Peer-reviewed version available at *Diagnostics* **2019**, *9*, 219; doi:10.3390/diagnostics9040219

Article

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

Osama Hamzeh [1,‡], Abedalrhman Alkhateeb [1,‡]*, Julia Zheng [1], Srinath Kandalam [1], Crystal Leung [2], Govindaraja Aikukke [3], Dora Cavallo-Medved [1], Nallasivam Palanisamy [4], and Luis Rueda [1,*]

1 School of Computer Science, University of Windsor 401 Sunset Ave, Windsor, ON, Canada, N9B 3P4; hamzeho, alkhate, fzheng12z, kandala1, dcavallo, lrueda@uwindsor.ca
2 Schulich School of Medicine and Dentistry, Western University 1151 Richmond St, London, ON, Canada N6A 5C1; cleung2021@meds.uwo.ca
3 ITOS Oncology 1453 Prince Rd, Windsor, ON N9C 3Z4 Ste: 4125 gatikukke@itosoncology.com
4 Department of Urology, Henry Ford Health System One Ford Place, Detroit, MI,USA 48202; NPALANI1@hfhs.org
* 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 and North American men is prostate cancer. Gleason score is a pathological grading system to examine the potential aggressiveness of the disease in the prostate tissue. The advancement in computing and next-generation sequencing technology now allow us to study the genomic profiles of patients in association with their different Gleason score more accurately and effectively. 2) Methods: In this study, we used a novel machine learning method to analyze gene expression of prostate tumors with different Gleason scores, and identify potential genetic biomarkers for each Gleason group. We obtained a publicly-available RNA-Seq dataset of a cohort of 104 prostate cancer patients from the National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO) repository, and categorized patients based on their Gleason scores to create a hierarchy of disease progression. A hierarchical model with standard classifiers in different Gleason groups, also known as *nodes*, was developed to identify and predict nodes based on their mRNA or gene expression. In each node, patient samples were analyzed via class imbalance and hybrid feature selection techniques to build the prediction model. The outcome from analysis of each node is a set of genes that can differentiate each 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 collected from cBioportal [1]. 3) Results: The overall accuracy of applying this novel method to the first dataset was 93.3%, and further validated to 87% accuracy using the second dataset. This method also identified genes that were not previously reported as potential biomarkers for specific Gleason groups. In particular, PIAS3 was identified as a potential biomarker for Gleason score 4+3=7, and UBE2V2 for Gleason score 6. 4) Insight: Previous reports show that the genes predicted by this newly proposed method strongly correlate with prostate cancer development and progression. Furthermore, pathway analysis shows that both PIAS3 and UBE2V2 share similar protein interaction pathways, the JAK/STAT signaling process.

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

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 cause of death worldwide. Among males, prostate cancer is the most incident cancer type, with 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 devised 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]. In prostate cancer, these characters are also used to assign a metric of tissue organization and disease aggressiveness called the 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 [6]. Epstein et al., however, indicated that Scores 2-5 are no longer assigned to the tissue and these multiple scores can be categorized together with score 6 as group 1, yielding to categories as depicted in Table 1, and used to determine prognosis of disease. 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 and groups. This categorization strategy has been shown to clearly indicate cancer recurrence, and improve the prognostic role of the Gleason score [7].

Recent prostate cancer research has greatly focused on identifying gene expression patterns that correlate with disease progression, and can be used as predictive tools for patient treatment and outcome. Moreover, 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 [10]. Based on the need for understanding the biological basis of the visual Gleason microscopic assessment, Roberto et al. conducted a gene expression profiling on 2 groups of Gleason score 6 and 7 or high using metabolic gene panel. The used panel consists of many gene members of JAK/STAT pathway [11]. In the present study, we analyze the transcription level of different Gleason scores to find genes that can identify a specific Gleason group from the others.

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 [12]. Arvaniti et al. proposed a deep learning approach to predict Gleason scores [13]. 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 pathologists. Although the study by Arvaniti et al. reported a decent performance measurements

Peer-reviewed version available at Diagnostics 2019, 9, 219; doi:10.3390/diagnostics9040219

(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 [14]. 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%).

The focus of this study is to identify genes that can be used to differentiate specific Gleason groups. This work is an extension of our previously proposed prediction model, which is based on analyzing the RNA-Seq data from patients with different Gleason scores [15]. 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 used in this study 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.

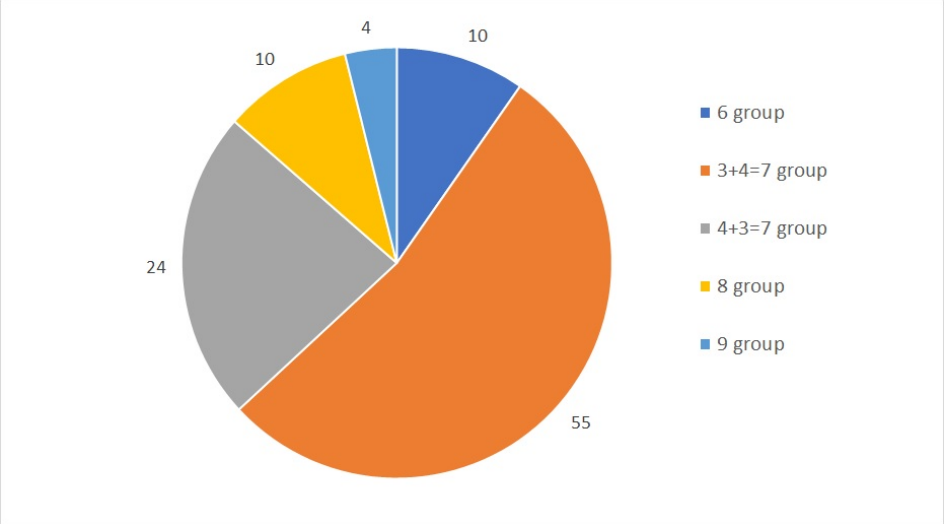

Figure 1. Gleason groups and their distributions.

The model also identified six gene transcripts that are differentially expressed in the five different Gleason scores. Of these, the corresponding genes shown in Tables 4, 5, 3 and 6 are the most relevant in identifying prostate cancer with the Gleason scores using the hierarchical method illustrated in Figure 2. Different classification methods for each stage within the hierarchy are shown in Table 2. The first node of the hierarchy yields 94% accuracy in identifying Gleason score 3+4=7 versus the other scores. The samples are then passed through node 2, in which Gleason score 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 score 6 is identified with 100% accuracy. The remaining samples are finally processed in the last node, where the Gleason score 8 is identified from the Gleason score 9 with 100% accuracy. Due to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 November 2019 doi:10.20944/preprints201911.0298.v2

Peer-reviewed version available at Diagnostics 2019, 9, 219; doi:10.3390/diagnostics9040219

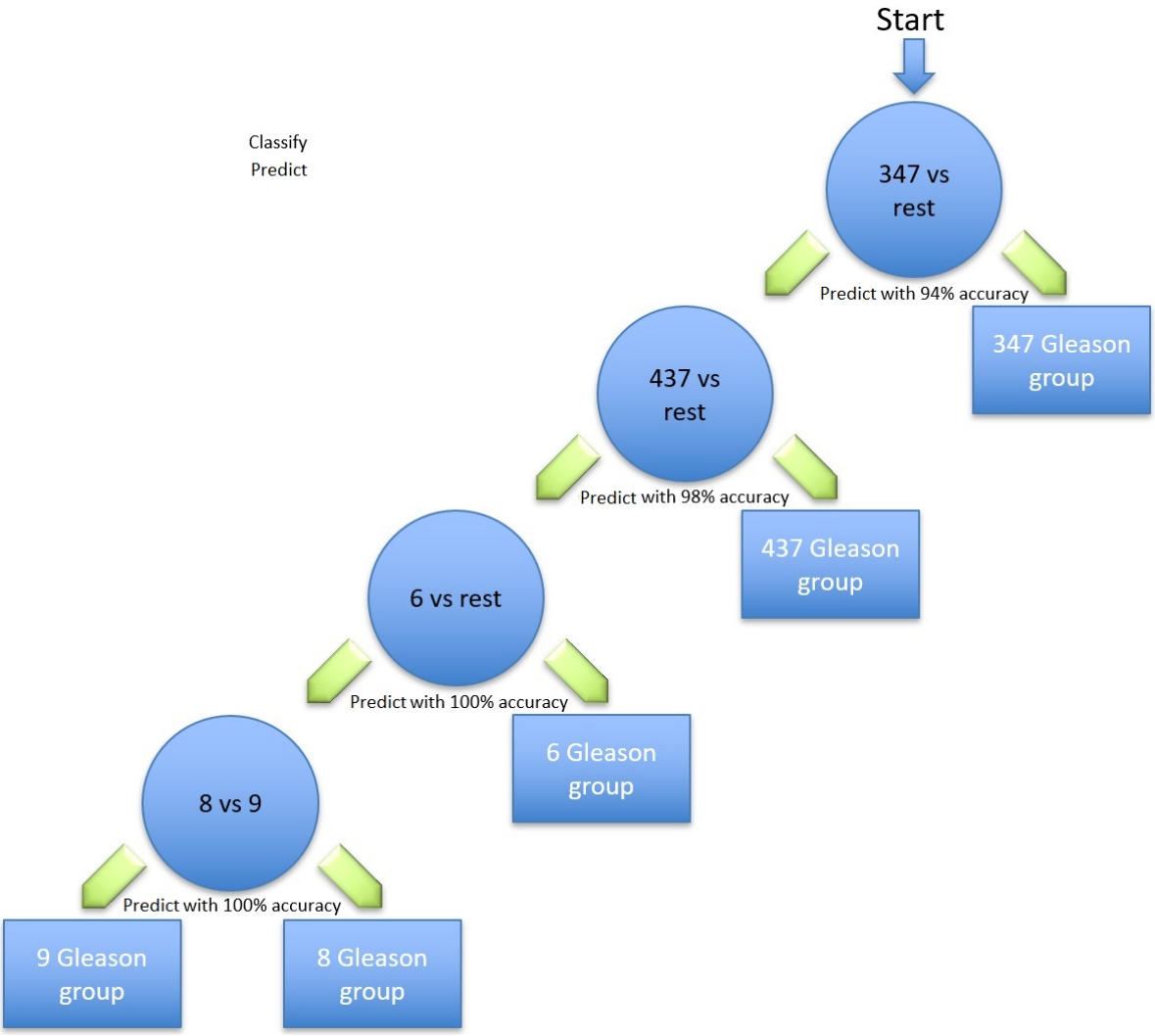

Figure 2. Hierarchical tree of classifications of Gleason groups against the rest, along with the corresponding classification accuracy.

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

Table 3. Set of resulting transcripts in Gleason group 1.

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 4. Set of resulting transcripts in Gleason group 2.

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 5. Set of resulting transcripts in Gleason group 3.

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)

Table 6. Set of resulting transcripts in Gleason group 4.

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

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 measurements are represented in the y-axis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 November 2019 doi:10.20944/preprints201911.0298.v2

Peer-reviewed version available at Diagnostics 2019, 9, 219; doi:10.3390/diagnostics9040219

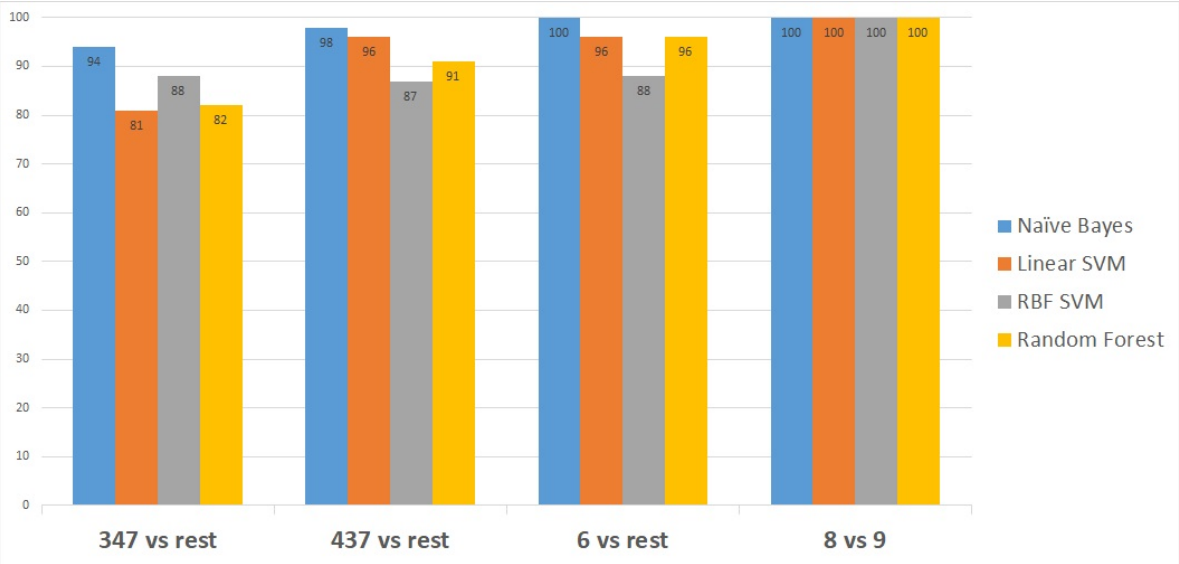

Figure 3. Accuracy obtained by each classifier for classifying one versus the rest for all five Gleason groups.

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

To further validate the model,we applied the method on a second publicly-available dataset [16] obtained from the National Center for Biotechnology Information (NCBI) portal [17].This second dataset contains gene expressions for 498 patient samples. The proposed model showed an excellent prediction accuracy on the 498 patients' gene expressions. The prediction accuracy for all the Gleason scores was above 90% except for the 4+3=7 Gleason score vs rest Figure (4).

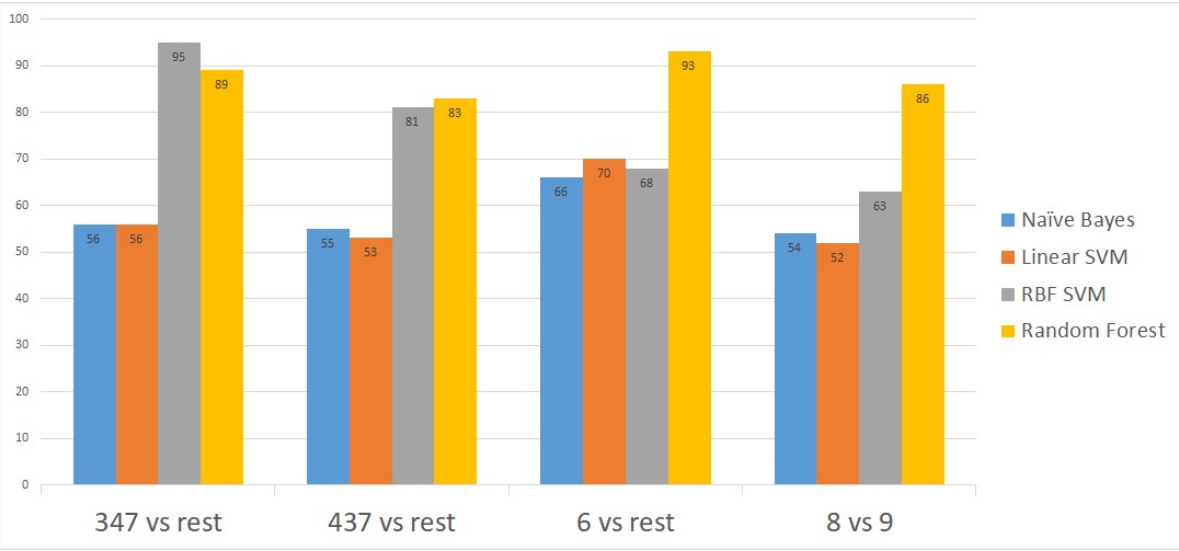

Figure 4. Classification accuracies obtained after applying the model on the second dataset.

3. Discussion

Many of the genes that encode the differentially expressed transcripts identified in this study have been previously shown to play various roles in cancer. Some have been shown to promote cancer

progression, while other play a protective role. For example, UBE2V2, whose gene's transcript was selected in the third node of our hierarchical model, has been shown to protect cells by mediating DNA repair functions [18]. In familial prostate cancer, however, a high frequency variant of UBE2V2 was identified and found to affect DNA repair and androgen signaling [19]. In our model study, different quantification of UBE2V2 transcript was able to predict Gleason score 6 (group 1) in the first dataset. Differential expression of UBE2V2 has also been associated with poor prognosis in breast cancer [20].

Our study also revealed that the differential expression of GPR137 expression and EPB41L1 is associated with tumors of Gleason scores 3+4=7 and 8, respectively. Earlier studies show that proteins encoded by EPB41L1 are associated with the proper organization of the cell cytoskeleton, and that EPB41L1 plays an important role in the negative regulation of cell metastasis, migration and invasion. Expression in EPB41L1 has been observed to be lower in prostate cancer compared to normal cells. Although it remains unclear, disruption of normal EPB41L1 expression may play an important role in disorganized cell and tissue structures associated with higher grade prostate cancer [21], and thus linking its deregulation to prostate cancer progression and prognosis. Furthermore, reduced expression of EPB41L1 plays an important role in recurrence and has been associated with highly metastatic lung and breast cancer [22]. EPB41L1 was also shown to be differentially expressed in gastric cancer [23]. On the other hand, GPR137 expression has been shown to be upregulated in prostate cancer tissues as compared with in paracancerous tissues. Moreover, knockdown of GPR137 resulted in decreased cell proliferation and colony formation in PC-3 and DU145 prostate cancer cell lines, and associated with cell cycle arrest at G0/G1 phase. GPR137 suppression also decreases the migration and invasive abilities of PC-3 cells, thus suggesting GPR137 plays in prostate cancer progression and metastasis [24].

Differential expression of PIAS3 and Rest Corepressor 3 (Rcor3) were both associated with tumors of Gleason score 4+3=7. While very little is known about the role of Rest Corepressor 3 (Rcor3) in prostate cancer, it has been shown to act as an antagonist of cell differentiation [25], a characteristic of prostate tumors with Gleason score 4+3=7 [26]. On the other hand, differential PIAS3 expression has been observed in a variety of human cancers including lung, breast, prostate, colon-rectum, and brain tumors [27]. PIAS3 is expressed in prostate cancer cells and its expression is induced in response to androgen [28,29]. Although PIAS has been shown to enhance the transcriptional activity of androgen receptors (AR) in prostate cancer cells, other studies have revealed that ectopic overexpression of PIAS3 suppresses AR-mediated gene activation induced by dihydrotestosterone (DHT) [30]. PIAS3 acts as a negative regulator of AR transcriptional activity and signaling through direct protein-protein interaction. Recent findings have also revealed that AR is also differentially correlated with Gleason score patterns in both primary and metastatic prostate cancer, where it is upregulated in Gleason group 4 and downregulated in Gleason pattern 5.

PIAS3 is a member of the mammalian PIAS family consists of four members: PIAS1, PIAS2, PIAS3 and PIAS4 [31]. PIAS3 protein directly binds to several transcription factors and either blocks or enhances their activity. PIAS3 is also specific inhibitor of signal transducer and activator of transcription 3 (STAT3), a transcription factor and member of the Janus Kinase (JAK)/STAT signaling pathway ([32,33]) This signaling pathway has been a target of interest in many cancer studies in recent years. In prostate cancer, the expression levels of JAK/STAT have been shown to impact the progression of the disease [34,35]. As an inhibitor of STAT3, PIAS3 blocks the transactivation and binding of STAT3 to specific DNA elements via protein-protein interactions, thereby inhibiting STAT3-mediated gene activation. Figure 5 depicts the protein-protein interaction among genes in 4+3=7 and 6 scores 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 UBE2V2 share the same protein interaction network.

PIAS3 is also the only member of the PIAS family that has been shown to directly interact with Stat5a/b and repress Stat5-mediated transcription [36]. Stat5a/b is constantly active in human prostate cancer [37], associated with high histological grade [38], and a predictor of early prostate cancer

Peer-reviewed version available at Diagnostics 2019, 9, 219; doi:10.3390/diagnostics9040219

recurrence [39]. Transcription factor Stat5a/b has been shown to regulate the viability and growth of human prostate cancer cells [40,41]. Moreover, in vitro inhibition of Stat5a/b induces apoptosis in human prostate cancer cells [37,42]. In vivo, Stat5a/b inhibition blocks prostate cancer subcutaneous and orthotopic xenograft tumor growth in nude mice [42]. Although, studies have revealed an inhibitory role for PIAS3 against Stat5a/b-driven gene transcription and disease progression in breast cancer, the predominant Stat5a/b protein that binds to DNA has been shown to be N-terminally truncated in human prostate cancer cells and clinical prostate cancers [43]. Further studies have demonstrated that the N-domain of Stat5a/b binds to PIAS3. Hence the truncated form of Stat5 in prostate cancer cells evades PIAS3-medeiated transcriptional inhibition thereby increasing prostate cancer growth and progression. Thus, the proteolytic cleavage of the N-terminus of Stat5a/b may be a mechanism by which Stat5 evades the transcriptional repression by PIAS3 in prostate cancer cells. This further indicates the complexity of intracellular protein interactions and its role in disease progression.

Our study applies a novel machine learning model to identify differentially expressed, prostate cancer stage-specific transcripts. Although the application of this model to other related datasets is required to further valid our findings, the use of this model in conjunction with in vitro and in vivo biological studies will aid in elucidating the intricate molecular relationships between the identified transcripts. Moreover, this will provide more insight into predicted prognostic outcomes and the development of effective therapeutic strategies against prostate cancer progression.

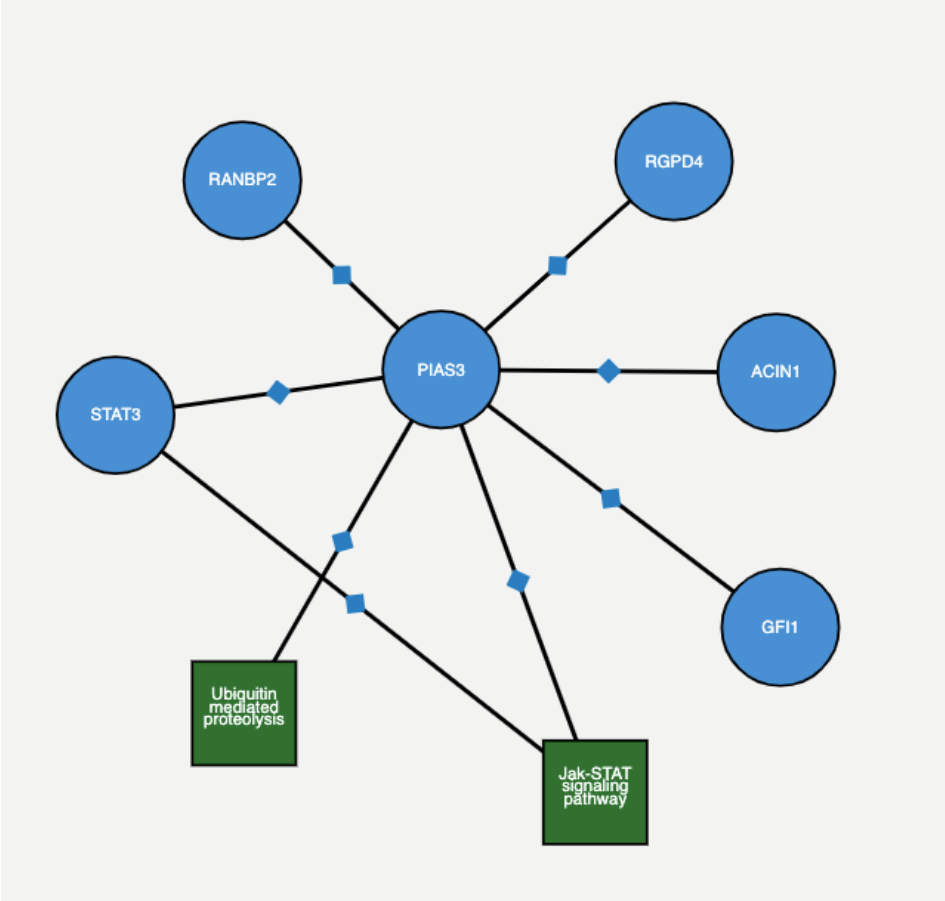

Figure 5. An interactive figure taken from proteomics database STRING, where it shows neighbouring protein binding and pathway interactions for a given gene using STRING and KEGG pathway analysis. Here, the gene of interest is PIAS3, an identified possible biomarker in the 4+3=7 score. The figure shows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

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.

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 GSE54460 [44]. This RNAseq prostatectomy dataset was generated 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 [45], as depicted in Figure 6. The process continues by incorporating the STAR aligner [10] to align the samples reads into the Human genome (hg19). Then, the process assembles the transcripts and quantifies the reads into the assembled transcripts using RSEM [46]. 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 significant amount of raw data in a snapshot. However, the underlying process yields artifacts, and pre-processing must be done before the downstream analysis. These artifacts include duplication and bias reads [47], among others. Counting the reads that are assembled by mapping them to the Human genome gives accurate indicators of 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) [48]. Also, the reason for choosing TPM instead of fragments per kilobase per million (FPKM) [49] is that TPM normalizes the reads to the length of the gene first, which makes it easier to compare the quantified reads among different samples.

4.1. Class Imbalance

Some classes have a markedly lower number of samples than the others, which may cause some classifiers to become biased towards the majority class. To solve this problem, multiple resampling methods were deployed and tested to identify the specific method that would yield the best solution for a particular dataset. After applying multiple oversampling and under-sampling methods, the best option was found to be synthetic minority oversampling technique (SMOTE) [50] for oversampling the minority class, while the neighborhood cleaning rule (NCL) [51] 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 that connects each sample and introduces 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

Peer-reviewed version available at Diagnostics 2019, 9, 219; doi:10.3390/diagnostics9040219

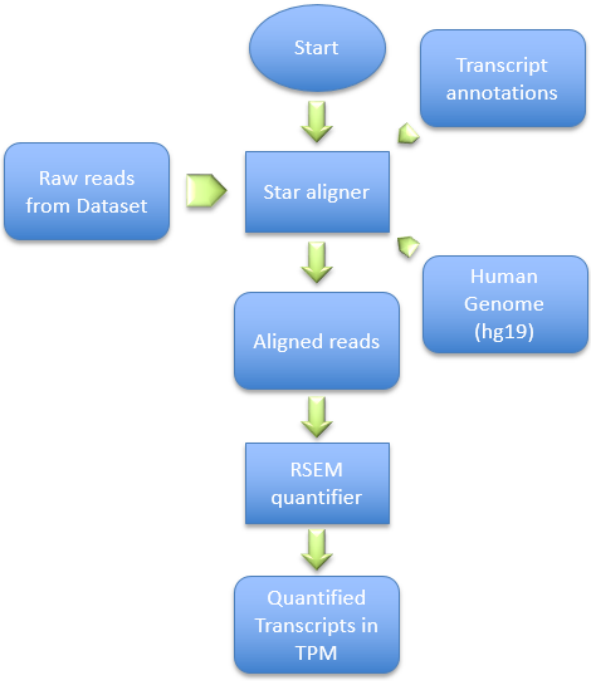

Figure 6. Pre-processing steps of the proposed method.

calculated by measuring the Euclidean distance between the two samples and multiplying that value by a random number between 0 and 1. Figure 7 shows a hypothetical example of 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. The blue points represent the original samples, while the amber points represent the synthetically generated samples.

4.2. Feature Selection

As the output of the pre-processing step, the method retrieved 41,971 transcripts along with their corresponding quantifications measured by TPM. Such a large number of transcripts leads to a complex classification model, mostly due to the curse of dimensionality [52]. 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, which is 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 (transcripts) with a score for each attribute showing its effect on the actual class. Then, 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 [53]. IG of attribute vector X concerning 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.

Peer-reviewed version available at Diagnostics 2019, 9, 219; doi:10.3390/diagnostics9040219

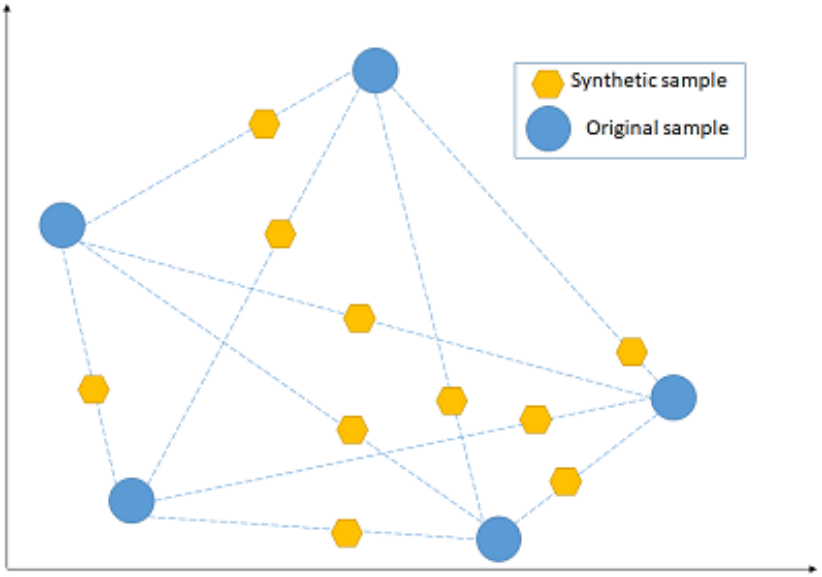

Figure 7. Hypothetical example that shows how synthetic minority oversampling technique (SMOTE) works.

After filtering the transcripts based on their IG score, a wrapper-based feature selection algorihtm that uses minimum redundancy maximum relevance (mRMR) is used to narrow down the most relevant, least redundant transcripts to a few per group; mRMR has the capability of incorporating any classifier to select features (transcripts) that minimize the redundancy while increasing the correlation to the class vector [54]. The wrapper method adds up the features that minimize redundancy (W_i), and maximize the relevance (V_i), with the best possible accuracy of a SVM classifier that uses a linear kernel, 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 is solved using the one-versus-rest approach. There are five different classes, which correspond to the five distinct Gleason groups. To apply a 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 to form the *negative* class. The classification pipeline resembles a binary tree structure, where each internal node is a binary classification problem (see Figure 2). Starting from the root, in the one-versus-rest classification, we remove the samples that belong to the chosen class earlier. We repeat the same steps of building datasets 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 the highest accuracy at each internal node.

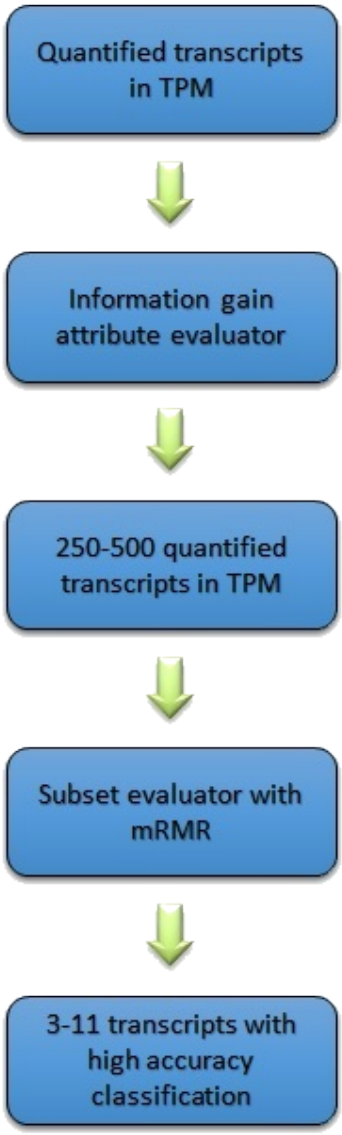

Figure 8. Machine learning pipeline used in the proposed method.

4.4. Identifying Transcripts within Different Gleason Scores

We used the Scitkit-learn [55] library to apply different classification algorithms on the final selected transcripts. This step identifies which transcripts can decide a Gleason group from the others based on their quantification values. Standard classifiers such as Naïve 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, while assuming that the features are independent of each other [56]. While being simple, Naïve Bayes has been shown to perform very well in many problems and avoid overfitting. An SVM classifier was also used to build a prediction model using the selected transcripts in the previous step [57]. The advantage of SVM is its exceptional generalization power, especially in high-dimensional data with a small number of samples. Figure 8 shows the pipeline followed in this study.

5. Conclusions and Future Directions

Identifying novel biomarkers that are clinically associated with specific Gleason groups in prostate cancer is vital for diagnosis and treatment of the disease. Utilizing NGS data and machine learning techniques, a supervised learning method is proposed to find group-specific sets of transcripts with significant different levels of quantification values. The transcripts, along with the corresponding genes, identified by the proposed machine learning method were found in the literature to play crucial roles in cancer pathogenesis; key transcripts were strongly related to prostate cancer. To validate the model, we also tested it on a gene expression dataset, showing that the resulting genes are related to prostate cancer progression.

The work presented in this paper opens the way for future direction of the research. One of these in to apply and adjust the same method 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. In addition, a multi-omics model based on different types of genomics data for this problem can be investigated, which 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. N. Palanisamy validated the idea, he shares 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 and clinical aspects of the problem. D. Cavallo and G. Atikkuke analyzed PIAS3 and UBE2VE roles in JAK/STAT pathway. All authors have participated in writing the paper and approved the final manuscript.

Funding: This research work has been 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
RPKM	reads per kilobase per million of reads
FPKM	Fragments per kilobase per million of reads
TPM	Transcripts per million of reads

References

1. cBioPortal for Cancer Genomics; 2019., 2019 https://cbioportal.org[Online; Last accessed July 2019].
2. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin D, Piñeros M, Znaor A, and Bray F.
3. M Gospodarowicz, L Benedet, RV Hutter, I Fleming, DE Henson, and LH Sobin.
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. Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403-410
6. Gordetsky J, Epstein J (2016) Grading of Prostatic Adenocarcinoma: Current State and Prognostic Implications. Diagnostic Pathology. 11:25

7. 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
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. 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
11. Domenica Roberto, Shamini Selvarajah, Paul C Park, David Berman, and Vasundara Venkateswaran. Functional validation of metabolic genes that distinguish Gleason 3 from Gleason 4 prostate cancer foci. *The Prostate*, 79(15):1777–1788, 2019.
12. 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.
13. 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.
14. 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.
15. 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.
16. Prostate Adenocarcinoma TCGA-PRAD dataset; 2019., 2019 https://portal.gdc.cancer.gov/projects/TCGA-PRAD[Online; Last accessed November 2019].
17. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov [Online; Last accessed July 2019].
18. 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.
19. 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.
20. Santarpia L, Iwamoto T, Di Leo A, Hayashi N, Bottai G, Stampfer M, André F, Turner F, Symmans W, Hortobágyi Get 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.
21. 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.
22. 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.
23. 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.
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. Liming Wang and Sipra Banerjee. Differential pias3 expression in human malignancy. *Oncology reports*, 11(6):1319–1324, 2004.
28. Vasily J Assikis, Kim-Anh Do, Sijin Wen, Xuemei Wang, Jeong Hee Cho-Vega, Shawn Brisbay, Remigio Lopez, Christopher J Logothetis, Patricia Troncoso, Christos N Papandreou, et al. Clinical and biomarker correlates of androgen-independent, locally aggressive prostate cancer with limited metastatic potential. *Clinical cancer research*, 10(20):6770–6778, 2004.
29. 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.
30. Liming Wang and Sipra Banerjee. Differential pias3 expression in human malignancy. *Oncology reports*, 11(6):1319–1324, 2004.
31. Nobuhide Ueki, Naohiko Seki, Kazuhiro Yano, Toshiyuki Saito, Yasuhiko Masuho, and Masa-aki Muramatsu. Isolation and chromosomal assignment of a human gene encoding protein inhibitor of activated stat3 (pias3). *Journal of human genetics*, 44(3):193–196, 1999.
32. D Schmidt and S Müller. Pias/sumo: new partners in transcriptional regulation. *Cellular and Molecular Life Sciences CMLS*, 60(12):2561–2574, 2003. Ke Shuai. Regulation of cytokine signaling pathways by pias proteins. *Cell research*, 16(2):196, 2006.
33. Ke Shuai. Regulation of cytokine signaling pathways by pias proteins. *Cell research*, 16(2):196, 2006.
34. Jason S Rawlings, Kristin M Rosler, and Douglas A Harrison. The JAK/Stat signaling pathway. *Journal of Cell Science*, 117(8):1281–1283, 2004.
35. 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.
36. Michael A Rycyzyn and Charles V Clevenger. The intranuclear prolactin/cyclophilin b complex as a transcriptional inducer. *Proceedings of the National Academy of Sciences*, 99(10):6790–6795, 2002.
37. Matti Ahonen, Minna Poukkula, Andrew H Baker, Masahide Kashiwagi, Hideaki Nagase, John E Eriksson, and Veli-Matti Kähäri. Tissue inhibitor of metalloproteinases-3 induces apoptosis in melanoma cells by stabilization of death receptors. *Oncogene*, 22(14):2121, 2003.
38. Hongzhen Li, Tommi J Ahonen, Kalle Alanen, Jianwu Xie, Matthew J LeBaron, Thomas G Pretlow, Erica L Ealley, Ying Zhang, Martti Nurmi, Baljit Singh, et al. Activation of signal transducer and activator of transcription 5 in human prostate cancer is associated with high histological grade. *Cancer Research*, 64(14):4774–4782, 2004.
39. Hongzhen Li, Ying Zhang, Andrew Glass, Tobias Zellweger, Edmund Gehan, Lukas Bubendorf, Edward P Gelmann, and Marja T Nevalainen. Activation of signal transducer and activator of transcription-5 in prostate cancer predicts early recurrence. *Clinical Cancer Research*, 11(16):5863–5868, 2005.
40. Yi-Chun Liao and Su Hao Lo. Deleted in liver cancer-1 (dlc-1): a tumor suppressor not just for liver. *The international journal of biochemistry & cell biology*, 40(5):843–847, 2008.
41. Shyh-Han Tan and Marja T Nevalainen. Signal transducer and activator of transcription 5a/b in prostate and breast cancers. *Endocrine-Related Cancer*, 15(2):367–390, 2008.
42. Ayush Dagvadorj, Robert A Kirken, Benjamin Leiby, James Karras, and Marja T Nevalainen. Transcription factor signal transducer and activator of transcription 5 promotes growth of human prostate cancer cells in vivo. *Clinical Cancer Research*, 14(5):1317–1324, 2008.
43. Ayush Dagvadorj, Shyh-Han Tan, Zhiyong Liao, Jianwu Xie, Martti Nurmi, Kalle Alanen, Hallgeir Rui, Tuomas Mirtti, and Marja T Nevalainen. N-terminal truncation of stat5a/b circumvents pias3-mediated transcriptional inhibition of stat5 in prostate cancer cells. *The international journal of biochemistry & cell biology*, 42(12):2037–2046, 2010.
44. 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.
45. 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.
46. Li B, Dewey C (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1), 1.

47. 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.
48. 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.
49. 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
50. 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.
51. Laurikkala, J. Improving Identification of Difficult Small Classes by Balancing Class Distribution. Tech. Rep. A-2001-2, University of Tampere, 2001.
52. Gerard V Trunk. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3):306–307, 1979.
53. Novakovic J (2009) Using information gain attribute evaluation to classify sonar targets. *In 17th Telecommunications forum TELFOR* (pp. 24-26)
54. 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.
55. Scikit-learn: Machine Learning in Python, Pedregosa et al.,*JMLR* 12, pp. 2825-2830, 2011.
56. 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.
57. Cortes C, Vapnik V (1995) Support-vector networks. *Machine learning*, 20(3), 273-297.