is the set of features, $I(i,j)$ is the mutual information between features (i,j), h is the class, in our case, the five Gleason groups.

4.3. Classification

The problem dealt with is multi-class classification, which we solved using the one-versus-rest approach. There are five different classes, which correspond to the five distinct Gleason groups. To apply the one-versus-rest approach, we created five different datasets from the actual data. For each dataset we set one of the classes to form the *positive* class, while the rest of the classes are combined together to form the *negative* class. The classification pipeline resembles a binary tree structure, in which each internal node is a two-class classification problem (see Figure 2). Starting from the root, in the one-versus-rest classification, we remove the samples that belong to the class chosen earlier. We repeat the same steps of building data sets for the remaining four different classes. At each node, the best class is chosen and the classification continues in the same fashion until two classes are left. To select the best class at each node, different performance measures can be used; accuracy, sensitivity and specificity are used here. Note that the hierarchical model involves list processing, and as such, any error at a particular node is propagated down the tree structure. In a greedy-like algorithm, we minimize the error propagation by choosing the class with best performance at each internal node.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0298.v1

Figure 8. Machine learning pipeline used in the proposed method.

4.4. Identifying Transcripts within Different Gleason Scores

We used the Scitkit-learn [46] library to apply different classification algorithms on the final selected transcripts to identify which transcripts can determine a Gleason group from the others based on their quantification values. Standard classifiers such as Naive Bayes and SVM are used in this study to build the classification model. Naive Bayes is a probability-based classifier that applies the well-known Bayes' theorem, and It assumes that the features are independent from each other [47]. Naive Bayes is simple and has been shown to perform very well in many problems, while avoiding overfitting. An SVM classifier was also used to build a prediction model using the selected transcripts in the previous step [48]. The advantages of SVM is their exceptional generalization power, especially in high-dimensional data with a small number of samples. Figure 8 shows the pipeline utilized in this study.

5. Conclusions and Future Directions

Identifying novel biomarkers that are clinically associated with certain Gleason groups in prostate cancer is vital for the diagnosis and the treatment process of the disease. Utilizing NGS technology and machine learning techniques, a supervised learning method is proposed to identify sets of transcripts with significantly different levels of quantification values amongst groups of prostate cancer samples with different Gleason scores. The gene transcripts identified by the propsoed machine learning method were shown in literature to hold crucial roles in cancer pathogenesis, and key transcripts were strongly related to prostate cancer. To validate the model, we also tested it on a gene expression dataset, where the results are very promising.

As a future direction, the same method can be applied to other cancer types. Another possible avenue of this work is to consider analyzing samples from patients who have progressed through more than one Gleason group. This method aims to eliminate confounding factors between patients, potentially leading to a clearer analysis of differential gene expression between different grades of prostate cancer. A multi-omics model based on different types of genomics data for this problem can be investigated too, and may provide a comprehensive analysis of progression, diagnosis, and treatment of the disease.

Author Contributions: L. Rueda is the principal investigator for this project who laid out the main ideas with N. Palanisamy; both share senior authorship. O. Hamzeh, A. Alkhateeb participated equally in implementing the methods, discussed the idea and the model with J. Zheng, C. CLeung and S. Kandalam, who investigated the biological findings. All authors have participated in writing the paper and approved the final manuscript.

Funding: This work was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

NGS	Next-Generation Sequencing
SVM	Support Vector Machine
mRMR	minimum redundancy maximum relevance
IG	Information Gain

References

1. cBioPortal for Cancer Genomics; 2017., 2017 https://cbioportal.org[Online; Last accessed July 201].
2. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin D, Piñeros M, Znaor A, and Bray F. Estimating the global cancer incidence and mortality in 2018: Globocan sources and methods. *International journal of cancer*, 144(8):1941–1953, 2019.
3. M Gospodarowicz, L Benedet, RV Hutter, I Fleming, DE Henson, and LH Sobin. History and international developments in cancer staging. *Cancer prevention & control: CPC= Prevention & controle en cancerologie: PCC*, 2(6):262–268, 1998.
4. Edge S, Compton C (2010) "The american joint committee on cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM," *Annals of Surgical Oncology*, vol. 17, no. 6, pp. 1471–1474
5. Gordetsky J, Epstein J (2016) Grading of Prostatic Adenocarcinoma: Current State and Prognostic Implications. Diagnostic Pathology. 11:25 https://www.overleaf.com/project/5bbcd6e10300fa2af19e3a24
6. Epstein J, Zelefsky M, Sjoberg D, Nelson J, Egevad L, Magi-Galluzzi C, et al. (2016) A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. *European Urology*. 69(3):428-35

7. Lexander H, Palmberg C, Hellman U, Auer G, Hellström M, Franzén B, Jörnvall H, Egevad L (2006) "Correlation of protein expression, gleason score and dna ploidy in prostate cancer," *Proteomics*, vol. 6, no. 15, pp. 4370–4380
8. Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403-410
9. Trapnell C, Pachter L, Salzberg S (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25876 1, 1105-1111
10. Fusun Citak-Er, Metin Vural, Omer Acar, Tarik Esen, Aslihan Onay, and Esin Ozturk-Isik. Final gleason score prediction using discriminant analysis and support vector machine based on preoperative multiparametric mr imaging of prostate cancer at 3t. *BioMed research international*, 2014, 2014.
11. Alkhateeb A, Rezaeian I, Singireddy S, Cavallo-Medved D, Porter L, and Rueda L. Transcriptomics signature from next-generation sequencing data reveals new transcriptomic biomarkers related to prostate cancer. *Cancer informatics*, 18:1176935119835522, 2019.
12. Arvaniti A, Fricker K, Moret M, Rupp N, Hermanns T, Fankhauser C, Wey N, Wild P, Rueschoff J, and Claassen M. Automated gleason grading of prostate cancer tissue microarrays via deep learning. *BioRxiv*, page 280024, 2018.
13. Citak-Er F, Vural M, Acar O, Esen T, Onay A, Ozturk-Isik E (2014) Final Gleason score prediction using discriminant analysis and support vector machine based on preoperative multiparametric MR imaging of prostate cancer at 3T. *BioMed Research International*.
14. Hamzeh O, Alkhateeb A, Rezaeian I, Karkar A, and Rueda L. Finding transcripts associated with prostate cancer gleason stages using next generation sequencing and machine learning techniques. In *International Conference on Bioinformatics and Biomedical Engineering*, pages 337–348. Springer, 2017.
15. Prostate Adenocarcinoma TCGA-PRAD dataset; 2017., 2017 https://portal.gdc.cancer.gov/projects/TCGA-PRAD[Online; Last accessed July 2017].
16. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov
17. Gross M, Liu B, Tan J, French F, Carey M, Shuai K (2001)"Distinct effects of PIAS proteins on androgen-mediated gene activation in prostate cancer cells," *Oncogene*, vol. 20, no. 29, p. 3880
18. Izumi K, Fang L, Mizokami A, Namiki M, Li L, Lin W, Chang C (2013) "Targeting the androgen receptor with sirna promotes prostate cancer metastasis through enhanced macrophage recruitment via ccl2/ccr2-induced stat3 activation," *EMBO Molecular Medicine*, vol. 5, no. 9, pp. 1383–1401
19. Zhang Q, Raghunath P, Xue L, Majewski M, Carpentieri D, Odum N, Morris S, Skorski T, Wasik M (2002) "Multilevel dysregulation of Stat3 activation in anaplastic lymphoma Kinase-positive t/null-cell Lymphoma," *The Journal of Immunology*, vol. 168, no. 1, pp. 466–474
20. Ogata Y, Osaki T, Naka T, Iwahori K, Furukawa M, Nagatomo I, Kijima T, Kumagai T, Yoshida M, Tachibana I, et al. (2006) "Overexpression of pias3 suppresses cell growth, restores the drug sensitivity of human lung cancer cells in association with pi3-k/akt inactivation," *Neoplasia*, vol. 8, no. 10, pp. 817–825
21. Nicolas E, Arora S, Zhou Y, Serebriiskii I, Andrake M, Handorf E, Bodian D, Vockley J, Dunbrack R, Ross E et al. (2015) "Systematic evaluation of underlying defects in dna repair as an approach to case-only assessment of familial prostate cancer," *Oncotarget*, vol. 6, no. 37, p. 39614
22. Santarpia L, Iwamoto T, Di Leo A, Hayashi N, Bottai G, Stampfer M, André F, Turner F, Symmans W, Hortobágyi G et al. (2013) "DNA repair gene patterns as prognostic and predictive factors in molecular breast cancer subtypes," *The Oncologist*, vol. 18, no. 10, pp. 1063–1073
23. Yi Zhao, Marcus JC Long, Yiran Wang, Sheng Zhang, and Yimon Aye. UBE2v2 is a rosetta stone bridging redox and ubiquitin codes, coordinating dna damage responses. *ACS Central Science*, 4(2):246–259, 2018.
24. Jizhong Ren, Xiuwu Pan, Lin Li, Yi Huang, Hai Huang, Yi Gao, Hong Xu, Fajun Qu, Lu Chen, Linhui Wang, et al. Knockdown of gpr137, g protein-coupled receptor 137, inhibits the proliferation and migration of human prostate cancer cells. *Chemical Biology & Drug Design*, 87(5):704–713, 2016.
25. Ghanshyam Upadhyay, Asif H Chowdhury, Bharat Vaidyanathan, David Kim, and Shireen Saleque. Antagonistic actions of rcor proteins regulate LSD1 activity and cellular differentiation. *Proceedings of the National Academy of Sciences*, 111(22):8071–8076, 2014.
26. Gordetsky J and Epstein J. Grading of prostatic adenocarcinoma: current state and prognostic implications. *Diagnostic pathology*, 11(1):25, 2016.

27. Schulz W, Ingenwerth M, Djuidje C, Hader C, Rahnenführer J, Engers R (2010) "Changes in cortical cytoskeletal and extracellular matrix gene expression in prostate cancer are related to oncogenic erg deregulation," *BMC Cancer*, vol. 10, no. 1, p. 505
28. Ji Z, Shi X, Liu X, Shi Y, Zhou Q, Liu X, Li L, Ji X, Gao Y, Qi Y, et al.(2012) "The membrane-cytoskeletal protein 4.1 n is involved in the process of cell adhesion, migration and invasion of breast cancer cells," *Experimental and Therapeutic Medicine*, vol. 4, no. 4, pp. 736–740
29. Seabra A, Araújo T, Mello F, Alcântara D, De Barros D, DE Assumpção P, Montenegro R, Guimarães A, Demachki S, Burbano R (2014) "High-density array comparative genomic hybridization detects novel copy number alterations in gastric adenocarcinoma," *Anticancer Research*, vol. 34, no. 11, pp. 6405–6415
30. Ralf Buettner, Linda B Mora, and Richard Jove. Activated stat signaling in human tumors provides novel molecular targets for therapeutic intervention. *Clinical Cancer Research*, 8(4):945–954, 2002.
31. Paula Kroon, Paul A Berry, Michael J Stower, Greta Rodrigues, Vincent M Mann, Matthew Simms, Deepak Bhasin, Somsundaram Chettiar, Chenglong Li, Pui-Kai Li, et al. Jak-stat blockade inhibits tumor initiation and clonogenic recovery of prostate cancer stem-like cells. *Cancer research*, 73(16):5288–5298, 2013.
32. Jason S Rawlings, Kristin M Rosler, and Douglas A Harrison. The JAK/Stat signaling pathway. *Journal of Cell Science*, 117(8):1281–1283, 2004.
33. Leslie Tam, Liane M McGlynn, Pamela Traynor, Rono Mukherjee, John MS Bartlett, and Joanne Edwards. Expression levels of the jak/stat pathway in the transition from hormone-sensitive to hormone-refractory prostate cancer. *British Journal of Cancer*, 97(3):378, 2007.
34. Qi Long, Jianpeng Xu, Adeboye O Osunkoya, Soma Sannigrahi, Brent A Johnson, Wei Zhou, Theresa Gillespie, Jong Y Park, Robert K Nam, Linda Sugar, et al. Global transcriptome analysis of formalin-fixed prostate cancer specimens identifies biomarkers of disease recurrence. *Cancer research*, 74(12):3228–3237, 2014.
35. Rasko Leinonen, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic acids research*, 39(suppl_1):D19–D21, 2010.
36. Dobin A, Davis C, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras T (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15-21
37. Li B, Dewey C (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1), 1
38. Trapnell C, Hendrickson D, Sauvageau M, Goff L, Rinn J, Pachter L (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, 31(1), 46-53. ISBN 0716776014
39. Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq.*Nature Methods* 5(7):621–8. doi:10.1038/nmeth.1226
40. Gerard V Trunk. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3):306–307, 1979.
41. Trapnell C, Williams B, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.*Nat Biotechnol* 28(5):511–5. doi:10.1038/nbt.1621
42. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.
43. Laurikkala, J. Improving Identification of Difficult Small Classes by Balancing Class Distribution. Tech. Rep. A-2001-2, University of Tampere, 2001.
44. Novakovic J (2009) Using information gain attribute evaluation to classify sonar targets. *In 17th Telecommunications forum TELFOR* (pp. 24-26)
45. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Aalysis and Machine Intelligence*, 27(8), 1226-1238
46. Scikit-learn: Machine Learning in Python, Pedregosa et al.,*JMLR* 12, pp. 2825-2830, 2011.
47. Domingos P, Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, pp. 29(2-3): 103-130
48. Cortes C, Vapnik V (1995) Support-vector networks. *Machine learning*, 20(3), 273-297
49. Canadian Cancer Society. Canadian Cancer Societys Advisory Committee on Cancer Statistics. Canadian Cancer Statistics 2016. Toronto, ON: Canadian Cancer Society; 2018., 2018. http:

//www.cancer.ca/media/cancer.ca/CW/cancerinformation/cancer101/Canadiancancerstatistics/Canadian-Cancer-Statistics-2016-EN.pdf [Online; Last accessed December 2016].

50. Xu J, Zheng S, Komiya A, Mychaleckyj C, Isaacs S, Chang B, Bleecker E (2003) Common sequence variants of the macrophage scavenger receptor 1 gene are associated with prostate cancer risk. *The American Journal of Human Genetics*, 72(1), 208-212
51. Loddsh H, Brrk A, Kaiser C, Krieger M, Scott M, Bretscher A, Ploegh H, and Mutsudaira P (2007). *Molecular Cell Biology*. W. H. Feeeman, 6th Edition, ISBN 0716776014
52. Clancy S, Brown W (2008) Translation: DNA to mRNA to protein. *Nature Education*, 1(1), 101
53. Li J, Yen C, Liaw D, Podsypanina K, Bose S, Wang S, et al. (1997). PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science*, 275(5308), 1943-1947
54. Trapnell C, Pachter L, Salzberg S (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9), 1105-1111
55. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11), 1851-1858
56. Zhang F, Drabier R (2012) IPAD: the integrated pathway analysis database for systematic enrichment analysis. *BMC Bioinformatics*, 13(Suppl 15):S7