7-7-2020

Machine Learning Approaches for Identifying Cancer Biomarkers Using Next Generation Sequencing

Osama Hamzeh
University of Windsor

Machine Learning Approaches for Identifying Cancer Biomarkers Using Next Generation Sequencing

by

Osama Hamzeh

A Dissertation
Submitted to the Faculty of Graduate Studies
through the School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy at the
University of Windsor

Windsor, Ontario, Canada
2020

Machine Learning Approaches for Identifying Cancer Biomarkers Using Next Generation Sequencing

by

Osama Hamzeh

APPROVED BY:

F. Wu , External Examiner

University of Saskatchewan

P. Karpowicz

School of Biomedical Sciences

A. Ngom

School of Computer Science

S. Samet

School of Computer Science

L. Rueda, Advisor

School of Computer Science

March 20, 2020

Declaration Of Co-Authorship / Previous Publication

I. Co-Authorship

I hereby declare that this thesis incorporates material that is result of joint research, as follows: Chapter 2 of the thesis was co-authored with A. Alkhateeb, J. Zheng, S. Kandalam,C. Leung, G. Atikukke, D. Cavallo-Medved, N. Palanisamy and L. Rueda. Everyone contributed in finalizing the idea and follow-up discussions, O. Hamzeh has implemented the method, pre-processed the data, experimental design, and the data analysis. J. Zheng, S. Kandalam,C. Leung, G. Atikukke, D. Cavallo-Medved and N. Palanisamy assisted in the biological significant of the finding of the study, O. Hamzeh wrote the contents of the chapter, and L. Rueda assisted in reviewing the content of this chapter.

Chapter 3 of the thesis was co-authored with A. Alkhateeb, J. Zheng, S. Kandalam, and L. Rueda. Everyone contributed in finalizing the idea and follow-up discussions. O. Hamzeh implemented the method, pre-processed the data, conducted experimental design, and performed data analysis. J. Zheng and S. Kandalam assisted in the biological significant of the finding of the study, O. Hamzeh wrote the contents of the chapter, L. Rueda and A. Alkhateeb elaborated the main ideas of the chapter and assisted in reviewing the content of the chapter.

Chapter 3 of the thesis was co-authored with L. Rueda. We both contributed in finalizing the idea and following-up discussions. O. Hamzeh implemented the method, pre-processed the data, conducted experimental design, and performed data analysis. O. Hamzeh wrote the contents of the chapter, and L. Rueda assisted in reviewing the content of the chapter. I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from each of the co-author(s) to include the above material(s) in my thesis. I certify that, with the above qualification, this thesis, and the research to which it refers,is the product of my own work.

II. Previous Publication

This thesis includes three original papers that have been previously published/submitted for publication in conferences and peer reviewed journals, as follows:

Dissertation chapter	**Publication title**	**Publication Status**
Chapter 2	O.Hamzeh, A. Alkhateeb, J. Zheng, S. Kandalam, C. Leung, G. Atikukke, D. Cavallo-Medved, N. Palanisamy and L. Rueda, A Hierarchical Machine Learning Model to Discover Gleason Grade-Specific Biomarkers in Prostate Cancer, Diagnostics(2019), 9(4), 219	Published
Chapter 3	O. Hamzeh, A. Alkhateeb, J. Zheng, S. Kandalam and L. Rueda, Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data, BMC Bioinformatics, Special_Issue(2020)	In print
Chapter 4	O. Hamzeh, & L. Rueda. A Gene-disease-based Machine Learning Approach to Identify Prostate Cancer Biomarkers. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (pp. 633-638). ACM., (2019, September)	Invited to a journal

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as a graduate student at the University of Windsor.

III. General

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix. I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution. I understand that my thesis may be made electronically available to the public.

Abstract

Identifying biomarkers that can be used to classify certain disease stages or predict when a disease becomes more aggressive is one of the most important applications of machine learning. Next generation sequencing (NGS) is a state-of-the-art method that enables fast sequencing of DNA or RNA samples. The output usually contains a very large file that consists of base pairs of DNA or RNA. The generated data can be analyzed to provide gene expression, chromosome counting, detection of mutations on the genes, and detecting levels of copy number variations or alterations in specific genes, just as examples. NGS is leading the way to explore the human genome, enabling the future of personalized medicine. In this thesis, a demonstration is done on how machine learning is used extensively to identify genes that can be used to predict prostate cancer stages with very high accuracy, using gene expression. We have also been successful in predicting the location of prostate tumors based on gene expression.

In addition, traditional biomarker identification approaches, typically, use machine learning techniques to identify a number of genes and macromolecules as biomarkers that can be used to diagnose specific diseases or states of diseases with very high accuracy, using molecular measurements such as mutations, gene expression, copy number variations, and others. However, experts' opinions and knowledge is required to validate such findings. We, therefore, also introduce a new machine learning model that incorporates a knowledge-assisted system used to integrate the findings of the DisGeNET database, which is a framework that contains proven relationships among diseases and genes. The machine learning pipeline starts by reducing the number of features using a filter-based feature selection method. The DisGeNET database is used to score each gene related to the given cancer name. Then, a wrapper-based feature-selection algorithm picks the best set of genes with the highest classification accuracy. The method has been able to retrieve key genes from multiple data sets that classify with very high accuracy, while being biologically relevant, and no human intervention needed. Initial results provide a high area-under-the-curve with a handful of genes that are already proven to be related to the relevant disease and state based on the latest published medical findings. The proposed methods results provide biomarkers that can be verified in wet lab environments and can then be further analyzed and studied for diagnostic purposes.

Dedication

This thesis is dedicated to the love of my life 'Nada', my family and to the soul of my mother; may she rest in peace.

Acknowledgements

I would like to thank all the people who made this thesis possible, especially my supervisor Prof. Luis Rueda. He was very close to everyone in the Pattern Recognition and Bioinformatics lab. He always supported us, guided us, and motivated us. Even when we lost hope in certain areas, he would help us pursue a new area that would enable us to strive again. The whole lab worked as a family; we were always holding our weekly meetings to discuss our work, and that meeting made us feel more supportive to each other. We were able to see what our colleagues were doing, which gave us an insight of things to come, and if we needed help, they would assist us in the best way possible.

I would also like to thank Prof. Alioune Ngom who acted as one of the internal committee members. He was also a great mentor to me; his guidance and his smiling face were always there to help me and my lab teammates. Big thanks go also to Prof. Saeed Samet, who was in my thesis committee, and he was also the Ph.D. coordinator, making sure that our needs and concerns were always heard. I would like to sincerely thank Prof. Phillip Karpowicz, who was the external reader of my thesis; he assisted me in understanding much more about the biological aspects of my thesis.

My tokens of appreciation go to Prof. FangXiang Wu, for his willingness to be the external examiner of this thesis, and for taking time to read this thesis. I was amazed by having such an outstanding researcher like Prof. Wu as my external reader.

I must also thank my dear colleagues in my lab, especially Dr. Abedlrahman Al-Khateeb who was more like a mentor to me during the past three years. He was the helping hand that was always there to guide me. A big thank you goes to my colleague

Quang Pham; we spent many days and nights in the lab. I remember entering the room multiple times and seeing Quang sleeping while his R and Java scripts were running on his PC!

Table of Contents

List of Tables

List of Figures

List of Acronyms

mRNA	messenger RNA
NGS	Next Generation Sequencing
TPM	Transcripts Per Kilobase Million per reads
AUC-ROC	Area Under the Receiver Operating Characteristics Curve
AUC	Area Under the Curve
ROC	Receiver Operating Characteristics Curve
TPR	True Positive Rate
FPR	False Positive Rate
LHGDN	Literature Human Gene Derived Network
TNM	Tumour, Node, Metastasis
TMA	Tissue MicroArray
CNN	Convolutional Neural Network
SVM	Support Vector Machine
NCBI	National Center for Biotechnology Information
GEO	Gene Expression Omnibus
MCC	Moffitt Cancer Center
AVAMC	Atlanta Veterans Administration Medical Center
hg19	human genome, version 19
RPKM	Reads Per Kilobase per Million of reads
FPKM	Fragments Per Kilobase per Million of reads
SMOTE	Synthetic Minority Oversampling Technique

NCL	Neighbourhood Cleaning rule
IG	Information Gain
mRMR	minimum Redundancy Maximum Relevance
Rcor3	Rest Corepressor 3
AR	Androgen Receptors
DHT	DiHydrotestosterone
PSA	Prostate Specific Antigen
TRUS	Transrectal Ultrasound image
DRE	Digital Rectal Exam
MRI	Magnetic Resonance Imaging
CRF	Conditional Random Fields
TCGA	The Cancer Genome Atlas
PRAD	Prostate Adenocarcinoma
ENN	Edited Nearest Neighbor
RBF	Radial Basis Function
ER	Endoplasmic Reticulum
PP2A	Protein Phosphatase 2A
OXPHOS	Oxidative Phosphorylation
HIF	Hypoxia-Inducible Factor
MRP	Mitochondrial RNA Processing
APC	Antigen-Presenting Cells
SRSF6	Serine and Arginine Rich Splicing Factor 6

EMT	Epithelial to Mesenchymal Transition
NED	Neuroendocrine Differentiation
CNA	Copy Number Alterations
CTD	Comparative Toxicogenomics Database
RGD	Rat Genome Database
MGD	Mouse Genome Database
GAD	Genetic Association Database
LHGDN	Literature Human Gene Derived Network
CSV	Command Separated Values
NCBI	National Center for Biotechnology Information
GEO	Gene Expression Omnibus
BCB	Conference on Bioinformatics, Computational Biology & Health Informatics

Chapter 1

Introduction

Machine learning provides tools and methods that help work on large data sets to find patterns that are usually hidden. The main idea behind machine learning is that we do not explicitly provide the rules, but examples of the data and the labels associated with it [1]. As such, the underlying algorithms will be able discover the hidden patterns and rules that can be used to predict the class labels for a new, unknown sample [2]. The methods used for of automatic classification and recognition of newly given samples are becoming very important and are used in many fields, including the fields of biology and clinical diagnosis.

Cancer is one of the main causes of death worldwide. Cancer is considered a genetic disorder [3], in which gene mutations and changes cause the cells to malfunction, which affects the cells growth and division. Roughly speaking, genes are “transformed” into proteins that are responsible for most of the work in biological processes and are required

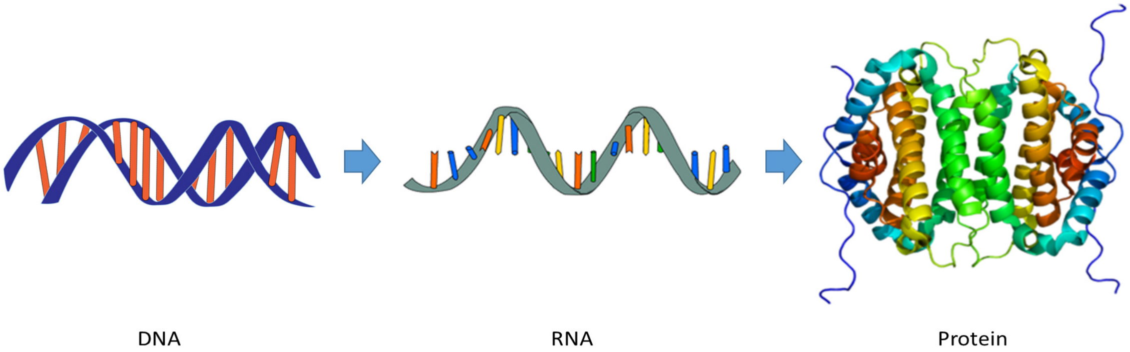

Figure 1.1: The central dogma of molecular biology.

for the structure, function, and regulation of the body's tissues and organs. Parts of the DNA are transcribed into messenger RNA (mRNA) through the process of *transcription* and then they will be translated into proteins in the process of *translation*. The full process is typically called the central dogma and is illustrated in Figure 1.1 Changes to the genes, mRNAs or the proteins may lead to some kind of malfunctions in the cell or tissue. As such, tissue might then go through uncontrolled growth and become cancer.

1.1 Prostate cancer

Prostate cancer is the cancer type with the highest incidence among males; around 1.276 million cases were newly diagnosed worldwide in 2019 [1]. In prostate cancer, the size of the main tumour and the lymphatic involvement are used to assign a metric of tissue organization and disease aggressiveness called the *Gleason score*.

The Gleason score is calculated by adding two numbers: the most common pattern

Table 1.1: Gleason groups as per the latest study from Epstein et al [4].

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

of the tumour 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 tumour. This number is determined by a pathologist, where the highest score means the most aggressive form of cancer [4]. For example a Gleason score of 3+3=6 is the first stage of prostate cancer, while a 5+5=10 is considered the last stage of the disease. 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 categories as depicted in Table 1.1.

Prostate cancer tumor can be located in three different locations, left, right or the middle of the prostate gland. A recent study by Akatsuka et al. [6] concluded that cancer incidence and prognosis varies based on the location within the prostate gland. In Chapter 3 we utilized gene expressions to predict the location of the tumor.

1.2 Next Generation Sequencing

The first successful attempts to sequence DNA started in the early 1970s by academic researchers using laborious methods based on two-dimensional chromatography. Then,

Figure 1.2: Illumnia Nextseq 550 model is one of the most powerful next generation sequencers; image taken from the Illumina website [9]

fluorescence-based sequencing methods were developed with a DNA sequencer [7]. The whole process of DNA sequencing became much easier and faster in the late 2000s [8].

Since then, DNA sequencing technology speeds increased dramatically, which cleared the way to start the process of sequencing the complete DNA of different species of life, including the Human genome. Next generation sequencing (NGS) parallel processing power enabled the sequencing of a massive number of DNA molecules at the same time, whose number can stretch up to the order of millions of molecules.

The high-throughput and the possibility of sequencing multiple samples in the same run enabled researchers to advance in the fields of clinical diagnostics, personalized medicine and genetic diseases, among others. Modern-day Sanger sequencing instruments use capillary-based automated electrophoresis, which typically analyzes 8–96 sequencing reactions simultaneously. NGS systems have been introduced in the past decade, allowing for massively parallel sequencing reactions. These systems are capable of analyzing millions or even billions of sequencing reactions at the same time. The major disadvantage of this technology is that to achieve a significant level of accuracy very short sequencing reads have to be generated.

These include whole genome sequencing, exome sequencing, RNA sequencing, disease panels, lane rentals, and many more. Illumina is one of the market leaders in providing life science tools and integrated systems for large-scale analysis of genetic variation and function. They produce multiple models of next generation sequencers, such as the Illumina Nextseq 550 sequencer model shown in Figure 1.2.

1.3 Gene Expression Data Analysis and Machine Learning Methods

The data produced by NGS are usually large raw data sets, which may include the whole genome DNA or messenger RNA from the tissues that are inspected. These raw data sets consist of millions of short sequence reads, which are used to measure gene expression at the nucleotide resolution level. The first step to obtain information from thes data consists of aligning these reads to a reference genome. There are many tools that can be used to perform this task. One of the most widely-used tools is Tophat2 [10], which aligns the given raw reads into annotated genes or transcripts. STAR [11] is another aligner tool that is well-known for its blazing aligning speed. The next step involves counting these reads. This task can be done using tools like Tophat2 or RSEM [12] to generate the "gene expressions", which are accounted for in terms of TPM (Transcripts Per Kilobase Million). Figure 1.3 shows the pipeline used to obtain the TPM expression levels from the raw reads that the NGS technology produces.

Once we obtain the gene expressions from each sample and the label of the corresponding sample, the next step is a direct implementation of machine learning methods.

Machine learning provides methods to handle data with given labels, which are called supervised learning. We also have unsupervised learning methods which deals with data without labels. Classifiers are methods that can utilize data that includes features and their corresponding labels to build a model that is capable of predicting the labels of new given, unlabelled samples. For example, a sample has gene expressions which are considered features and also has a Gleason score that can be considered the label.

There are many classifiers that can be used, but there is no specific classifier that can solve all the problems efficiently. In Chapter 2, we used multiple classifiers to build a model that predicts the Gleason group of a sample prostate tumour given it's gene expressions, while in Chapter 3, we used different classifiers to predict the location of the tumour.

The data generated from the NGS includes gene and transcript data for each sample. A single sample can contain up to 70,000 transcripts or more than 30,000 genes. Dealing with this huge number of features would make the classifiers struggle with processing all the features; this problem is known as the *curse of dimensionality.* Machine

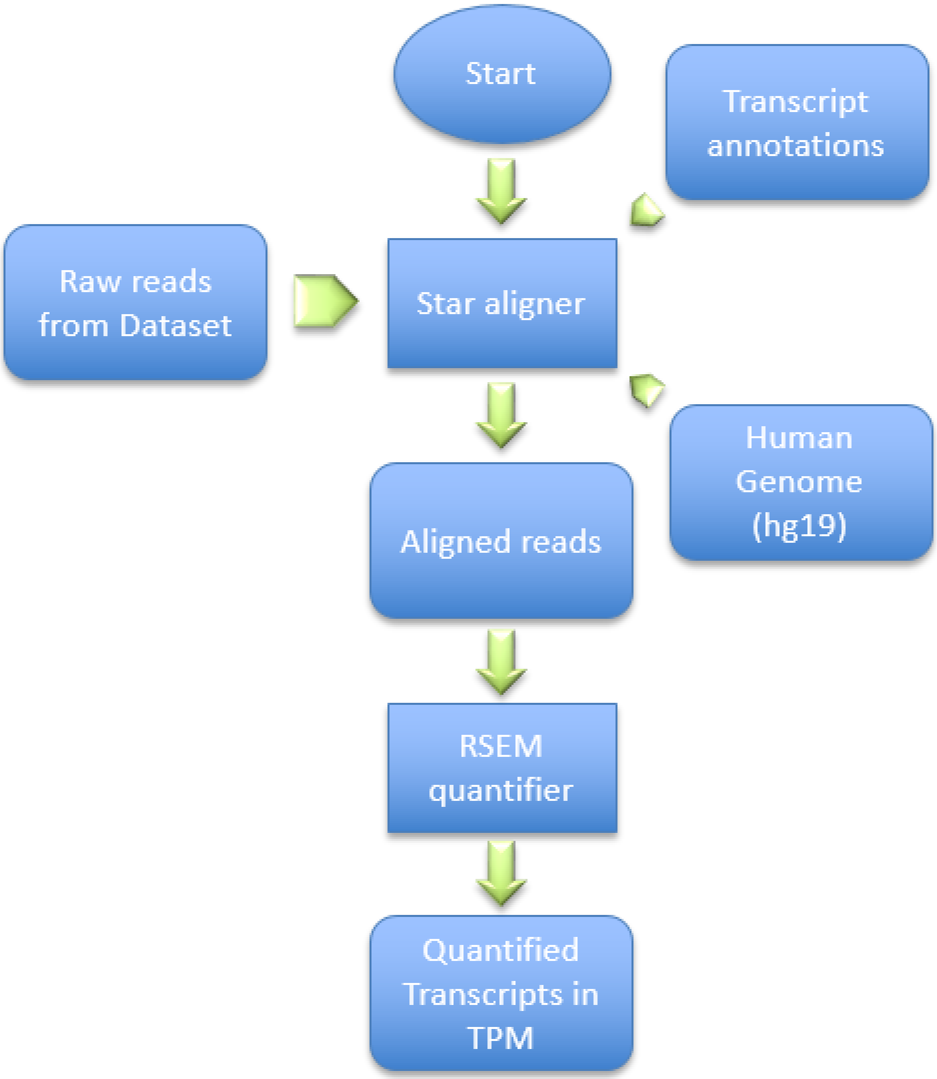

Figure 1.3: Pipeline used to obtain the TPM expression levels from the raw reads that the NGS technology produces.

learning has tools to reduce the number of features using feature selection methods.

1.4 Traditional and Integrative Machine Learning Feature Selection Methods

Traditional feature selection is generally done in two steps. The initial step involves filter-based the feature selection, which aims at giving a score to each feature based on its effect on the predicted target for the main classifier. Each attribute, which in our case, are the genes, is assigned a score that depends on how relevant the feature is to perform the classification task. The second step is a wrapper-based feature selection method, which involves identifying sets of attributes that can be used to categorize certain biological features (the target) [13]. An example of a clinical feature that can be used as a target for classification is the Gleason score in prostate cancer or progression stages in breast cancer. The constructed model can then be assessed using certain performance measures, such as accuracy, specificity, or Area Under the Receiver Operating Characteristics (AUC-ROC) [14].

The final step of traditional machine learning entails a trained individual determining the validity of the results based on the newest literature.

In order to evaluate the model, certain metrics are used, such as accuracy which calculates the ratio of correctly classified samples against the total number of samples [15]. Another two metrics used are sensitivity and specificity, where sensitivity indicates, how well the test predicts one category and specificity measures how well the test predicts the other category. Another important metric is the AUC (Area Under The Curve), where the curve is the ROC (Receiver Operating Characteristics) curve. It is a graph that shows the performance of a classification model by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR), where TPR is the the number of samples classified correctly as a positive class divided by the sum of both the number of samples classified incorrectly as the negative class and number of samples classified correctly as the positive class. FPR is the number of samples classified correctly as a negative class divided by the sum of both the number of samples classified incorrectly as the negative class and the number of samples classified correctly as the positive class. This area indicates the capability of the model to distinguish between different classes. A higher AUC value means a better predictive model.

On the other hand, integrative techniques add domain knowledge from trusted external knowledge databases during feature selection, which might lead to better ability to interpret the data and might give a better predictive outcome. A recent study by Perscheid et al. proposed a framework that utilizes domain knowledge from different databases to generate a list of genes related to the disease of study [16].

The way genes affect certain diseases is an area that is being extensively studied, and numerous discoveries in this regard have already been published. Taking that aspect into account, DisGeNET [17] is a database that can gather knowledge and offer a tool that can be used to find established relations between genes and diseases.

DisGeNET incorporates data from expert-curated sources, GWAS catalogues, animal models and the literature. DisGeNET data are consistently annotated with controlled terms and community-driven ontologies. It also integrates the literature directly using text-mining approaches like The Literature Human Gene Derived Network (LHGDN) [18] and BeFree data, obtained using the BeFree System, which obtains gene-disease associations from MEDLINE abstracts [19] [20].

This database provides several ways to gather its findings, whether it be through the main web portal, a web API, a SQL database or an all in one file. The results obtained are a score that associates a gene to a disease.

There are other databases that are publicly available , like DISEASES [21], Poly-Search2 [22], and DigSee [23]. However, DisGeNET contains more resources than any other database, and has a higher number of citations in the latest publications. In Chapter 4 of this thesis, we propose a knowledge-base integrated approach that enhances conventional methods such as those proposed in Chapters 2 and 3, and it can be used with any machine learning project that deals with cancer data sets.

1.5 Thesis Organization

This thesis is organized in five chapters. The first chapter is an introduction to the relevant fields and the main terms used in the thesis. Chapter 2 discusses how to utilize machine learning techniques to identify Gleason Groups based on mRNA transcripts and gene expressions. This involves a multi-class classification problem that was solved using a hierarchical model:

Chapter 2: Hamzeh, O., Alkhateeb, A., Zheng, J. Z., Kandalam, S., Leung, C., Atikukke, G. & Rueda, L. (2019). A Hierarchical Machine Learning Model to Discover Gleason Grade-Specific Biomarkers in Prostate Cancer. Diagnostics, 9(4), 219.

Chapter 3 covers the implementation of a machine learning approach that uses feature selection methods and classification models to predict the location of the prostate tumours based on gene expression:

Chapter 3: Hamzeh, O., Alkhateeb, A., & Rueda, L. (2018, April). Predicting Tumor Locations in Prostate Cancer Tissue Using Gene Expression. In International Conference on Bioinformatics and Biomedical Engineering (pp. 343-351). Presented at the 6th International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO 2018, Granada, Spain, April 25–27, 2018.

In Chapter 4, we propose an integrative feature selection method that utilizes literature from online databases to integrate knowledge of gene to disease relation to enhance the feature selection methods:

Chapter 4: Hamzeh, O., & Rueda, L. (2019, September). A Gene-disease-based

Machine Learning Approach to Identify Prostate Cancer Biomarkers. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (pp. 633-638). Presented at the Machine Learning Models for Multi-omics Data Integration MODI 2019, a workshop held at the 10th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB), in the , Niagara Falls, New York, September 7-10, 2019.

Finally, Chapter 5 concludes the thesis and highlights the contributions and some of the drawbacks of the implementations covered inside this thesis, and also discusses possible avenues for extension of the proposed approaches and future work.

References

[1] Bishop, C. (2006). Pattern Recognition and Machine Learning. Springer.

[2] Zhang, D., & Zhou, L. (2004). Discovering golden nuggets: data mining in financial application. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34(4), 513-522.

[3] National Cancer Institute, The Genetics of Cancer,https://www.cancer.gov/about-cancer/causes-prevention/genetics [Online; accessed on November 11, 2019].

[4] Ferlay, J.; Colombet, M.; Soerjomataram, I.; Mathers, C.; Parkin, D., Piñeros, M.; Znaor, A.; Bray, F. Estimating the global cancer incidence and mortality in 2018: Globocan sources and methods. Int. J. Cancer 2019, 144, 1941–1953.

[5] Gordetsky, J.; Epstein, J. Grading of Prostatic Adenocarcinoma: Current State and Prognostic Implications. Diagn. Pathol. 2016, 11, 25.

[6] Jun, A., Go, K., Kotaro, O., Masayuki, S., Masato, Y., Yuki, E., Hayato, T., Tatsuro, H., Ichiro, M., Yasutomo, S., Tsutomu, H., and Yukihiro, K. Does tumor location affect prostate cancer prognosis. Journal of Clinical Oncology 2019 37:7_suppl, 45-45.

[7] Olsvik, O., Wahlberg, J., Petterson, B., Uhlén, M., Popovic, T., Wachsmuth, I., Fields, P. (January 1993). Use of automated sequencing of polymerase chain reaction-generated amplicons to identify three types of cholera toxin subunit B in Vibrio cholerae O1 strains. J. Clin. Microbiol. 31 (1): 22–25.

[8] Pettersson, E., Lundeberg, J., Ahmadian, A. (February 2009). Generations of sequencing technologies. Genomics. 93 (2): 105–11

[9] Illumnia Incorporation, https://www.illumina.com/content/dam/illumina-marketing/images/systems/v2/banners/systems/system-banner-nextseq-550.png . [Online; accessed December 20, 2019].

[10] Trapnell, C., Pachter, L., & Salzberg, L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25(9), 1105-1111.

[11] Dobin A., Davis C., Schlesinger F., Drenkow, J., Zaleski C., Jha, S., Batut, P., Chaisson, M., Gingeras, T. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013, 29, 15–21.

[12] Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12(1), 323.

[13] Kotsiantis, S., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging Artificial Intelligence Applications in Computer Engineering, 160, 3-24.

[14] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145-1159.

[15] Flach, P. (2003). The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In Proceedings of the 20th International Conference on Machine Learning (ICML-03) (pp. 194-201).

[16] Perscheid, C., Grasnick, B., & Uflacker, M. (2019). Integrative gene selection on gene expression data: providing biological context to traditional approaches. Journal of Integrative Bioinformatics, 16(1).

[17] Pinero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M. & Furlong, L. I. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database, 2015.

[18] Bundschus, M., Dejori, M., Stetter, M., Tresp, V., & Kriegel, H. P. (2008). Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics, 9(1), 207.

[19] Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., ...& Furlong, L. I. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research, gkw943.

[20] Bravo, À., Piñero, J., Queralt-Rosinach, N., Rautschka, M., & Furlong, L. I. (2015). Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinformatics, 16(1), 55.

[21] Liu, Y., Liang, Y., & Wishart, D. (2015). PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. Nucleic Acids Aesearch, 43(W1), W535-W542.

[22] Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, X., & Jensen, J. (2015). DISEASES: Text mining and data integration of disease–gene associations. Methods, 74, 83-89.

[23] Kim, J., So, S., Lee, H. J., Park, C., Kim, J., & Lee, H. (2013). DigSee: disease gene search engine with evidence sentences (version cancer). Nucleic Acids Research, 41(W1), W510-W517.

Chapter 2

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

2.1 Introduction

Cancer is among the main causes of death worldwide. Among males, prostate cancer is the cancer type with the highest incidence; 1.276 million new cases were diagnosed in 2019 [1]. To date, most cancer studies have concentrated on finding biomarkers that enable differentiating malignant tumours from benign ones. More recent studies, though, have focused on specific clinical aspects of tumours, such as recurrence, progression, survivability, and metastasis, among others.

In the 1950s, Denoix devised a system that categorises solid tumours into different stages [2]. The classification (TNM) of cancer progression is done by utilising (T) the extension and the size of the main tumour, (N) the lymphatic involvement, and (M) the metastasis levels [3]. In prostate cancer, these characteristics are also used to assign a metric of tissue organisation and disease aggressiveness called the Gleason score. That score is calculated by adding two numbers: the most common pattern of the tumour 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 tumour, where the highest score means the most aggressive form of cancer [4]. 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 categories as depicted in Table 1.1. They are 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 [5].

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 have made genomic data analysis widely available. The output of NGS sequencers requires preprocessing algorithms to do things such as align the reads to a reference human genome and assemble 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 [6]. TopHat2 is a widely used, open-source tool that incorporates Bowtie sequence alignment to align reads [7]. STAR is the fastest RNA-Seq sequence alignment algorithm to date, although it requires huge computational resources to perform efficiently [8]. Based on the need for understanding the biological basis of the visual Gleason microscopic assessment, Roberto et al. conducted a gene expression profiling on two groups of Gleason scores 6 and 7, or high, using a metabolic gene panel. The panel consists of many gene members of the JAK/STAT pathway [9]; this pathway is involved in processes such as immunity, cell division, cell death and tumour formation. In this study, we analysed the transcription level of different Gleason scores to find genes that can identify one specific Gleason group from the others.

In addition, machine learning applications in genomic analysis have become a solid approach to analysing 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 tumour will progress to the next stage [10]. Arvaniti et al. proposed a deep learning approach to predict Gleason scores [11]. 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 decent performance measurements (average accuracy 85.72%, and recall 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 [12]. 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 was to identify genes that can be used to differentiate specific Gleason groups. This work is an extension of our previously proposed prediction model, which was based on analysing the RNA-Seq data from patients with different Gleason scores [13]. The method can track transcripts associated with specific genes, in addition to their corresponding expression values. The results of the initial trial show great potential to build a simple system to diagnose Gleason scores based on NGS data.

2.2 Materials and Methods

The primary data set used in this study was retrieved from the National Center for Biotechnology Information (NCBI) and is referenced with Gene Expression Omnibus (GEO) number GSE54460 [40]. This RNAseq prostatectomy data set was generated from 106 prostate cancer tissue samples and validated on an independent data set with 140 patients. Several health sciences centres provided data samples as well. The Moffitt Cancer Center (MCC) contributed ten samples from patients who underwent radical prostatectomies 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

Table 2.1: Numbers of samples in different Gleason groups.

Gleason Score	**Number of Samples**
6	10
3 + 4 = 7	55
4 + 3 = 7	24
8	10
9	4

samples from patients who underwent radical prostatectomy between the years 1990 and 2000. Table 2.1 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 data set was generated by using the Illumina HiSeq 2000 NGS on paired-end sequences of length 51 bp each. The pre-processing pipeline starts by obtaining the RNA-Seq samples and pre-processing them using SRAtools [41], as depicted in Figure 1.3. The process continues by incorporating the STAR aligner [8] 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 [42]. 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 artefacts, and pre-processing must be done before the downstream analysis. These artefacts include duplication and bias reads [43], 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 used to measure the read quantification rather than reads per kilobase per million of reads (RPKM) [44]. Additionally, the reason for choosing TPM instead of fragments per kilobase per million (FPKM) [45] is that TPM normalises the reads to the length of the gene first, which makes it easier to compare the quantified reads among different samples.

2.2.1 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 data set. After applying multiple oversampling and under-sampling methods, the best option was found to be the synthetic minority oversampling technique (SMOTE) [46] for oversampling the minority class, while the neighbourhood cleaning rule (NCL) [47] 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 neighbours. SMOTE, instead, introduces a new way of creating new samples, by utilising 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 calculated by measuring the Euclidean distance between the two samples and multiplying that value by a random number between 0 and 1. Figure 2.1 shows a hypothetical example of the mechanism followed by SMOTE, by adding new synthetic samples randomly along the line that connects each of two original samples in a minority class. The blue points represent the original samples, while the amber

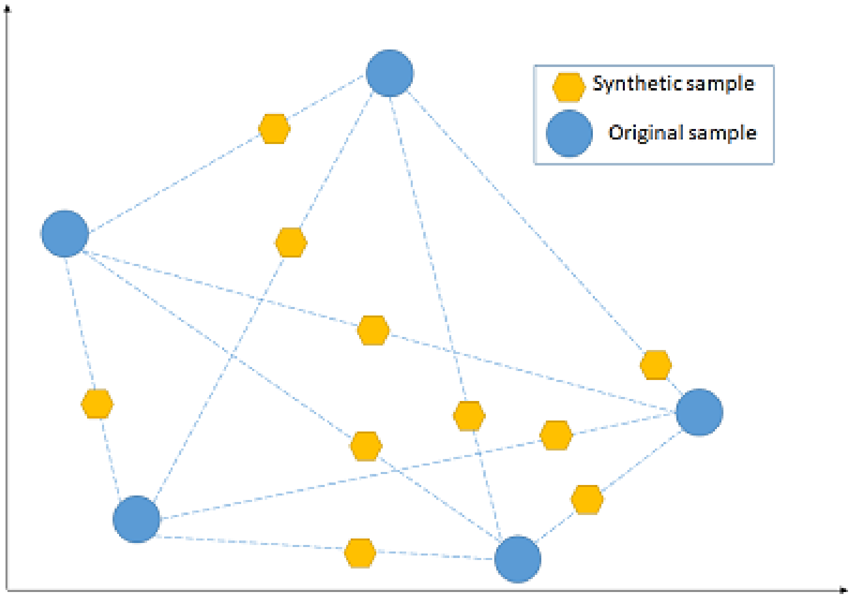

Figure 2.1: Hypothetical example that shows how the synthetic minority oversampling technique (SMOTE) works.

points represent the synthetically generated samples.

2.2.2 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 [48]. Thus, feature selection was applied to reduce the dimensionality of the problem. The first step of the feature selection is to filter the transcripts based on their information gain values by selecting the ones with the highest scores. The filter method, which is called attribute evaluator, is the procedure by which each attribute (transcript) in the data set is assessed with regard 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) was used as an attribute evaluator to rank each attribute vector [49]. The IG of attribute vector X concerning class vector A is defined as follows:

$$IG(A, X) = H(A) - H(A|X), \tag{2.1}$$

Where, $H(A)$ 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 scores, a wrapper-based feature selection algorithm 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 minimise the redundancy while increasing the correlation to the class vector [50]. The wrapper method adds up the features that minimise redundancy (W), and maximize the relevance (V), with the best possible accuracy of an SVM classifier that uses a linear kernel, as per the following equations:

$$W = \frac{1}{|S|^2} \sum_{i,j \in S} I(i, j)\,, \tag{2.2}$$

and

$$V = \frac{1}{|S|} \sum_{i \in S} I(h, i)\,, \tag{2.3}$$

where S is the set of features, $I(i, j)$ is the mutual information between features (i, j), and

h is the class, in our case, the five Gleason groups.

2.2.3 Classification

The problem dealt with is multi-class classification, which was 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 data sets from the actual data. For each data set, we set one of the classes to form the positive class, while the rest of the classes were 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 2.4). 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 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; The accuracy, sensitivity, and specificity are used in this study. 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 minimise the error propagation by choosing the class with the highest accuracy at each internal node.

2.2.4 Identifying Transcripts within Different Gleason Scores

We used the Scitkit-learn [51] library to apply different classification algorithms to the final transcripts selected. 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 were 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 [52]. 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 transcripts selected in the previous step [53]. The advantage of SVM is its exceptional generalisation power, especially in high-dimensional data with a small number of samples. Figure 2.2 shows the pipeline followed in this study.

2.3 Results

The first data set 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 2.3. The data set 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 was calculated from the total number of correctly classified samples divided by the total number of samples.

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 2.2–2.5 are the most relevant for identifying prostate cancer; the Gleason scores using the

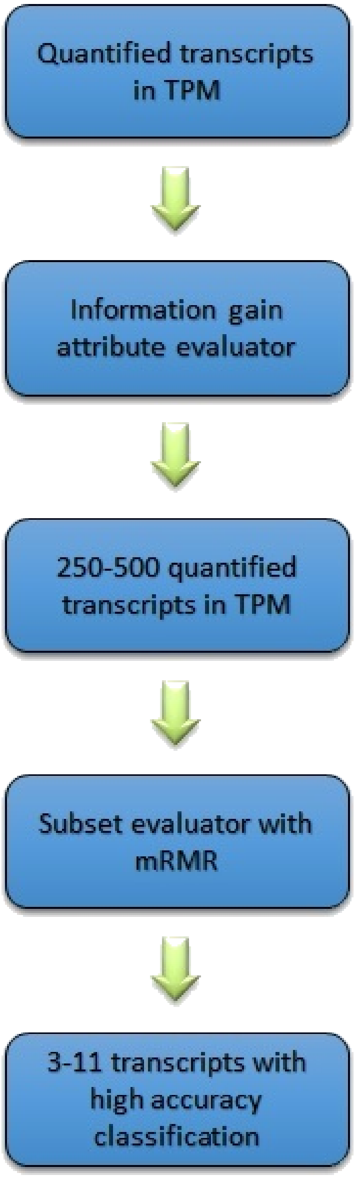
Quantified transcripts in TPM
Information gain attribute evaluator
250-500 quantified transcripts in TPM
Subset evaluator with mRMR
3-11 transcripts with high accuracy classification

Figure 2.2: Machine learning pipeline used in the proposed method.

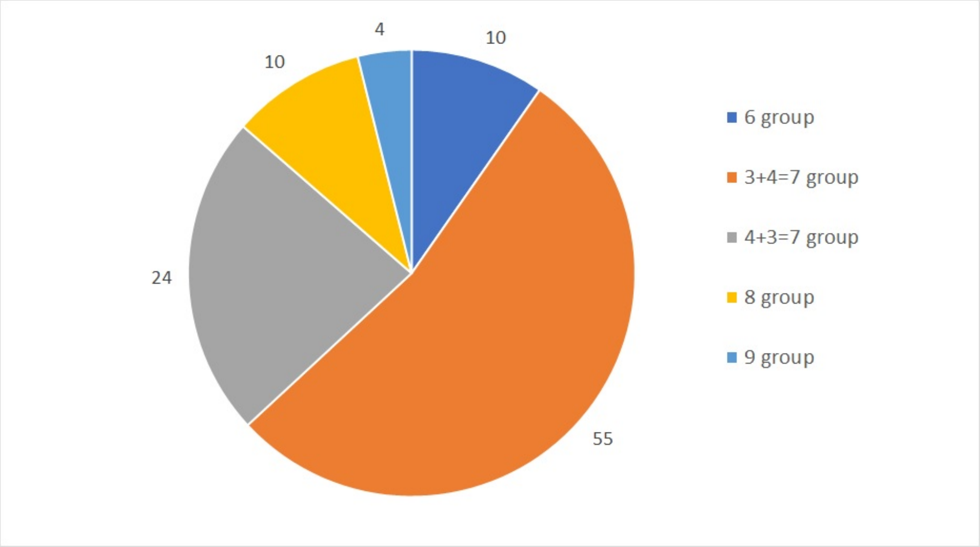

Figure 2.3: Gleason groups and their distributions.

hierarchical method are illustrated in Figure 2.4. Different classification methods for each stage within the hierarchy are shown in Table 2.6.

The first node of the hierarchy yields 94% accuracy in identifying Gleason score 3 + 4 = 7 compared to the other scores. The samples are then passed through node 2, in which Gleason score 4 + 3 = 7 was identified from the rest with a prediction accuracy of 98%. The other samples were then passed through node 3, where Gleason score 6 was identified with the accuracy of 100%. The remaining samples were finally processed in the last node, where the Gleason score 8 was identified from the Gleason score 9 with the accuracy of 100%. Due to the similarity in the aggressiveness of the tumour and the low number of samples, all the other Gleason scores were merged in the last node.

Figure 2.5 shows the classifiers that have been utilised to identify the set of transcripts that differentiate specific Gleason groups against the rest. The classifiers are rep-

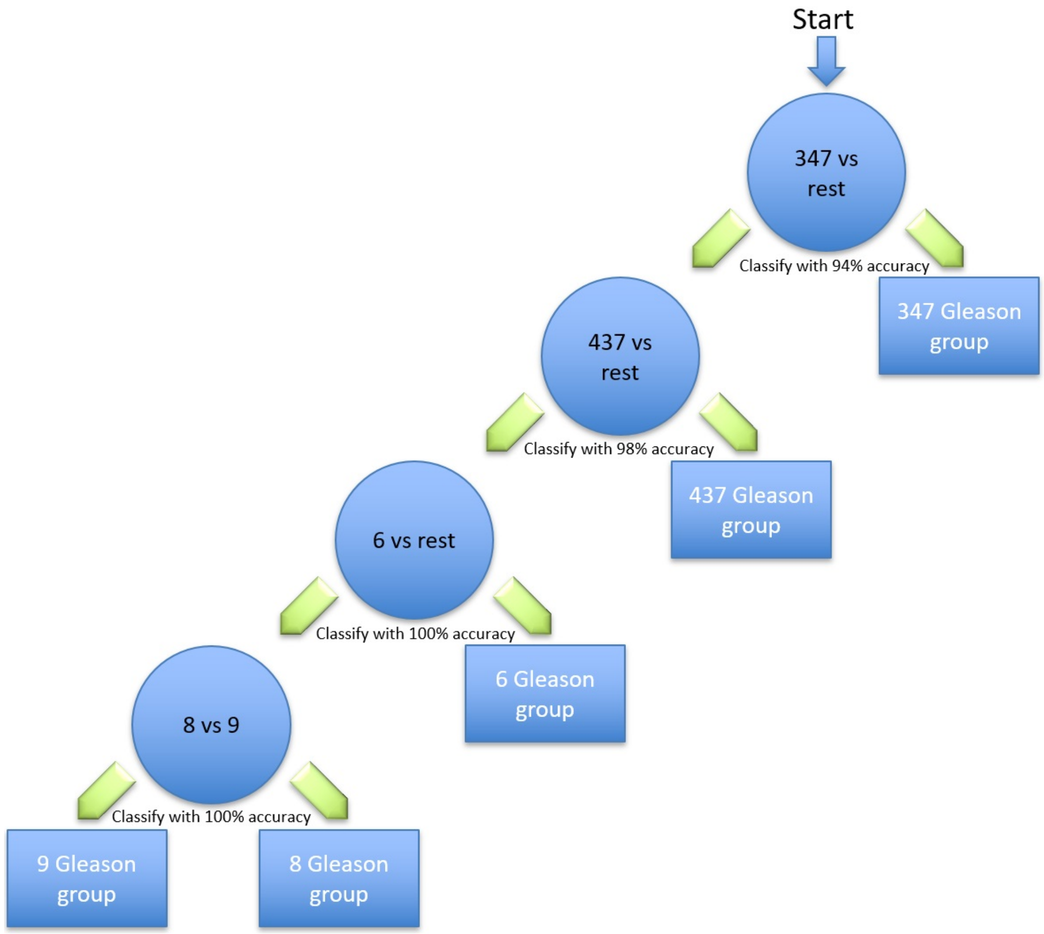

Figure 2.4: Hierarchical tree of classifications of Gleason groups against the rest, along with the corresponding classification accuracies.

Table 2.2: 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 2.3: 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 2.4: 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 2.5: 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

Table 2.6: Classification performance for each step in the hierarchy.

Gleason Group	**Accuracy**	**Sensitivity**	**Specificity**	**F-Measure**	**MCC**	**ROC**
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

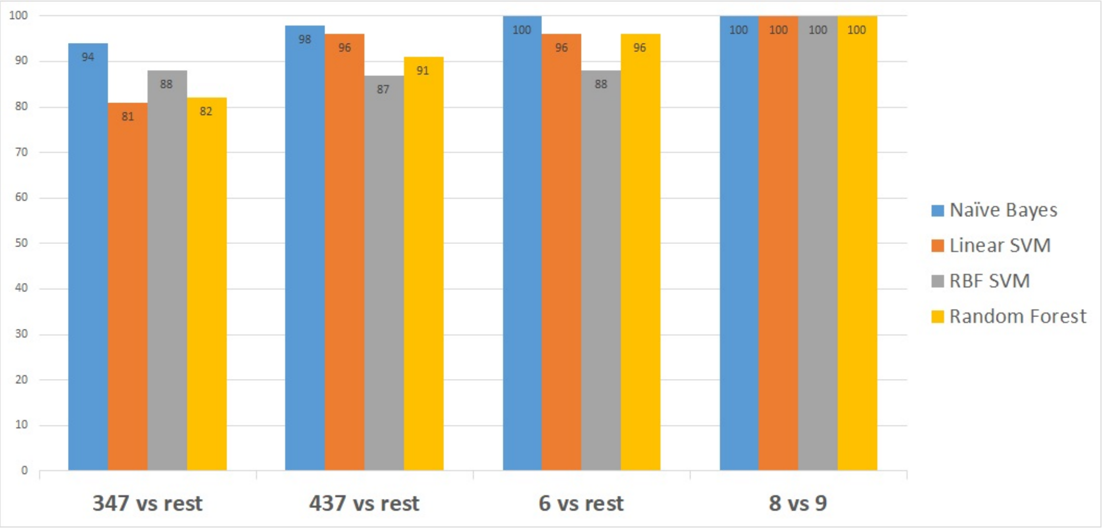

Figure 2.5: Accuracy obtained by each classifier for classifying one versus the rest for all five Gleason groups.

resented on the x-axis, while the classification performance measurements are represented on the y-axis.

Naïve Bayes outperformed the other classifiers, as it distinguished the first Gleason score node from the rest with the accuracy of 94%, the second node with a higher accuracy of 98%, and the last two Gleason score nodes with the accuracy of 100% accuracy, as shown in Figure 2.5.

2.4 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*,

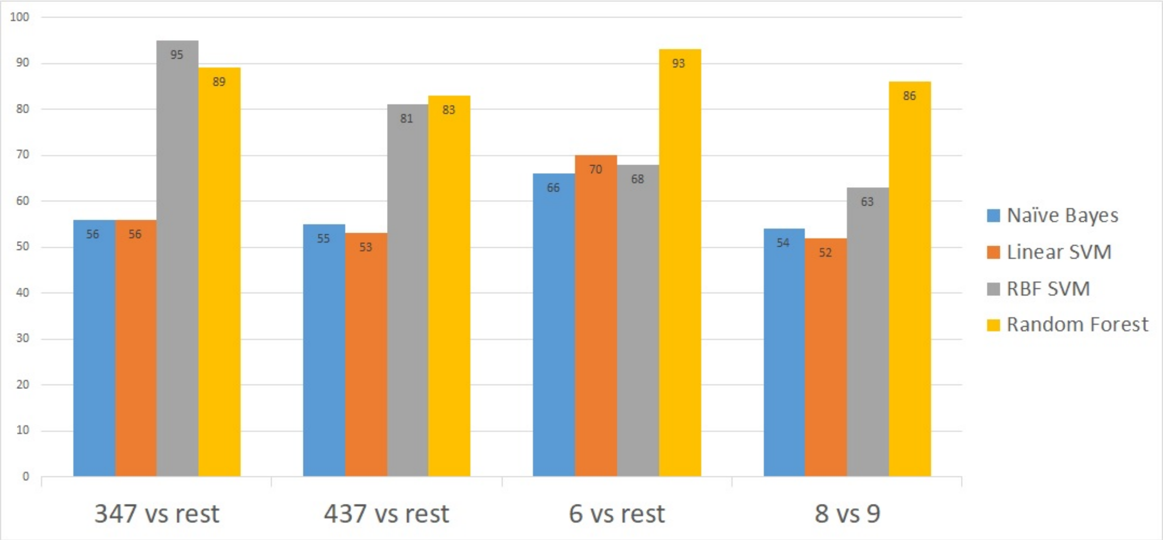

Figure 2.6: Classification accuracies obtained after applying the model on the second data set.

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 [16]. In familial prostate cancer, however, a high frequency variant of UBE2V2 was identified and found to affect DNA repair and androgen signaling [17]. In our model study, a different quantification of the UBE2V2 transcript was able to predict Gleason score 6 (group 1) in the first data set. Differential expression of UBE2V2 has also been associated with poor prognosis in breast cancer [18].

Our study also reveals that the differential expression of *GPR137* expression and EPB41L1 is associated with tumours of Gleason scores 3 + 4 = 7 and 8, respectively. Earlier studies show that proteins encoded by *EPB41L1* are associated with the proper organisation of the cell cytoskeleton, and that *EPB41L1* plays an important role in the negative regulation of cell metastasis, migration, and invasion. Expression of *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 disorganised cell and tissue structures associated with higher grade prostate cancer [19], and thus link 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 [20]. *EPB41L1* was also shown to be differentially expressed in gastric cancer [21]. On the other hand, *GPR137* expression has been shown to be upregulated in prostate cancer tissues compared with paracancerous tissues. Moreover, knockdown of *GPR137* resulted in decreased cell proliferation and colony formation in PC-3 and DU145 prostate cancer cell lines, and was associated with cell cycle arrest at G0/G1 phase. *GPR137* suppression also decreases the migration and invasive abilities of PC-3 cells, suggesting that *GPR137* plays a role in prostate cancer progression and metastasis [22].

Differential expression of *PIAS3* and Rest Corepressor 3 (Rcor3) were both associated with tumours 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 [23], a characteristic of prostate tumours with Gleason score 4 + 3 = 7 [4]. On the other hand, differential *PIAS3* expression has been observed in a variety of human cancers, including lung, breast, prostate, colorectal, and brain [24]. *PIAS3* is expressed in prostate cancer cells, and its expression is induced in response to androgens [26, 25]. 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) [24]. *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 consisting of four members: *PIAS1*, *PIAS2*, *PIAS3*, and *PIAS4* [27]. *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 [28, 29]. 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 [30, 31]. 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 2.7 depicts the protein–protein interaction among genes with $4 + 3 = 7$ and 6 scores, as extracted from ProteomicsDB (`https://www.proteomicsdb.org/proteomicsdb/#human/proteinDetails/86810/interactions`) based on experimental and epidemiological 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 [32]. Stat5a/b is constantly active in human prostate cancer [33], associated with high histological grades [34], and a predictor of early prostate cancer recurrence [35]. Transcription factor Stat5a/b

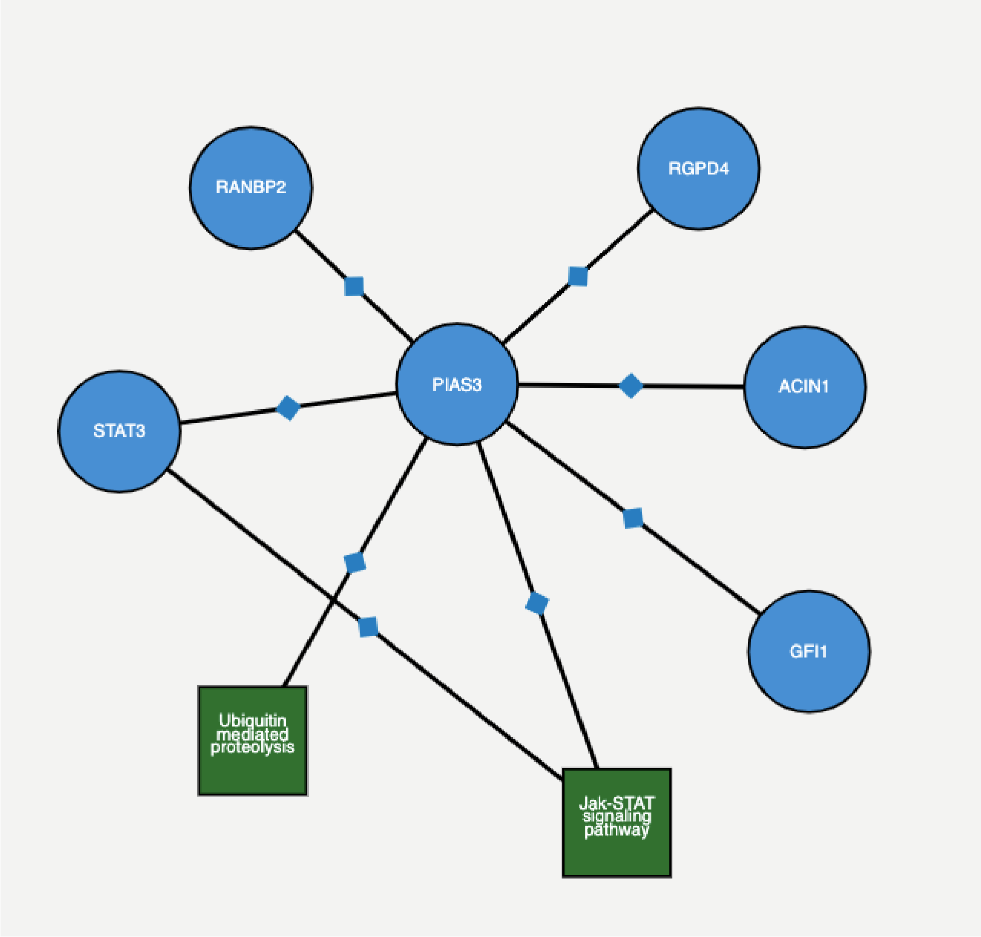

Figure 2.7: An interactive figure taken from proteomics database STRING. It shows neighbouring protein binding and pathway interactions for a given gene using STRING and KEGG pathway analysis. Here, the gene of interest is *PIAS3*, an identified possible biomarker in the $4 + 3 = 7$ score. The figure shows the interaction between other proteins and pathways associated with it.

has been shown to regulate the viability and growth of human prostate cancer cells [36, 37]. Moreover, in vitro inhibition of Stat5a/b induces apoptosis in human prostate cancer cells [33, 38]. In vivo, Stat5a/b inhibition blocks prostate cancer subcutaneous and orthotopic xenograft tumour growth in nude mice [38]. 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 [39]. 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*-mediated 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 applied a novel machine learning model to identify differentially expressed, prostate cancer stage-specific transcripts. Although the application of this model to other related data sets 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.

2.5 Conclusions and Future Directions

Identifying novel biomarkers that are clinically associated with specific Gleason groups in prostate cancer is vital for the diagnosis and treatment of the disease. Utilising NGS data and machine learning techniques, a supervised learning method was 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 correlated to prostate cancer. To validate the model, we also tested it on a gene expression data set, showing that the resulting genes are related to prostate cancer progression.

The work presented in this chapter opens the way for future directions of research. One of these is to apply and adjust the same method to other cancer types. Another possible avenue would be to consider analysing 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 could be investigated, which may provide a comprehensive analysis of the progression, diagnosis, and treatment of the disease.

References

[1] Ferlay, J.; Colombet, M.; Soerjomataram, I.; Mathers, C.; Parkin, D., Piñeros, M.;

Znaor, A.; Bray, F. Estimating the global cancer incidence and mortality in 2018: Globocan sources and methods. *Int. J. Cancer* **2019**, 144, 1941–1953.

[2] Gospodarowicz, M.; Benedet, L.; Hutter, R.V.; Fleming, I.; Henson, E.; Sobin, H. History and international developments in cancer staging. Cancer Prev. Control CPC Prev. Controle en Cancerol. PCC **1998**, 2, 262–268.

[3] Edge, S.; Compton, C. The American Joint committee on cancer: The 7th edition of the AJCC cancer staging manual and the future of TNM. Ann. Surg. Oncol. **2010**, 17, 1471–1474.

[4] Gordetsky, J.; Epstein, J. Grading of Prostatic Adenocarcinoma: Current State and Prognostic Implications. Diagn. Pathol. **2016**, 11, 25.

[5] Epstein, I.; Zelefsky, J.; Sjoberg, D.; Nelson, B.; Egevad, L.; Magi-Galluzzi, C.; Vickers, J.; Parwani, V.; Reuter, E.; Fine, W.; et al. A contemporary prostate cancer grading system: A validated alternative to the Gleason score. Eur. Urol. **2016**, 69, 428–435.

[6] Altschul, S.; Gish, W.; Miller, W.; Myers, E.; Lipman, D. Basic local alignment search tool. J. Mol. Biol. **1990**, 215, 403–410.

[7] Trapnell, C.; Pachter, L.; Salzberg, S. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics **2009**, 25, 1105–1111.

[8] Dobin, A.; Davis, C.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T. STAR: ultrafast universal RNA-seq aligner. Bioinformatics **2013**, 29, 15–21.

[9] Roberto, D.; Selvarajah, S.; Park, C.; Berman, D.; Venkateswaran, V. Functional validation of metabolic genes that distinguish Gleason 3 from Gleason 4 prostate cancer foci. Prostate **2019**, 79, 1777–1788.

[10] Alkhateeb, A.; Rezaeian, I.; Singireddy, S.; Cavallo-Medved, D.; Porter, L.; Rueda, L. newblock Transcriptomics signature from next-generation sequencing data reveals new transcriptomic biomarkers related to prostate cancer. *Cancer Inform.* **2019**, 18, 1176935119835522.

[11] Arvaniti, A.; Fricker, K.; Moret, M.; Rupp, N.; Hermanns, T.; Fankhauser, C.; Wey, N.; Wild, P.; Rueschoff, J.; Claassen, M. Automated gleason grading of prostate cancer tissue microarrays via deep learning. BioRxiv **2018**, 280024.

[12] Citak-Er, F.; Vural, M.; Acar, O.; Esen, T.; Onay, A.; Ozturk-Isik, E. Final gleason score prediction using discriminant analysis and support vector machine based on preoperative multiparametric mr imaging of prostate cancer at 3T. BioMed Res. Int. **2014**, 2014, 690787.

[13] Hamzeh, O.; Alkhateeb, A.; Rezaeian, I.; Karkar, A.; Rueda, L. Finding transcripts associated with prostate cancer gleason stages using next generation sequencing and machine learning techniques. In Proceedings of the International Conference on Bioinformatics and Biomedical Engineering, 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 337–348.

[14] Prostate Adenocarcinoma TCGA-PRAD Data set. 2019. Available online: `https://portal.gdc.cancer.gov/projects/TCGA-PRAD` (accessed on November 29, 2019).

[15] National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov (accessed on July 23, 2019).

[16] Zhao, Y.; Long, M.J.; Wang, Y.; Zhang, S.; Aye, Y. UBE2v2 is a rosetta stone bridging redox and ubiquitin codes, coordinating dna damage responses. ACS Cent. Sci. **2018**, 4, 246–259.

[17] Nicolas, E.; Arora, S.; Zhou, Y.; Serebriiskii, G.; Andrake, D.; Handorf, D.; Bodian, L.; Vockley, G.; Dunbrack, L.; Ross, A.; et al. Systematic evaluation of underlying defects in dna repair as an approach to case-only assessment of familial prostate cancer. Oncotarget **2015**, *6*, 39614.

[18] Santarpia, L.; Iwamoto, T.; Di Leo, A.; Hayashi, N.; Bottai, G.; Stampfer, M.; André, F.; Turner, C.; Symmans, F.; Hortobágyi, N.; et al. DNA repair gene patterns as prognostic and predictive factors in molecular breast cancer subtypes. Oncologist **2013**, *18*, 1063–1073.

[19] Schulz, W.; Ingenwerth, M.; Djuidje, C.; Hader, C.; Rahnenführer, J.; Engers, R. Changes in cortical cytoskeletal and extracellular matrix gene expression in prostate cancer are related to oncogenic erg deregulation. BMC Cancer **2010**, *10*, 505.

[20] Ji, Z.; Shi, X.; Liu, X.; Shi, Y.; Zhou, Q.; Liu, X.; Li, L.; Ji, X.; Gao, Y.; Qi, Y.; et al. The membrane-cytoskeletal protein 4.1 n is involved in the process of cell adhesion, migration and invasion of breast cancer cells. Exp. Ther. Med. **2012**, *4*, 736–740.

[21] Seabra, A.; Araújo, T.; Mello, F.; Alcântara, D.; De Barros, D.; De Assumpção, P.; Montenegro, R.; Guimarães, A.; Demachki, S.; Burbano, R. High-density array

comparative genomic hybridization detects novel copy number alterations in gastric adenocarcinoma. Anticancer Res. **2014**, *34*, 6405–6415.

[22] Ren, J.; Pan, X.; Li, L.; Huang, Y.; Huang, H.; Gao, Y.; Xu, H.; Qu, F.; Chen, L.; Wang, L.; et al. Knockdown of gpr137, g protein-coupled receptor 137, inhibits the proliferation and migration of human prostate cancer cells. Chem. Biol. Drug Des. **2016**, *87*, 704–713.

[23] Upadhyay, G.; Chowdhury, A.H.; Vaidyanathan, B.; Kim, D.; Saleque, S. Antagonistic actions of Rcor proteins regulate LSD1 activity and cellular differentiation. Proc. Natl. Acad. Sci. USA **2014**, *111*, 8071–8076.

[24] Wang, L.; Banerjee, S. Differential pias3 expression in human malignancy. Oncol. Rep. **2004**, *11*, 1319–1324.

[25] Vassikis, J.; Do, A.; Wen, S.; Wang, X.; Cho-Vega, H.; Brisbay, S.; Lopez, R.; Logothetis, J.; Troncoso, P.; Papandreou, N.; et al. Clinical and biomarker correlates of androgen-independent, locally aggressive prostate cancer with limited metastatic potential. Clin. Cancer Res. **2004**, *10*, 6770–6778.

[26] Gross, M.; Liu, B.; Tan, J.; French, F.; Carey, M.; Shuai, K. Distinct effects of PIAS proteins on androgen-mediated gene activation in prostate cancer cells. Oncogene **2001**, *20*, 3880.

[27] Ueki, N.; Seki, N.; Yano, K.; Saito, T.; Masuho, Y.; Muramatsu, M. Isolation and chromosomal assignment of a human gene encoding protein inhibitor of activated STAT3 (PIAS3). J. Hum. Genet. **1999**, *44*, 193–196.

[28] Schmidt, D.; Müller, S. PIAS/SUMO: new partners in transcriptional regulation Cell. Mol. Life Sci. **2003**, *60*, 2561–2574.

[29] Shuai, K. Regulation of cytokine signaling pathways by pias proteins. Cell Res. **2006**, *16*, 196.

[30] Rawlings, S.; Rosler, M.; Harrison, A. The JAK/Stat signaling pathway. J. Cell Sci. **2004**, *117*, 1281–1283.

[31] Tam, L.; McGlynn, L.M.; Traynor, P.; Mukherjee, R.; Bartlett, M.; Edwards, J. Expression levels of the JAK/STAT pathway in the transition from hormone-sensitive to hormone-refractory prostate cancer. Br. J. Cancer **2007**, *97*, 378.

[32] Rycyzyn, M.A.; Clevenger, C.V. The intranuclear prolactin/cyclophilin B complex as a transcriptional inducer. Proc. Natl. Acad. Sci. USA **2002**, *99*, 6790–6795.

[33] Ahonen, M.; Poukkula, M.; Baker, H.; Kashiwagi, M.; Nagase, H.; Eriksson, E.; Kähäri, M. Tissue inhibitor of metalloproteinases-3 induces apoptosis in melanoma cells by stabilization of death receptors. Oncogene **2003**, *22*, 2121.

[34] Li, H.; Ahonen, T.J.; Alanen, K.; Xie, J.; LeBaron, J.; Pretlow, G.; Ealley, L.; Zhang, Y.; Nurmi, M.; Singh, B.; et al. Activation of signal transducer and activator of transcription 5 in human prostate cancer is associated with high histological grade. Cancer Res. **2004**, *64*, 4774–4782.

[35] Li, H.; Zhang, Y.; Glass, A.; Zellweger, T.; Gehan, E.; Bubendorf, L.; Gelmann, P.; Nevalainen, T. Activation of signal transducer and activator of transcription-5 in prostate cancer predicts early recurrence. Clin. Cancer Res. **2005**, *11*, 5863–5868.

[36] Liao, C.; Lo, H. Deleted in liver cancer-1 (DLC-1): A tumor suppressor not just for liver. Int. J. Biochem. Cell Biol. **2008**, *40*, 843–847.

[37] Tan, H.; Nevalainen, T. Signal transducer and activator of transcription 5a/b in prostate and breast cancers. Endocr. -Related Cancer 2008, *15*, 367–390.

[38] Dagvadorj, A.; Kirken, A.; Leiby, B.; Karras, J.; Nevalainen, T. Transcription Factor Signal Transducer and Activator of Transcription 5 Promotes Growth of Human Prostate Cancer Cells In vivo. Clin. Cancer Res. **2008**, *14*, 1317–1324.

[39] Dagvadorj, A.; Tan, H.; Liao, Z.; Xie, J.; Nurmi, M.; Alanen, K.; Rui, H.; Mirtti, T.; Nevalainen, T. N-terminal Truncation of STAT5a/b Circumvents PIAS3-mediated Transcriptional Inhibition of STAT5 in Prostate Cancer cells. Int. J. Biochem. Cell Biol. **2010**, *42*, 2037–2046.

[40] Long, Q.; Xu, J.; Osunkoya, O.; Sannigrahi, S.; Johnson, A.; Zhou, W.; Gillespie, T.; Park, Y.; Nam, K.; Sugar, L.; et al. Global Transcriptome Analysis of Formalin-Fixed Prostate Cancer Specimens Identifies Biomarkers of Disease Recurrence. Cancer Res. **2014**, *74*, 3228–3237.

[41] Leinonen, R.; Sugawara, H.; Shumway, M.; International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. **2010**, *39* (Suppl. 1), D19–D21.

[42] Li, B.; Dewey, C. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinform. **2011**, *12*, 1.

[43] Trapnell, C.; Hendrickson, D.; Sauvageau, M.; Goff, L.; Rinn, J.; Pachter, L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. **2013**, *31*, 46–53.

[44] Mortazavi, A.; Williams, B.; McCue, K.; Schaeffer, L.; Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods **2008**, *5*, 621–628.

[45] Trapnell, C.; Williams, A.; Pertea, G.; Mortazavi, A.; Kwan, G.; Van Baren, J.; Salzberg, L.; Wold, J.; Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **2010**, *28*, 511–515. doi:10.1038/nbt.1621.

[46] Chawla, V.; Bowyer, W.; Hall, O.; Kegelmeyer, P. SMOTE: synthetic minority oversampling technique. J. Artif. Intell. Res. **2002**, *16*, 321–357.

[47] Laurikkala, J. Improving Identification of Difficult Small Classes by Balancing Class Distribution; Tech. Rep. A-2001-2; University of Tampere: Tampere, Finland, 2001.

[48] Trunk, V. A problem of dimensionality: A simple example. IEEE Trans. Pattern Anal. Mach. Intell. **1979**, *1*, 306–307.

[49] Novakovic, J. Using information gain attribute evaluation to classify sonar targets. In Proceedings of the 17th Telecommunications forum TELFOR, Serbia, Belgrade, 24–26 November 2009; pp. 24–26.

[50] Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **2005**, *27*, 1226–1238.

[51] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. JMLR **2011**, *12*, 2825–2830.

[52] Domingos, P.; Pazzani, M. On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn. **1997**, *29*, 103–130.

[53] Cortes, C.; Vapnik, V. Support-vector networks. Mach. Learn. **1995**, *20*, 273–297.

Chapter 3

Prediction of Tumor Location in Prostate Cancer Tissue Using a Machine Learning System on Gene Expression Data

3.1 Introduction

Cancer is among the leading causes of death worldwide. In 2013, there were 8.2 million deaths, and 14.9 million cases of cancer incidence [1]. As with all cancer diseases, investigating prostate cancer at the molecular level reveals transcriptional and regulatory mechanisms of the tumour biology. Traditionally, prostate cancer studies centered primarily on finding biomarkers for differentiation between benign and cancerous tumors. Recently, studies have considered some other aspects of the tumours including progression, metastasis, location, and recurrence, among others.

Traditional methods for detecting prostate cancer such as prostate specific antigen (PSA) blood test, transrectal ultrasound image (TRUS) guided biopsy, and digital rectal exam (DRE) do not measure up to the medical standards. PSA blood test statistical results shows a specificity of 61% and a low sensitivity of 34.9%, while TRUS-guided biopsy and DRE are invasive [2].

In addition, multiparametric magnetic resonance imaging (MRI) of the prostate is a functional form of imaging used to augment standard T1- and T2-weighted imaging. Multiparametric MRI may miss up to 12% of cancer cases [3]. In addition to the need for reducing the number of biopsies come most of the time with pain, fever, bleeding, infection, transient urinary difficulties, or other complications that require hospitalization [4]. Finding gene biomarkers of prostate cancer location and analyzing their proteomics can help clinically understand the development of the disease and improve treatment efficiency.

Machine learning approaches, on the other hand, have been successfully applied on prostate cancer data to identify gene biomarkers of the disease [5, 6]. Using next generation sequencing and the power of machine learning, Singireddy et al. devised a support vector machine (SVM) classifier to identify biomarker genes associated with prostate cancer progression. The biomarkers were able to discriminate consecutive prostate cancer stages with high performance [5]. Earlier, Hamzeh et al. proposed a method for finding groups of transcripts that are differentially expressed among the different Gleason stages

[7]. The identified transcripts can be used to predict the actual Gleason score for new samples, and these transcripts belong to genes that are well known to play important roles in prostate and other types of cancer. Yu et al. demonstrated that their method is efficient for predicting prostate cancer aggressiveness based on gene expression patterns [8].

Similarly, machine learning approaches have been used for cancer localization prediction [10, 9]. Artan et al. proposed a prediction model based on a cost-sensitive SVM. The model is used to analyze a large data set of multispectral magnatic resonance imaging (MRI). This method improves the cost-sensitive SVM using a segmentation method by combining conditional random fields (CRF) with a cost-sensitive framework. Incorporating spatial information leads to better localization accuracy [9]. As stated earlier, prediction by imaging is still inaccurate, not specific and hence needs more improvement. In an attempt to find different gene expression levels between two lists, the first contains the expression levels of colon tumor cells, while the latter for rectal tumor cells, Sanz-Pamplona et al. applied agglomerative hierarchical clustering to display the classification ability between both lists. Both lists have very similar gene expression levels except for several HOX genes which are found to be associated with tumor location [10].

In this work, we are extending our previous method for classifying different laterality prostate samples which are left unary, right unary, or bilateral [11]. The results of this multi-class model are set of genes that can determine a specific class from the others. The literature shows that these genes are related to prostate cancer, which may lead to be a potential biomarkers for prostate cancer laterality.

3.2 Materials and Methods

RNA-sequencing data from The Cancer Genome Atlas (TCGA) Prostate Adenocarcinoma (PRAD) was used. This data set consists of 450 samples for different patients with different cancer locations. There are three primary locations that the tumor might be located within the prostate: left, right and bilateral. Figure 3.1 shows the actual possible locations, while Table 3.1 describes the number of samples in each location.

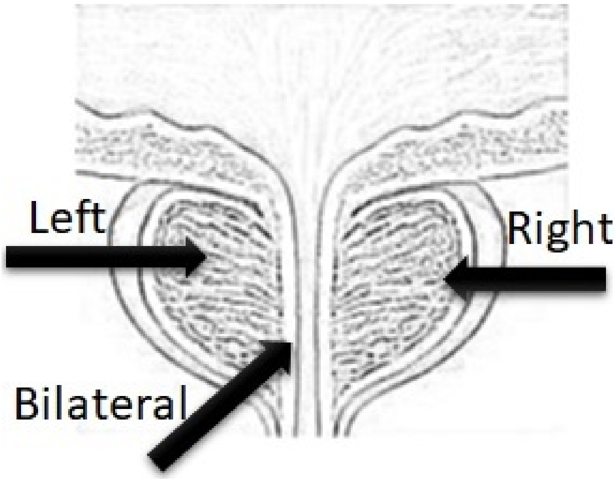

Figure 3.1: Possible locations of the tumor in prostate cancer.

Table 3.1: Number of samples in each prostate cancer tumor location.

Left	**Bilateral**	**Right**
18	431	38

Gene expression data was downloaded through the cBioPortal for cancer genomics database [12]. Each sample contains expression levels for each of the 60,488 genes; the gene expressions are given in terms of Transcripts Per Kilobase Million (TPM) values. The aim of this study is to identify genes which are associated with specific tumor locations, and hence we need to use the genes as features and the actual locations as classes to build a model to predict locations for future samples. Since most of the samples are bilateral, we deal with a class imbalance problem. We used the resampling method proposed in [13] as measure to lower the effect of this imbalance.

3.2.1 Resampling

By observing Table 3.1, we clearly notice that there is a class imbalance problem, where the number of samples in the right class (38) is almost twice as large as that of the left class (18). while the number of samples of the bilateral class (431) is more than twenty times larger than the left class and more than ten times larger than the right class.

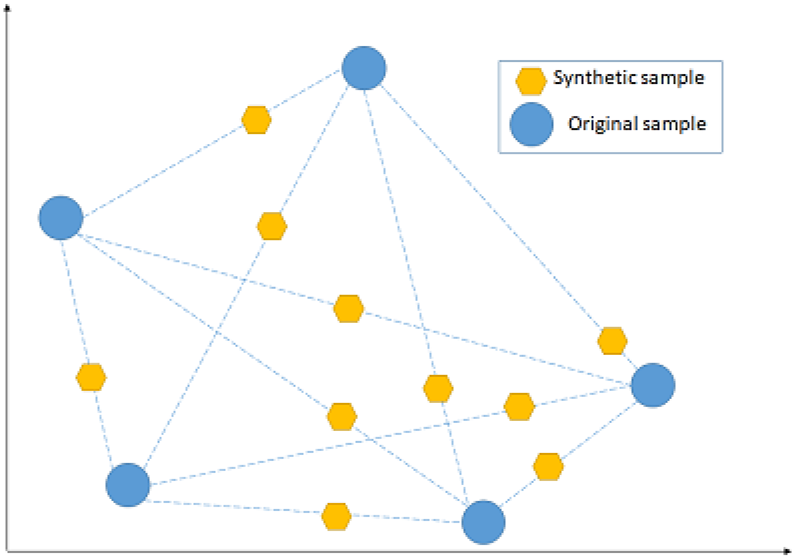

Figure 3.2: Synthetic Minority Oversampling Technique (SMOTE) works by adding new synthetic sample randomly along the line that connects each of the two original samples.

To solve this problem, multiple resampling methods were deployed and tested to identify a method that would yield the best solution for our data set. Oversampling provides a fast solution for classes left and right. This method duplicates samples from the minority classes and adds them until yielding a similar number of samples for each class. Applying oversampling directly did resolve the class imbalance problem and provided high accuracy for classifiers, although after taking a closer look at the samples used in these classifiers, we noticed that there was a major overfitting. Based on the literature [24, 25], we selected the combination of oversampling Synthetic Minority Oversampling Technique (SMOTE) [26] and Neighborhood Cleaning Rule (NCL)[27] for under-sampling the majority class. Junsomboon et al. reported that the combination (NCL+SMOTE) outperfomed another set of methods for handling the imbalance data sets. They have applied this combination on different health related data sets [24]. NCL uses the Wilson's Edited Nearest Neighbor Rule (ENN) to remove majority class outliers [28]. Batista et al. reported a high performance for SMOTE+ENN in handling imbalance data set [25].

NCL works by removing any sample whose class is different from the class of at least two of its three nearest neighbors. SMOTE 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 distance between the two samples and multiplying that value by a random number between 0 and 1. Figure 3.2 shows the behavior of SMOTE.

Applying these two methods allowed us to use three classes that are balanced. Table 3.2 shows the number of samples after applying the SMOTE+ENN resampling methods.

Table 3.2: Number of samples in each prostate cancer tumor location after applying the SMOTE+ENN resampling methods.

70 vs 70	240 vs 240	40 vs 40
Left vs rest	**Bilateral vs rest**	**Right vs rest**

3.2.2 Feature Selection

Dealing with a huge number of features lead us to the problem of curse of dimensionality. As such, we use machine learning techniques to lower the number of features used for classification. We applied the information gain (IG) feature selection method to rank all the genes with a score that relates to the highest information gain against the different classes. We then chose the attributes with the highest scores, discarding those with lower scores. In this chapter, the IG attribute evaluator [14] is used to evaluate each attribute. IG of feature X with respect to class Y is calculated as follows:

$$IG(Y, X) = H(Y) - H(Y|X) \tag{3.1}$$

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

The next step is to choose the best set of attributes (genes) that provide good classification among the different classes.

A wrapper that binds feature selection and classification methods is used. The feature selection method is the minimum redundancy maximum relevance (mRMR), which takes features that contain minimum redundancy while at the same time have high correlation to the classification variable [15]. The equation for minimizing redundancy (W) and maximizing the relevancy (V) is the following:

$$min\{W(S), W\} = \frac{1}{|S|^2} \sum_{i,j \in S} I(i,j), \tag{3.2}$$

and

$$max\{V(S,h), V\} = \frac{1}{|S|} \sum_{i \in S} I(h,i), \tag{3.3}$$

where S is the set of features, $I(i,j)$ is mutual information between features (i,j), h is the class.

3.2.3 Classification

We deal with a multi-class classification problem which is solved by using the one-versus-all approach. We have three different classes which are the three different locations. To apply the one-versus-all approach, we need to create three separate copies from the actual data set. For each data set, we set one of the classes to positive, and the rest of the classes are combined together to form the negative class. We used accuracy, sensitivity and specificity to choose the best classification method.

Multiple classification methods were applied on the data to identify which methods separate the locations better. Accordingly, the probabilistic classifier Naive Bayes that applies Bayes' theorem with the assumption of independence between the features [16] was tested. SVM was also used to build a classification model based on the features selected in the previous step [17]. The other classifier that was tested is random forest [18], which attempts to build multiple decision tree models with different samples and different initial variables.

The Weka open source libraries were used to run different classification algorithms on the minimized number of features to identify which genes are differentially expressed in the different locations [19].

3.3 Results and Discussion

The different classifiers produced varied results as observed in Table 3.3 and Figure 3.3. The classifiers were chosen based on accuracy and precision, as leading high accuracy with low precision is not a good criterion at all. The accuracy measures the number of correctly classified samples divided by the number of all samples, while the precision is the true positive rate which measures the number of true positive calls divided by all positive calls. Table 3.3 shows the actual accuracy and precision for each classifier. The highest accuracy and precision for the different classifiers came from the SVM Radial basis function kernel (SVM-RBF) classifier. Grid search optimization was applied to fine tune the RBF classifier, it was able to separate the different locations by an accuracy of 99%. Random forest managed to result in high accuracy too, while the naive Bayes classifier results were not satisfactory.

Table 3.4 show the actual genes that were identified by SVM-RBF. These genes can be used to predict the location of the prostate cancer tumor very accurately from gene expression data.

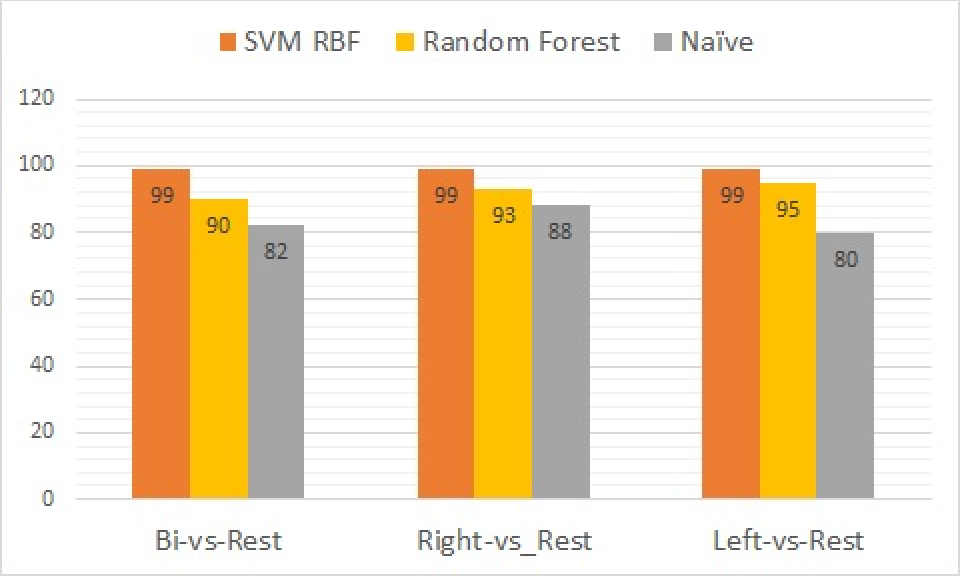

Figure 3.3: Different classifiers accuracy for the different locations.

Table 3.3: Accuracy and precision for classifying each class versus the rest.

Classifier	**Accuracy**	**Precision**	**Accuracy**	**Precision**	**Accuracy**	**Precision**
SVM RBF	99	97	99	97	99	97
Naive Bayes	88	78	82	78	80	78
Random Forest	93	85	90	85	95	85
	Left vs rest		**Bilateral vs rest**		**Right vs rest**	

Table 3.4: Genes that can predict tumors in each location class of the prostate tumor.

Gene	**Gene**	**Gene**
FBXO21	ALG5	HLA-DMB
RTN1	Z99129	SRSF6
NDUFA5	SNAI2	EIF4G2
POP7	MRI1	
	TAF7	
Left vs rest	**Bilateral vs rest**	**Right vs rest**

Throughout our model 10-fold cross-validation was used. The proposed method identified 12 genes that are differentially expressed among the three different possible locations.

It is important to highlight that most of the genes identified in this work have been previously characterized and described to play some role in prostate cancer as well as other types of cancer. SNAI2 is a gene shown [20] to be silenced in prostate cancer and regulates neuroendocrine differentiation, metastasis-suppressor, and pluripotency gene expression.

Likewise, the results shown in [21, 22] indicate that increased TAF1/7 expression is associated with progression of human prostate cancers to the lethal castration-resistant state. In a similar way, the results reported in [23] found that tumor cell expression of HLA-DMB is associated with increased numbers of tumor-infiltrating CD8 T lymphocytes and both are associated with improved survival in advanced serous ovarian cancer.

Figures 3.4-3.6 depict the ROC curves for all the classes versus the rest at each node. The area under the curve AUC for SVM-RBF tends to be further towards the north west with 0.99 value in the three figures, which means the best overall performance across all classes versus the rest. All other classifiers were inconsistent in the three figures. However, random forest performed very well in later false positive rates for both left and right classes with overall performance 0.87, 0.84 in order for both classes. it slightly outperformed the SVM-RBF in one point at both classes. but as we stated earlier, it was inconsistent through out different running parameters for false positive rates.

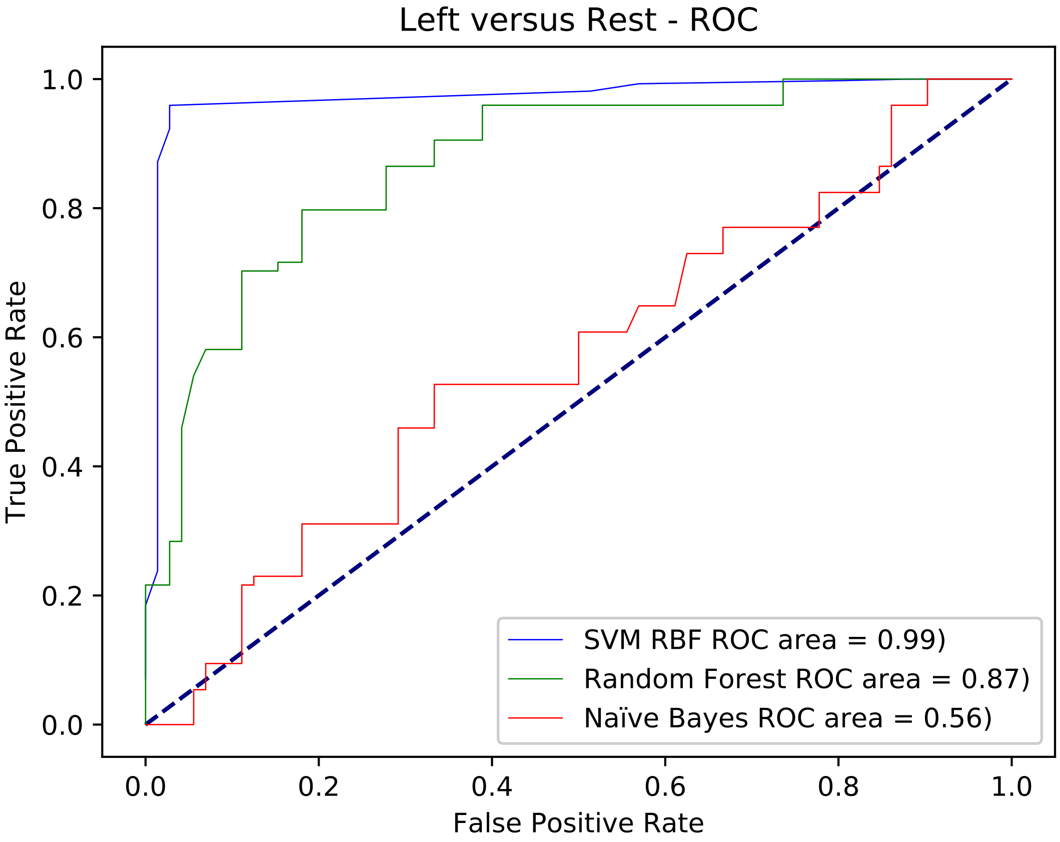

Figure 3.4: The ROC curve for left versus the rest using different classifiers.

3.3.1 Biological insight

We have conducted a thoroughly literature review on the most up to date classification, as well as in the relevant databases and gathered valuable information about the most relevant genes that we have found in our study. A summary for each gene is given below and opens the avenue for further studies as well as additional lab experiments that can corroborate our studies and lead to novel ways of diagnosis, treatment and prognosis of

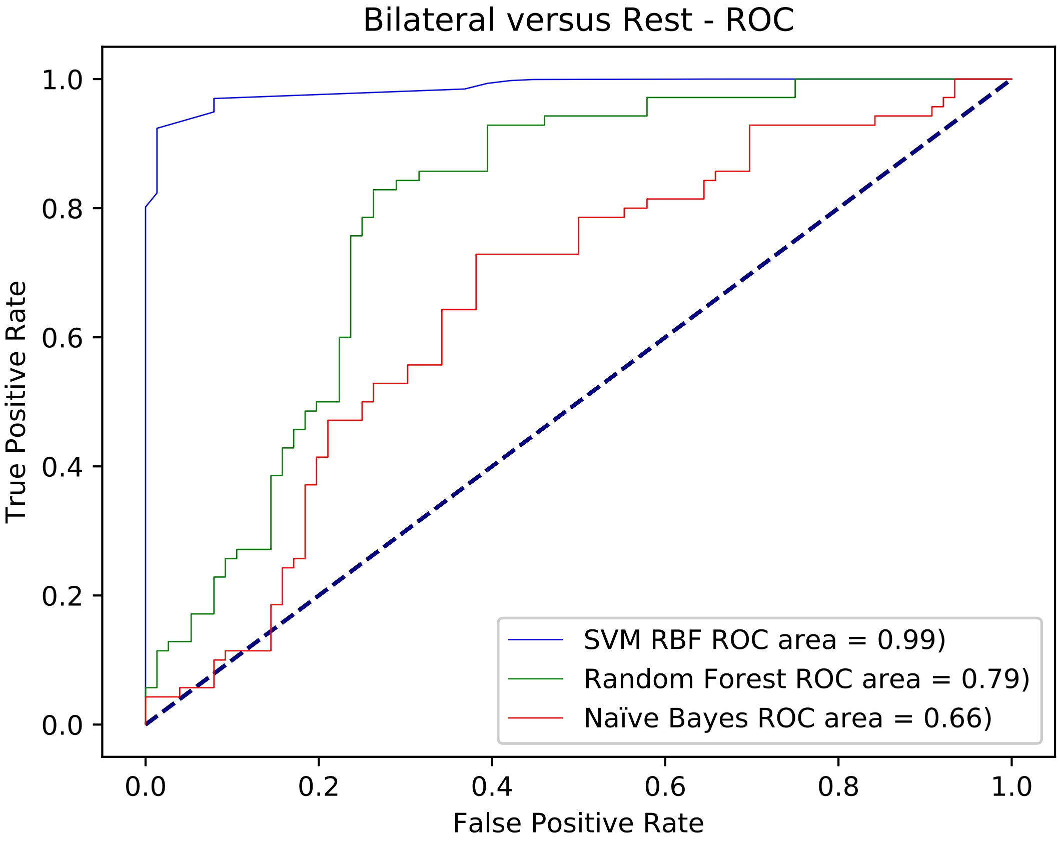

Figure 3.5: The ROC curve for bilateral versus the rest using different classifiers.

the disease.

FBXO21 (F-box protein 21) is part of the multiprotein complex, SCF E3-ligase, which functions in phosphorylation-dependent ubiquitination. FBXO21 may affect prostate cancer through different mechanisms, and here we hypothesize two possibilities. Firstly, ABCB1 is a known tumour drug resistance biomarker because it is a multi-drug efflux pump linked with the development of metastases [29]. FBXO21 tags ABCB1 for proteaso-

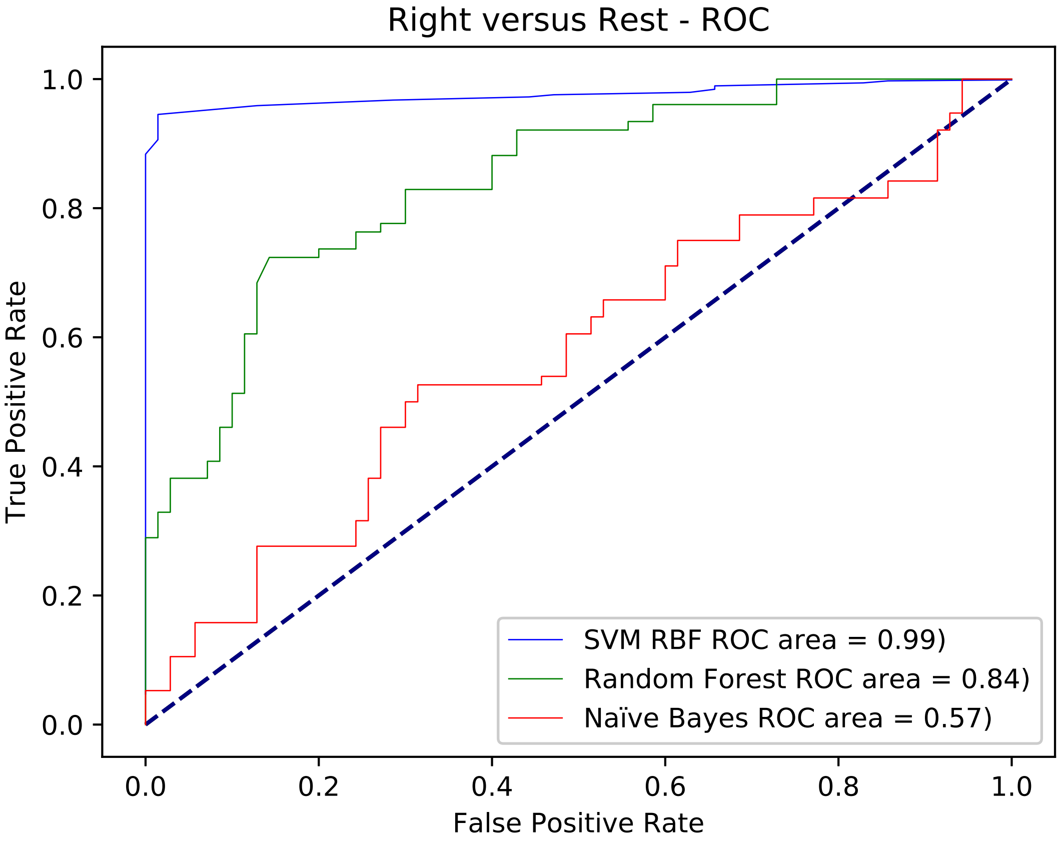

Figure 3.6: The ROC curve for right versus the rest using different classifiers.

mal degradation, whereas inhibition of FBXO21 leads to higher expression level of ABCB1. Secondly, FBXO21 recognizes EID1 in cycling and G0 stage cells and targets it for degradation. EID1 interacts with retinoblastoma tumour suppressor (pRB), melanoma-associated antigen (MAGE), and E1A binding protein p300 (EP300) as well as being involved in the coupling cell cycle exit to cellular differentiation. All available evidence suggests that FBXO21 may be downregulated in prostate cancer, although further research is desirable [30].

RTN1 (reticulon 1) is associated with the endoplasmic reticulum (ER) and is involved in neuroendocrine secretions and membrane trafficking. RTN1 has been known exert a cancer-specific proapoptotic function. Specifically, RTN1-C regulates the two mutually exclusive ER stress-induced apoptosis and DNA damage-induced cell death. Overexpression of RTN1-C results in ER stress-induced cell death mediated by aberrantly increased cytosolic Ca^{2+} due to depletion of ER calcium stores [31]. A recent publicaiton on prostate cancer shows that silencing RTN1 by siRNA enabled androgen-independent proliferation of androgen-dependent prostate cancer tumours. The knockdown of RTN1 increases the nuclear concentration of HDAC8, a multifunctional histone deacetylase that regulates activity of transcription factors such as nuclear hormone receptors [32]. In particular, it is known that ceramide inhibits androgen receptor activity and inhibits androgen-independent growth by activation of protein phosphatase 2A (PP2A) [33]. However, HDAC8-induced depletion of SPTSSA in the ER compromises the ER-localized ceramide biosynthesis pathway, leading to downregulation of ceramide, partial inhibition of PP2A and androgen receptor activation in androgen-deprived conditions [32]. Consequently, RTN1 may be a proto-oncogene associated with aggressive, malignant and androgen-independent prostate cancer.

NDUFA5 (NADH:ubiquinone oxidoreductase subunit A5) is localized to the inner mitochondrial membrane and functions in the NADH two-electron reduction of ubiquinone [34]. Complex I, also known as NADH-ubiquinone oxidoreductase, is the first complex of the mitochondrial oxidative phosphorylation (OXPHOS) system. The energy released is coupled with generation of the electrochemical gradient necessary for ATP synthesis [35]. As expected, NDUFA5 activity is lower in hypoxic cells [36]. The Warburg effect states that tumour cells demonstrate drastically increased glycolysis activity compared to oxidative phosphorylation due to target genes upregulated by hypoxia-inducible factor (HIF) [37]. On the other hand, NDUFA5 is upregulated in HPV+ cervical cancer and its overexpression may play a role in carcinogenesis through acquiring growth advantage and resistance against an apoptotic signal [34]. In a recent publication, NDUFA5 also gained copy numbers in both low-grade and high-grade gliomas. Therefore, NDUFA5 may also be upregulated in prostate cancer, although further research is necessary to confirm this hypothesis [38].

POP7 (POP7 homolog, ribonuclease P/MRP subunit) is discovered in S. cerevisiae. POP7 heterodimerizes to POP6 and binds to the P3 domain of catalytic ribonucleoproteins RNase MRP (mitochondrial RNA processing) and Rpr1 RNA [39]. RNase MRP is critically important to the viability of eukaryotic cells because it is localized in the nucleolus and is involved in processing mitochondrial RNAs and regulating mitochondrial DNA replication [40]. POP1/POP6/POP7 complex is required for telomere elongation protein (Est1) to associate with the RNP, which is critical during the process of mitosis for the cell lifespan before its senescence [41]. Despite the critical importance of POP7, no known human diseases are associated with this gene currently. Further research will be important to explore the biological significance of POP7.

HLA-DMB (major histocompatibility complex class II, DM beta) is a subunit of the HLA class II heterodimer found embedded in intracellular vesicles. In antigen-presenting cells (APC), HLA-DMB is critical in the antigen-presentation machinery by releasing class II-associated invariant chain peptide (CLIP) from MHC class II molecules so that the peptide binding site is free to interact with antigenic peptides [42]. A recent publication on prostate cancer research found that HLA-DMB is coexpressed with ERG and silencing ERG led to significant underexpression of HLA-DMB. Thus, HLA-DMB is an upregulated tumour-associated gene in prostate cancer [43].

SRSF6 (Serine and Arginine rich Splicing Factor 6) modulates a splicing factor protein called SFRS12 to determine alternative splicing of mRNA. In a recent publication on colorectal cancer, SRSF6 targeted ZO-1 (tight junction protein 1) exon23 for alternative splicing, consequentially disrupting ZO-1 from regulating tight junctions between adjacent cells [44]. Furthermore, SRSF6 is the direct target of LINC01133, a key SRSF6 modulates a splicing factor protein called SFRS12 to determine alternative splicing of mRNA. In a 2017 paper on colorectal cancer, SRSF6 targeted ZO-1 (tight junction protein 1) exon23 for alternative splicing, consequentially disrupting ZO-1 from regulating tight junctions between adjacent cells. In addition, SRSF6 is the direct target of LINC01133, a key downstream protein of TGF-β signaling pathway which is critical for cell growth and differentiation [45]. Silencing SRSF6 in colorectal cancer tissues inhibited epithelial-mesenchymal transition, tissue invasion, and metastasis. A study on wound healing found that overexpression of SRSF6 induces skin hyperplasia due to SRSF6 upregulating Tenascin C and suppressing the normal epithelial differentiation mechanism. Therefore, SRSF6 may be upregulated in prostate cancer [44].

EIF4G2 gene, Eukaryotic Translation Initiation Factor 4 Gamma 2 is a cap - binding protein complex which has three sub units – eiF4A, eiF4E eiF4G. The gene is known to upregulate p21, a cyclin dependant kinase inhibitor and interleukin 6 [46]. Higher expression levels of p21 oncogene protein are found with increasing prostate cancer tumor grade [47]. Interleukin 6 is involved in the progression of prostate cancer [48], and is used as a clinicopathological feature by detecting the levels in serum [49]. With the upregulated expression levels of EIF4G2 gene in prostate cancer, it can be used as a potential marker for studying the progression of the disease.

Interestingly, EIF4G2 and HLA-DMB which are part of the gene set that can identify right side from the rest, they are both part of Allograft rejection SuperPath pathway [50].

The discovery of fusion protein transcripts in the recent times have helped studying prostate cancer development with much detail. ALG5, Dolichyl-Phosphate Beta-Glucosyltransferase and PIGU, Phosphatidylinositol Glycan Anchor Biosynthesis Class forms a chimeric-fusion protein transcript in which glucosyltransferase, the head from ALG5 is retained but GPI transamidase, the tail has been eliminated in PIGU resulting in the loss of functionality of both the genes [51]. The uncommon joining of the genes would result in serious complications in the overall environment of the cell causing further progression of the cancer. The transcription of the fused ALG5-PIGU is androgen independent [52]. Fusion protein transcripts will serve as an important biomarker both in detection and treatment of Prostate Cancer.

SNAI2, Snail Family Transcriptional Repressor 2 encodes zinc-finger protein of the Snail family transcription factors, is involved in the generation and migration of neural crest cells in embryonic stages which is driven by epithelial to mesenchymal transition (EMT). Presence of neuroendocrine cells in nests - neuroendocrine differentiation (NED) is a known histological marker for prostate Cancer. SNAI2 expression is down regulated in prostate cancer and silencing of the gene may turn on neuroendocrine differentiation, pluripotent genes and turn on specific metastasis suppressors [53]. SNAI2 knockdown initiating metastatic suppressor genes involves many pathways and further research is needed to derive a conclusion. Studies of SNAI2 gene regulation properties will help us in understanding the development of prostate cancer.

MRI1, Methylthioribose-1-Phosphate Isomerase 1 gene helps in catalyses of methionine, an important amino acid, in methionine salvage pathway. Development of certain cancers like prostate, glioma, bladder, breast, melanoma are dependent on methionine [54, 55]. To understand the dependency of methionine in prostate cancer a study has been conducted on patients who were not receiving any conventional treatment and were undergoing an intensive lifestyle program with a restricted methionine vegan diet. Analysis of serum samples revealed that there was a 70% inhibition of the growth androgen sensitive prostate adenocarcinoma (LNCaP) cells [56]. The data suggests that methionine restricted diet and lifestyle changes may help in slowing down the development of prostate cancer.

3.4 Conclusion

Understanding gene activity in the prostate cancer laterality may help to guide the diagnosis and treatment of the disease. In this work, we have proposed a machine learning method that is capable of predicting with a high accuracy the tumor location in a cancer infected prostate. As a result, we have found genes as indicators that can differentiate the three locations of prostate cancer with high accuracy. The contributions of this study are two-fold. The proposed machine learning system can be used as a protocol for other types of cancer and other clinical problems in cancer studies. It also open the doors for potential biomarkers that can be further tested in wet-lab scenarios with the hope to move to clinical trials in order to replace the invasive biopsy or inaccurate image scanning.

The literature shows strong relations between prostate cancer metastasis and the computationally derived genes. Wet-lab experiments and RNA-seq profiling of those genes will better explore the relation between the findings and the prostate cancer laterality, which will potentially help the prognosis of the disease.

References

[1] B. Stewart, P. Wild, et al. World cancer report 2014. Health, 2017.

[2] S. Parpart, A. Rudis, A. Schreck, N. Dewan, and P. Warren. Sensitivity and specificity in prostate cancer screening methods and strategies. *J Young Investig*, 2007.

[3] W. Stewart, S. Lizama, K. Peairs, F. Sateia, and Y. Choi. Screening for prostate cancer. In Seminars in Oncology. Elsevier, 2017.

[4] J. Rosario, J. Lane, C. Metcalfe, L. Donovan, A. Doble, L. Goodwin, M. Davis, W. Catto, K. Avery, E. Neal, et al. Short term outcomes of prostate biopsy in men tested for cancer by prostate specific antigen: prospective evaluation within protect study. Bmj, 344:d7894, 2012.

[5] S. Singireddy, A. Alkhateeb, I. Rezaeian, L. Rueda, D. Cavallo-Medved, and L. Porter. Identifying differentially expressed transcripts associated with prostate cancer progression using RNA-seq and machine learning techniques. In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on, pages 1–5. IEEE, 2015.

[6] Alkhateeb, A., Rezaeian, I., Singireddy, S., & Rueda, L. (2015, November). Obtaining biomarkers in cancer progression from outliers of time-series clusters. In 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 889-896). IEEE.

[7] Hamzeh, O., Alkhateeb, A., Rezaeian, I., Karkar, A., & Rueda, L. (2017, April). Finding transcripts associated with prostate cancer gleason stages using next generation sequencing and machine learning techniques. In International Conference on Bioinformatics and Biomedical Engineering (pp. 337-348). Springer, Cham.

[8] Y. Ping, D. Landsittel, L. Jing, J. Nelson, B. Ren, L. Liu, C. McDonald. Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. Journal of Clinical Oncology 22.14 (2004): 2790-2799.

[9] Y. Artan, A. Haider, L. Langer, H. Kwast, J. Evans, Y. Yang, N. Wernick, J. Trachtenberg, and I. Yetik. Prostate Cancer Localization With Multispectral MRI Using Cost-Sensitive Support Vector Machines and Conditional Random Fields. IEEE Transactions on Image Processing, 19(9):2444–2455, 2010.

[10] R. Sanz-Pamplona, D. Cordero, A. Berenguer, F. Lejbkowicz, H. Rennert, R. Salazar, S. Biondo, X. Sanjuan, A. Pujana, L. Rozek. Gene expression differences between colon and rectum tumors. Clinical Cancer Research, 2011.

[11] O. Hamzeh, A. Alkhateeb, and L. Rueda. Predicting tumor locations in prostate cancer tissue using gene expression. In International Conference on Bioinformatics and Biomedical Engineering, pages 343–351. Springer, 2018.

[12] GDC, Portal.gdc.cancer.gov, 2017. [Online]. Available: https://portal.gdc.cancer.gov/. [Accessed: 15- Aug- 2017].

[13] A. Estabrooks, T. Jo, N. Japkowicz. A multiple resampling method for learning from imbalanced data sets. Computational Intelligence, 20(1), 18-36.

[14] Novakovic J (2009) Using information gain attribute evaluation to classify sonar targets. In 17th Telecommunications forum TELFOR (pp. 24-26)

[15] H. Peng, F. Long, C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226-1238

[16] P. Domingos, M. Pazzani (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, pp. 29(2-3): 103-130

[17] C. Cortes, V. Vapnik (1995). Support-vector networks. Machine Learning, 20(3), 273-297

[18] F. Rodriguez, B. Ghimire, J. Rogan, M. Olmo, & P. Sanchez. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67, 93-104.

[19] E. Frank, M. Hall, I. Witten. (2016) The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, Fourth Edition

[20] S. Esposito, V. Russo, I. Airoldi, G. Tupone, C. Sorrentino, G. Barbarito, E. Di Carlo (2015). SNAI2/Slug gene is silenced in prostate cancer and regulates neuroendocrine differentiation, metastasis-suppressor and pluripotency gene expression. Oncotarget, 6(19), 17121–17134.

[21] P. Tavassoli, L. Wafa, H. Cheng, A. Zoubeidi, L. Fazli, M. Gleave, R. Snoek, P. Rennie; TAF1 Differentially Enhances Androgen Receptor Transcriptional Activity via Its N-Terminal Kinase and Ubiquitin-Activating and -Conjugating Domains, Molecular Endocrinology, Volume 24, Issue 4, 1 April 2010, Pages 696–708, https://doi.org/10.1210/me.2009-0229

[22] S. Bhattacharya, X. Lou, P. Hwang, K. Rajashankar, X. Wang, J. Gustafsson, R. Fletterick, R. Jacobson, and P. Webb. Structural and functional insight into TAF1–TAF7, a subcomplex of transcription factor II D PNAS 2014 111 (25) 9103-9108; published ahead of print June 10, 2014, doi:10.1073/pnas.1408293111

[23] M. Callahan , Z. Nagymanyoki,T. Bonome, et al. Increased Hla-Dmb Expression In The Tumor Epithelium Is Associated With Increased Cytotoxic T Lymphocyte Infil-

tration And Improved Prognosis In Advanced Serous Ovarian Cancer. Clinical Cancer Research. 2008;14(23):7667-7673. doi:10.1158/1078-0432.CCR-08-0479.

[24] N. Junsomboon and T. Phienthrakul. Combining over-sampling and under-sampling techniques for imbalance data set. In Proceedings of the 9th International Conference on Machine Learning and Computing, pages 243–247. ACM, 2017.

[25] G. Batista, R. Prati, and M. Monard. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1):20–29, 2004.

[26] N. Chawla, K. Bowyer, O. Hall, & P. Kegelmeyer. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[27] J. Laurikkala. Improving Identification of Difficult Small Classes by Balancing Class Distribution. Tech. Rep. A-2001-2, University of Tampere, 2001.

[28] D. Wilson. Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, (3):408–421, 1972.

[29] A. Ravindranath, S. Kaur, R. Wernyj, M. Kumaran, K. Gonzalez, R. Chan, E. Lim, K. Madura, and L. Rodriguez. CD44 promotes multi-drug resistance by protecting P-glycoprotein from FBXO21-mediated ubiquitination. Oncotarget, 6(28):26308, 2015.

[30] Z. Cuiyan, X. Li, G. Adelmant, J. Dobbins, C. Geisen, M. Oser, K. Wucherpfenning, J. Marto, and W. Kaelin. Peptidic degron in EID1 is recognized by an SCF E3 ligase complex containing the orphan F-box protein FBXO21. Proceedings of the National Academy of Sciences 112, no. 50 (2015): 15372-15377.

[31] L. Chen, L. Wan, J. Du, and Y. Shen. Identification of MANF as a protein interacting with RTN1-C. Acta Biochimica et Biophysica Sinica, 47(2):91–97, 2014.

[32] E. Levina, H. Ji, M. Chen, M. Baig, D. Oliver, P. Ohouo, C. Lim, G. Schools, S. Carmack, Y. Ding, et al. Identification of novel genes that regulate androgen receptor signaling and growth of androgen-deprived prostate cancer cells. Oncotarget, 6(15):13088, 2015.

[33] A. Bhardwaj, S. Singh, S. Srivastava, R. Honkanen, E. Reed, and A. Singh. Modulation of Protein Phosphatase 2A Activity Alters Androgen-Independent Growth of Prostate Cancer Cells: Therapeutic Implications. Molecular Cancer Therapeutics, pages molcanther–1096, 2011.

[34] K. Denninger, T. Litman, T. Marstrand, K. Moller, L. Svensson, T. Labuda, and A. Andersson. Kinetics of gene expression and bone remodelling in the clinical phase of collagen-induced arthritis. Arthritis Research & therapy, 17(1):43, 2015.

[35] Y. Sato, M. Inoue, T. Yoshizawa, and K. Yamagata. Sato, Yoshifumi, Masahiro Inoue, Tatsuya Yoshizawa, and Kazuya Yamagata. Moderate hypoxia induces Beta-cell dysfunction with HIF-1–independent gene expression changes. PloS One, 9(12):e114868, 2014.

[36] S. Peralta, A. Torraco, T. Wenz, S. Garcia, F. Diaz, and C. Moraes. Partial complex I deficiency due to the CNS conditional ablation of Ndufa5 results in a mild chronic encephalopathy but no increase in oxidative damage. Human molecular genetics, 23(6), 1399-1412. Human Molecular Genetics, 23(6):1399–1412, 2013.

[37] P. Hsu and D. Sabatini. Cancer cell metabolism: Warburg and beyond. Cell, 134(5):703–707, 2008.

[38] Y. Li, D. Wang, L. Wang, J. Yu, D. Du, Y. Chen, P. Gao, D. Wang, . Yu, F. Zhang, et al. Distinct genomic aberrations between low-grade and high-grade gliomas of chinese patients. PLoS One, 8(2):e57168, 2013.

[39] R. Reiner, N. Alfiya-Mor, M. Demma, D. Wesolowski, S. Altman, and N. Jarrous. RNA binding properties of conserved protein subunits of human RNase P. Nucleic Acids Research, 39(13):5704–5714, 2011.

[40] O. Esakova and A. Krasilnikov. Of proteins and RNA: The RNase P/MRP family. RNA, 2010.

[41] K. Collins. The biogenesis and regulation of telomerase holoenzymes. Nature Reviews Molecular Cell Biology, 7(7):484, 2006.

[42] S. Theodoridis, A. Pikrakis, K. Koutroumbas, and D. Cavouras. Introduction to Pattern Recognition: A Matlab Approach: A Matlab Approach. Academic Press, 2010.

[43] P. Paulo, F. Ribeiro, J. Santos, D. Mesquita, M. Almeida, J. Silva, H. Itkonen, R. Henrique, C. Jerónimo, A. Sveen, et al. Molecular Subtyping of Primary Prostate Cancer Reveals Specific and Shared Target Genes of Different ETS Rearrangements. Neoplasia (New York, NY), 14(7):600, 2012.

[44] L. Wan, W. Yu, E. Shen, W. Sun, Y. Liu, J. Kong, Y. Wu, F. Han, L. Zhang, T. Yu, et al. Srsf6-regulated alternative splicing that promotes tumour progression offers a therapy target for colorectal cancer. Gut, pages gutjnl–2017, 2017.

[45] H. Kim, G. Lee, K. Choi, D. Kim, J. Ryu, K. Hwang, K. Na, C. Choi, Ja Hong Kuh, M. Chung, et al. SRSF5: a novel marker for small-cell lung cancer and pleural metastatic cancer. Lung Cancer, 99:57–65, 2016.

[46] R. Mori, S. Xiong, Q. Wang, C. Tarabolous, H. Shimada, E. Panteris, K. Danenberg, P. Danenberg, and J. Pinski. Gene profiling and pathway analysis of neuroendocrine transdifferentiated prostate cancer cells. The Prostate, 69(1):12–23, 2009.

[47] M. Viola, F. Fromowitz, S. Oravez, S. Deb, G. Finkel, J. Lundy, P. Hand, A. Thor, and J. Schlom. Expression of RAS oncogene P21 in prostate cancer. New England Journal of Medicine, 314(3):133–137, 1986.

[48] T. Chung, J. Yu, M. Spiotto, M. Bartkowski, and J. Simons. Characterization of the role of il-6 in the progression of prostate cancer. The Prostate, 38(3):199–207, 1999.

[49] V. Michalaki, K. Syrigos, P. Charles, and J. Waxman. Serum levels of IL-6 and TNF-α correlate with clinicopathological features and patient survival in patients with prostate cancer. British Journal of Cancer, 90(12):2312, 2004.

[50] F. Belinky, N. Nativ, G. Stelzer, S. Zimmerman, T. Stein, M. Safran, and D. Lancet. Pathcards: multi-source consolidation of human biological pathways. Database, 2015, 2015.

[51] J. Luo, S. Liu, Z. Zuo, R. Chen, G. Tseng, and P. Yan. Discovery and classification of fusion transcripts in prostate cancer and normal prostate tissue. The American Journal of Pathology, 185(7):1834–1845, 2015.

[52] D. Pflueger, S. Terry, A. Sboner, L. Habegger, R. Esgueva, P. Lin, M. Svensson, N. Kitabayashi, B. Moss, Theresa Y MacDonald, et al. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Research, 2010.

[53] S. Esposito, M. Russo, I. Airoldi, M. Tupone, C. Sorrentino, G. Barbarito, S. Di Meo, and E. Di Carlo. SNAI2/sLUG gene is silenced in prostate cancer and regulates neuroendocrine differentiation, metastasis-suppressor and pluripotency gene expression. Oncotarget, 6(19):17121, 2015.

[54] P. Cavuoto and M. Fenech. A review of methionine dependency and the role of methionine restriction in cancer growth control and life-span extension. Cancer Treatment Reviews, 38(6):726–736, 2012.

[55] F. Breillout, E. Antoine, and M. Poupon. Methionine dependency of malignant tumors: a possible approach for therapy. JNCI: Journal of the National Cancer Institute, 82(20):1628–1632, 1990.

[56] D. Ornish, G. Weidner, W. Fair, R. Marlin, E. Pettengill, C. Raisin, S. Dunn-Emke, L. Crutchfield, F. Jacobs, R. Barnard, et al. Intensive lifestyle changes may affect the progression of prostate cancer. The Journal of Urology, 174(3):1065–1070, 2005.

Chapter 4

An Integrative Knowledge-based Method to Identify Cancer Biomarkers Based on Gene-Disease Relations

4.1 Introduction

Next generation sequencing (NGS) is one of the most important technologies to explore genetic associations in medical studies. NGS technologies large data sets, which provides a detailed view of the human genome [1]. The sensitivity, speed and reduced cost per sample make it an attractive option, especially when compared to older technologies. These details include the actual DNA or RNA in various forms. The amount of raw data produced require many computational steps to produce high-quality data that can be later used to obtain information [2]; this information can be in the form of mutations, copy number alterations (CNA), and others. Cancer is known to be a genetic disorder which might be heritable or in terms of somatic mutations. NGS has a significant impact on the detection, and treatment of this disease [3].

The generated data sets from this technology are huge and involve a big challenge. Machine learning techniques, on the other hand, have proved to be useful for such large data sets and provided excellent results for classifying cancer states based on gene expression, CNA levels or mutations of certain genes. Machine learning techniques construct models that can be used to predict certain biological characteristics from multi-dimensional data sets, these predictive models are becoming essential to modern biological research [4] [5]. Machine learning offers so many techniques that can be used to extract information from these data sets. The data sets are made of a huge number of genes or transcripts expressions. These gene/transcripts are called attributes in machine learning. One of the first steps in machine learning is to reduce the number of attributes. In a typical data-set that contains gene expressions, the number of attributes is approximately 35,000 to 40,000 genes, and can reach up to 65,000 transcripts. The reduction is done using a procedure called feature selection [6], in which the attributes that do not affect the performance of the model are removed.

Feature selection is done usually in two steps. The first step is to give a score for each attribute against the medical feature that the model tries to predict. This step is called filter-based feature selection. The second step is to find sets of attributes that can be used together to classify a certain clinical sign, this step is called wrapper-based feature selection. The classifiers usually provide a model that can classify a certain clinical sign [7]. For example a certain Gleason score in prostate cancer or a certain breast cancer stage. The produced model can be evaluated using certain measures, for e.g. accuracy, specificity or Area Under the Curve (AUC) [8].

The last step would require an expert to look into the findings and to search the latest literature to verify the results. Hamzeh et al. applied a method to identify biomarkers that can predict Gleason score stages for prostate cancer patients using machine learning techniques [9]. Gleason score is a grading system which is widely used to describe the aggressiveness of prostate cancer, it was first introduced by Dr. Donald Gleason back in the 1960's. A recent study proposed to join certain Gleason scores together, which will create a new 5 Grade Group system, this simplifies the Gleason score[10]. Disease to gene relation is a field that has been studied widely, and the findings have been published in medical journals and scientific papers. In this regard, DisGeNET [11] is a database that collects such knowledge and provides a tool that can be queried to find proven relations between diseases and genes. In this work, we integrate the latest knowledge from the literature as a step in the feature selection method. The results show an increase in the number and relevance of cancer related genes that can be used in a predictive model.

DisGeNET integrates data from expert-curated repositories, GWAS catalogues, animal models and the literature. DisGeNET data are homogeneously annotated with controlled vocabularies and community-driven ontologies. DisGeNET utilizes the following resources to provide the disease to gene relations: The Comparative Toxicogenomics Database (CTD) [12], UniProt [13], ClinVar [14], Orphanet [15], The GWAS Catalog [16], The Rat Genome Database (RGD) [17], The Mouse Genome Database (MGD) [18], and The Genetic Association Database (GAD) [19]. It also incorporates the literature directly using text-mining approaches like The Literature Human Gene Derived Network (LHGDN) [20] and BeFree data, obtained using the BeFree System, which extracts gene-disease associations from MEDLINE abstracts [21] [22].

This database provides many ways to collect its findings, either through the main Web portal, a Web API, a SQL database or an all in one file. The results would be a score that relates a gene to disease.

We are proposing a new machine learning pipeline that utilizes proven literature knowledge to identify bio-markers that can be used to classify certain clinical attributes.

4.2 Methods

The proposed machine learning method starts with a basic data pre-procession step followed by two different filter-based feature selection methods. The output of the two methods are two different lists, each one includes the gene names that affect the actual class. We then create a new list from the combination of the two lists and create a third list from the intersection of the two lists. The combined version is sent to DisGeNET to obtain the scores against the actual disease that we are looking for, and the intersection list is processed in another function that checks Pearson's correlation [26] among the different genes depending on their expression levels. The two methods produce different scores for each gene. The final score for all the genes is calculated, and then sorted in a descending

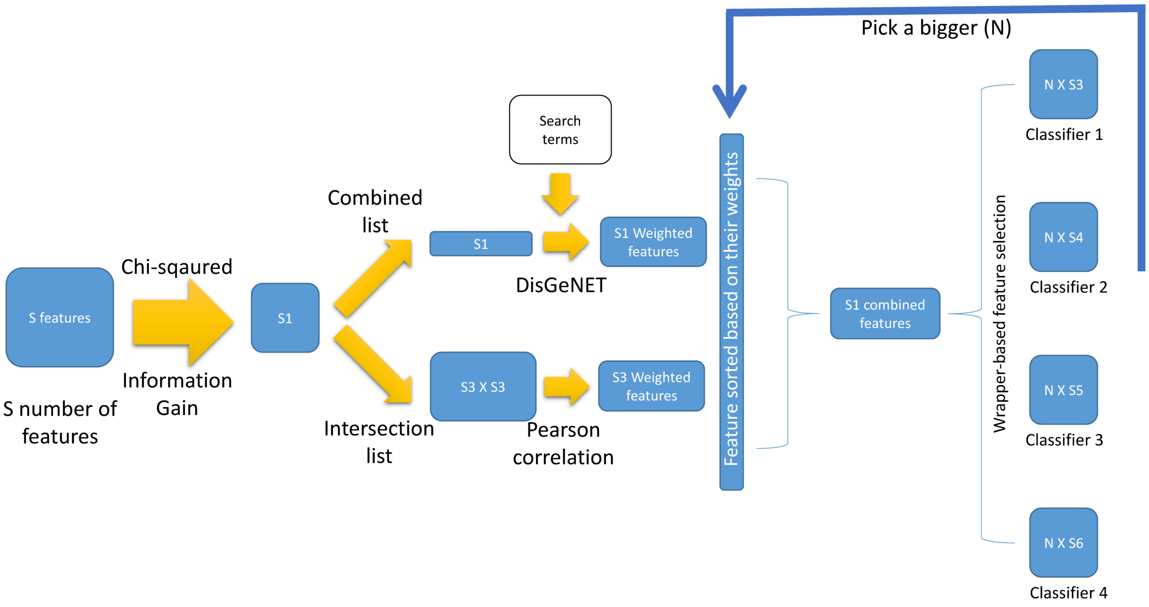

Figure 4.1: Machine learning pipeline used on the proposed method and testing.

order according to the final score. A wrapper-based filter selection method picks the top N number of genes in the list and tries to find the best subset with the best classifier according to their ROC. After this step, the wrapper-based feature selection picks a larger N to find another subset and calculates the ROC for the new subset and the new classifier. This process continues until all the genes in the sorted final list are considered. We then pick the best subset based on ROC and relations to the disease that we are looking for, as the first generated subset will have genes that are more related to the given disease. Thus, the final choice would be specific to each researcher. Figure 4.1 illustrates the machine learning pipeline used by the proposed method. The details of each step are explained below.

4.2.1 Pre-processing

This step checks for missing data and fills the missing cells with the median values. In the same step, we also look for attributes whose values are not changing throughout the samples. These attributes are usually irrelevant for the model, and so they are deleted at this stage. A specific step that is needed for DisGeNET makes sure that the gene names are actually following the HGNC [23] naming schema. If the gene names correspond to the Ensembl genes naming codes [24], an actual name converter is used to convert these into the HGNC. If the data-set includes transcript names, they need to be converted to the HGNC naming schema, and again the developed converter applied the required conversion.

4.2.2 Filter-based feature selection

As the data for gene expressions, CNA or mutation are numerical and continuous, a good filter-based scoring criterion is Information Gain (IG) [25].

We used IG to rank all the attribute with a score that relates to the highest information gain against the different classes of choice.

IG of feature X with respect to class Y is calculated as follows:

$$IG(Y, X) = H(Y) - H(Y|X), \tag{4.1}$$

Here, $H(Y)$ is the entropy of class Y, $H(Y|X)$ is the conditional entropy of Y given X, $p(y)$ is the probability of y, $p(x)$ is the probability of x and $p(y|x)$ is the probability of y given x.

After obtaining the scores for each attribute using IG, we save the names of the gene with scores higher than zero in a list.

We also use another filter based method, Chi-squared which measures the degree of independence of each feature:

$$\chi^2(Y, X) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \tag{4.2}$$

where A is the number of times feature X and class Y co occur, B in the number of times X occurs without Y , C in the number of times Y occurs without X , D in the number of times neither X and Y occurs, and N is the total number of samples.

We run Chi-squared on the original data to generate another list that contains the gene name with scores higher than zero in a list. The two lists are combined into a new combined list that includes all the gene names that have a positive score from each of the methods. Another list is created from the intersection of the two lists.

The intersection list is used in a Pearson's correlation test, which is a statistical method that finds the correlation between two genes based on their expression levels as follows:

$$p(X,Y) = \frac{conv(X,Y)}{\sigma x \sigma y}, \tag{4.3}$$

where *cov* is the co-variance between X and Y, σx is the standard deviation of x, and σy is the standard deviation of y.

4.2.3 DisGeNET

The combined list that is generated in the filter based feature selections step is used during this step. We noticed that using the online API version of DisGeNET to query each gene in the list against the given diseases did not work well as the number of genes was more than 5,000 genes and it took a very long time to query each gene. We used the offline Command Separated Values (CSV) version, which provides instant responses to each query. DisGeNET provides a single score against each query, the query itself requires the gene name in HGNC gene naming schema and requires the disease name. As in real life, when an expert verifies the biomarkers, they usually check if the gene is related to a particular disease (specific cancer type or stage). They also check if the gene is related to cancer in general. As such, we had to do the same, each gene will be DisGeNET is queried twice for each gene. The first query includes the gene name and the particular cancer type we are looking for, while the second query includes the gene name and the word ”cancer”. Our method utilizes RegEx [27] to find any disease name that includes the word ”cancer”. This means that each gene has two scores from each query. We give the specific cancer score a high weight δ and the score returned for cancer a lower weight γ. Thus, the total score for the gene is:

$$s(X) = s(a) * \delta + s(b) * \gamma + s(c) * \beta \tag{4.4}$$

where $s(x)$ is the final score, $s(a)$ is the score returned for the specific cancer type, $s(b)$ is the score returned for the word ”cancer”, $s(c)$ is the score obtained from the Pearson correlation, δ, γ, β are user-defined weights.

δ, γ, β are weights defined by the researcher depending on the actual use of the method itself, increasing the value of δ will prioritize genes that are related to the specific key-word given, while increasing the value of γ will increase the priority of genes that are known to be related to cancer in general, and increasing the value of β will increase the priority of genes that are correlated based on their gene expression values.

A new data set is generated from the set of genes that are part of the final calculated score, this list is sorted in descending order based on the calculated score.

4.2.4 Wrapper-based feature selection

We choose the top N genes from the previously scored and sorted list to start this step. We use a wrapper-based feature selection method that utilizes the minimum redundancy maximum relevance (mRMR) method. This method fuses feature selection and a classification method to find a subset that can classify with high accuracy and specificity. This produces a good AUC. It does this by taking features that contain minimum redundancy while at the same time have a high correlation to the classification variable [28]. The equation for minimizing redundancy W and maximizing the relevancy V is the following:

$$W = \frac{1}{|S|^2} \sum_{i,j \in S} I(i,j), \tag{4.5}$$

and

$$V = \frac{1}{|S|} \sum_{i \in S} I(h,i), \tag{4.6}$$

where S is the set of features, $I(i,j)$ is mutual information between features (i,j), and h is the class.

In this step, we make sure to use multiple classifiers, since it is not guaranteed that a specific classifier would perform on all data sets. The classifiers that were used during this step are: Support Vector Machine with the Radial Basis Function kernel (SVB-RBF) [29], Naive Bayes [30], Random Forest [31] and K-Nearest Neighbour [32].

The output of this step would be a single subset generated with each of the four classifiers, and each one will have its own accuracy, specificity, and ROC.

4.2.5 Back-propagation

In this step, we go back to the wrapper-based feature selection step and choose a larger value for N and repeat the last wrapper-based feature selection method again to generate another subset for each of the classifiers, the results are saved for each value of N. We continue doing this until the value of N is equal to the number of genes in the scored list.

By the end of this final step, we will have scores from each of the four classifiers for each of the values of N used. The final decision on which subset and classifier to choose can be automated based on the larger value of the AUC and the smallest value of N used. Instead, the researcher can pick which set of genes to choose. Taking into considerations, that the lower number of N means that genes are most likely related to the given key-word and that provide genes that are related to the clinical attributes in study.

4.3 Results and Discussion

In our previous work [9], an RNA-Seq data set of 104 prostate cancer patients was analyzed. This data set is publicly available from the National Center for Biotechnology Information (NCBI) with Gene Expression Omnibus (GEO) number GSE54460 [33]. It includes samples with different Gleason stages, and has been analyzed using machine learning techniques to identify transcripts that are linked to prostate progression.

In that study, filter-based feature selection is performed first using IG. Then, a wrapper-based feature selection was performed on the resulting genes to find the best possible subset that are able to predict the Gleason score group with the highest accuracy possible. After the subset is identified for each of the Gleason score, an expert has to take revise the genes that were used to classify each stage and look into the literature for a proof that these genes are already linked to prostate cancer or at least related to types of cancer. The expert found that only seven out of the 26 genes that we identified were already found in the literature. We applied the new method on the same data set.

We started with the pre-processing step, and noticed that the transcripts were used in this data set. These needed to be converted to the NHGC naming schema. We then ran the two filter-based feature selection methods and identified the combined list and the intersected list. The combined list was processed with the DigGeNET and the scores for each gene were calculated. The other intersected list was processed with Pearson's correlation test and the final scores for each gene were calculated. The genes were sorted in descending order based on their score.

We then picked $N = 500$ to start the wrapper-based feature selection with the top 500 genes. Wrapper-based feature selection provided a different subset for each of the four classifiers used. We then started increasing N by 500 genes each time and continued repeating the wrapper-based feature selection step with the a new value of N, until all the genes in the sorted list were used. With the new method, we were able to obtain similar results to the original method. The results for the original method are shown in Table 4.3, and the results for the new method are shown in Table 4.5.

The results shown in Table 4.5 are for $N = 500$. We were able to classify with high accuracy using a subset of 21 genes, 20 of these genes are already known to be related to prostate cancer directly, or to another type of cancer. Increasing the value of N increased the accuracy for the '347 vs rest', but the number of genes related to cancer went down to the original number that was discovered in the first study.

The accuracy for the rest of the Gleason scores did not increase as it is already very high. We can notice in Table 3 how the accuracy of the class '347 vs rest' increased while the number of genes related to cancer becomes smaller.

Table 4.1: Results for running the laterality study.

	Total/Average	347 vs rest	437 vs rest	336 vs rest	448 vs 538
Accuracy	98	94	98	100	100
Number of Genes	26	15	7	3	1
Cancer related	7	3	2	1	1

Table 4.2: Results for running the laterality study with the new proposed method

	Total/Average	347 vs rest	437 vs rest	336 vs rest	448 vs 538
Accuracy	98	94	98	100	100
Number of genes	26	15	7	3	1
Cancer related	7	3	2	1	1

Table 4.3: Results for running the previous method.

	Total/Average	347 vs rest	437 vs rest	336 vs rest	448 vs 538
Accuracy	98	94	98	100	100
Number of genes	26	15	7	3	1
Cancer related	7	3	2	1	1

4.4 Conclusion

Using feature selection is an important step in any machine learning problem that has a large number of attributes. In the field of biology and especially studying gene expression,

Table 4.4: Results for running the proposed method.

	Total/Average	347 vs rest	437 vs rest	336 vs rest	448 vs 538
Accuracy	95	83	98	100	100
Number of genes	21	7	8	4	2
Cancer related	20	7	8	3	2

Table 4.5: The effect of changing the value of (number of genes chosen) on the classification performance and the number of genes that are related to cancer.

N	Classification accuracy	Number of genes	Cancer related
500	83	7	7
1000	86	10	6
1500	90	10	5
2000	93	13	4
2500	94	15	3

the number of features is huge, and many of the features/genes that are eliminated during this step might be a gene that is proven to be related to cancer that is being studied or to another related cancer type. With the new method, genes that are proven in literature are priority and are part of the first phase of the wrapper-based feature selection step. In fact, they are part of each step after the selection. At the same time, if a gene is not at all expressed, or if it is not deferentially expressed, it is discarded at the first pre-processing step. The results for each iteration in the last step are stored, so that the researcher is able to choose which sub set has either a larger AUC or that the subset that has the highest number of genes related to cancer.

We are planning to incorporate deep learning into future versions of this method, either as an additional feature selection step, or as another classifier within the wrapper-base feature selection step.

Future versions of this method are expected to integrate deep learning techniques. This will be done either as a supplementary feature selection step or as an additional classifier within the wrapper-based feature selection step.

References

[1] Mardis, E. (2008). The impact of next-generation sequencing technology on genetics. (Vol. 24) Elsevier.

[2] DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl, C., Philippakis, A., Del Angel, G., Rivas, M., Hanna, M., & others (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43(5), 491.

[3] Meldrum, C., Doyle, M., & Tothill, R. (2011). Next-generation sequencing for cancer diagnostics: a practical perspective. The Clinical Biochemist Reviews, 32(4), 177.

[4] Libbrecht, M., & Noble, W. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321.

[5] Alkhateeb, A., Atikukke, G., Porter, L., Fifield, B. A., Cavallo-Medved, D., Facca, J., & Kanjeekal, S. M. (2020). Comprehensive targeted gene profiling to determine the genomic signature likely to drive progression of high-grade nonmuscle invasive bladder cancer to muscle invasive bladder cancer.

[6] Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3(Mar), 1157–1182.

[7] Kotsiantis, S., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging Artificial Intelligence Applications in Computer Engineering, 160, 3–24.

[8] Bradley, A. (1997). The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognition, 30(7), 1145–1159.

[9] Hamzeh, O., Alkhateeb, A., Rezaeian, I., Karkar, A., & Rueda, L. (2017). Finding transcripts associated with prostate cancer Gleason stages using next generation sequencing and machine learning techniques. In International Conference on Bioinformatics and Biomedical Engineering (pp. 337–348).

[10] Gordetsky, J.; Epstein, J. Grading of Prostatic Adenocarcinoma: Current State and Prognostic Implications. Diagn. Pathol. 2016, *11*, 25.

[11] Piñero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., & Furlong, L. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database, 2015.

[12] Davis, A., Grondin, C., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B., Wiegers, T., & Mattingly, C. (2014). The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic Acids Research, 43(D1), D914–D920.

[13] UniProt Consortium (2014). UniProt: a Hub for Protein Information. Nucleic Acids Research, 43(D1), D204–D212.

[14] Landrum, M., Lee, J., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., & others (2015). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic acids research, 44(D1), D862–D868.

[15] Rath, A., Olry, A., Dhombres, F., Brandt, M., Urbero, B., & Ayme, S. (2012). Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. Human Mutation, 33(5), 803–808.

[16] Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., & others (2013). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research, 42(D1), D1001–D1006.

[17] Shimoyama, M., De Pons, J., Hayman, G., Laulederkind, S., Liu, W., Nigam, R., Petri, V., Smith, J., Tutaj, M., Wang, S.J., & others (2014). The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. Nucleic Acids Research, 43(D1), D743–D750.

[18] Eppig, J., Blake, J., Bult, C., Kadin, J., Richardson, J., & Mouse Genome Database Group (2014). The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. Nucleic Acids Research, 43(D1), D726–D736.

[19] Becker, G., Barnes, C., Bright, J., & Wang, A. (2004). The genetic association database. Nature Genetics, 36(5), 431-432.

[20] Bundschus, M., Dejori, M., Stetter, M., Tresp, V., & Kriegel, H.P. (2008). Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics, 9(1), 207.

[21] Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., & Furlong, L. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research, 45(D1), D833-D839.

[22] Bravo, ., Piñero, J., Queralt-Rosinach, N., Rautschka, M., & Furlong, L. (2015). Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinformatics, 16(1), 55.

[23] Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., & Wain, H. (2001). The HUGO gene nomenclature committee (HGNC). Human Genetics, 109(6), 678–680.

[24] Aken, B., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., C., Hourlier, T., & others (2016). The Ensembl gene annotation system. Database, 2016.

[25] Novakovic, J. (2009). Using information gain attribute evaluation to classify sonar targets. In 17th Telecommunications Forum TELFOR (pp. 1351–1354).

[26] Benesty J., Chen J., Huang Y., Cohen I. (2009) Pearson Correlation Coefficient. In: Noise Reduction in Speech Processing. Springer Topics in Signal Processing, vol 2. Springer, Berlin, Heidelberg.

[27] Yu, S. (1997). Regular languages. In Handbook of formal languages (pp. 41-110). Springer, Berlin, Heidelberg.

[28] 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 Analysis and Machine Intelligence, 27(8), 1226-1238.

[29] Chang, C.C., & Lin, C.J. (2011). LIBSVM: A library for support vector machinesACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 27.

[30] Rish, I. (2001). An Empirical Study of the Naïve Bayes Classifier. IJCAI 2001 Work Empir Methods Artif Intell, 3.

[31] Pal, M. (2005). Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1), 217–222.

[32] Cunningham, P., & Delany, S. (2007). k-Nearest neighbour classifiers. Multiple Classifier Systems, 34(8), 1–17.

[33] Geer, L., Marchler-Bauer, A., Geer, R., Han, L., He, J., He, S., Liu, C., Shi, W., & Bryant, S. (2009). The NCBI biosystems database. Nucleic Acids Research, 38(suppl_1), D492–D496.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The main contribution of this thesis is to provide a generic pipeline for modeling RNA-Seq data as a supervised learning scheme used to obtain meaningful biomarkers in cancer. While the main models were created for different cancer problems, the proposed models showed high performance and throughput. In Chapter 2, the proposed models we were able to extract transcriptomic biomarkers that can predict, with very high accuracy, certain Gleason groups for prostate cancer. Using the model described in Chapter 3, we were able to detect prostate cancer location based on the gene expressions provided. Additionally, we performed biological validation using the literature in collaboration with biologists to investigate the obtained biomarkers' significance. In Chapter 4, we proposed a machine learning framework that is capable of enhancing conventional models to identify and validate biomarkers. This framework enhances the feature selection method by utilizing knowledge extracted from Medline and other public resources. The method can be used in any kind of cancer and can also be used as an integrative multi-omics model that utilizes mutations, copy number alterations or any other clinical data that is available. The methods were able to provide a number of genes, which can be used to classify samples accurately, and these genes are already proven to be related to cancer by the latest literature.

The main methods proposed were able to handle different machine learning problems, such as class-imbalance and multi-class classification. In the work presented in Chapter 3, we faced the the class-imbalance problem, in which the number of samples from one class is much higher than the number of samples from the other class. We were able to solve this problem using a combination of machine learning techniques.

To summarize, the contributions of this thesis are listed below:

- ⇒ Developing a framework that enhances the feature selection method for a classification problem, and applying this method to enhance earlier work.
- ⇒ Proposing a machine learning pipeline that takes raw RNA-Seq data and provides a number of biomarkers that can classify prostate cancer Gleason groups.
- ⇒ Developing a machine learning pipeline that takes gene expressions and provides a model that detects the location of the prostate cancer.

⇒ Handling the multi-class problem using the one-versus-all approach for prostate cancer Gleason stages.

⇒ Proposing a generic pipeline that can be proved to work with any kind of cancer, as long as gene expressions or other types of data are used.

Even the though the proposed work is valuable, it does have some limitations. The framework proposed in Chapter 4 provides excellent results, but it is sensitive to the availability of information on the knowledge-base used. For example, if the gene name is not in the database, then that gene will have the same priority as the genes that are not related to cancer. Although this is an issue that is beyond our control, it is something that could be investigated further. The proposed framework takes a long time to run, especially when it creates the weights, though this task is performed only once at the beginning of the pipeline. To solve the class-imbalance problem faced in Chapter 3, we used machine learning techniques to create synthetic samples, which solves the class imbalance, while it introduces extra samples that are not in the original sample set.

Most parts of this work have been published in conferences in collaboration with my lab mates and other researchers from different disciplines, who have jointly co-authored these publications. Chapter 2 has been published, by invitation, in Diagnostics journal as part of the Special Issue on Next Generation Sequencing in Tumor Diagnosis and Treatment, in 2019. Chapter 2 was also presented at the 5th International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO 2017, Granada, Spain, April 26–28, 2017. Chapter 3 has been accepted in a special edition of BMC Bioinformatics journal, and is currently in press. This publication was the result of an invitation to submit an extended version after presenting our work at the 6th International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO 2018, Granada, Spain, April 25–27, 2018., which was presented by L. Rueda as a keynote talk. Chapter 4 of this thesis was presented at the Machine Learning Models for Multi-omics Data Integration, MODI 2019, a workshop held at the 10th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB), in the , Niagara Falls, New York, September 7-10, 2019., in which we received an invitation to submit an extended version to a special collection of Evolutionary Bioinformatics journal. The draft has been submitted already and is currently being reviewed.

5.2 Future work

Even though this work provides the scientific community with meaningful contributions, there is some room for improvement.

- In Chapter 2, the hierarchical model utilized one-versus-all approach, which can be enhanced by implementing other methods such as one-versus-rest and then a comparison can be done against the original results to provide a comprehensive solution to the multi-class problem.

- The method implemented in Chapter 2 can be utilized to study other kinds of cancer, as long as we utilize gene expressions or transcriptomic data.

- In Chapter 3, we utilized SMOTE to introduce synthetic samples to the minority class and we used NCL to remove samples from the majority class, a future enhancement would add boosting and bagging methods and then benchmark the two methods.

- In the framework introduced in Chapter 4 we used the DisGeNET database. However, other resources are available, accordingly, the framework can be enhanced by utilizing plugins. These plugins can be used to connect the framework to any source that can be utilized as a knowledge-base.

APPENDICES

Appendix A

Information about data sets used

A.1 List of aligning percentage from 104 samples in the Hierarchical method

Table A.1: The percentage of uniquely aligned mapped reads to the human genome for each sample of the 104 prostate cancer patients sample using the STAR genome alignment tool.

SRA Sample	Run name	Gleason score	Pstage	Uniquely mapped reads
SRS554892	SRR1164787	347	pT3	91.20%
SRS554893	SRR1164788	336	pT2	89.76%
SRS554894	SRR1164789	336	pT2	89.01%
SRS554895	SRR1164790	347	pT3	88.81%
SRS554896	SRR1164791	347	pT2	88.27%
SRS554897	SRR1164792	347	pT2	89.32%
SRS554898	SRR1164793	437	pT1C	86.10%
SRS554899	SRR1164794	347	pT2	94.36%
SRS554900	SRR1164795	448	pT2B	93.49%
SRS554901	SRR1164796	336	pT1C	94.33%
SRS554902	SRR1164797	347	pT1C	92.03%
SRS554903	SRR1164798	347	pT2C	94.88%
SRS554904	SRR1164799	347	pT2	92.58%
SRS554905	SRR1164800	437	pT2C	91.19%
SRS554906	SRR1164801	538	pT3A	93.76%
SRS554907	SRR1164802	347	pT2A	93.75%
SRS554908	SRR1164803	347	pT3A	93.33%
SRS554909	SRR1164804	347	pT2C	93.91%

Continued on next page –

– continued from previous page

SRA Sample	Run name	Gleason score	Pstage	Uniquely mapped reads
SRS554910	SRR1164805	336	pT2C	92.92%
SRS554911	SRR1164806	347	pT2C	92.77%
SRS554912	SRR1164807	437	pT2C	92.28%
SRS554913	SRR1164808	448	pT2C	91.58%
SRS554914	SRR1164809	347	pT2C	95.75%
SRS554915	SRR1164810	347	pT2A	93.42%
SRS554916	SRR1164811	347	pT2C	96.01%
SRS554917	SRR1164812	347	pT2	90.93%
SRS554918	SRR1164813	347	pT2	95.03%
SRS554919	SRR1164814	347	pT2A	95.59%
SRS554920	SRR1164815	437	pT2C	94.04%
SRS554921	SRR1164816	347	pT2C	92.04%
SRS554922	SRR1164817	347	pT3A	92.72%
SRS554923	SRR1164818	347	pT3A	92.92%
SRS554924	SRR1164819	347	pT2C	93.50%
SRS554925	SRR1164820	347	pT2C	91.91%
SRS554926	SRR1164821	336	pT2A	89.54%
SRS554927	SRR1164822	347	pT2A	92.63%
SRS554928	SRR1164823	437	pT2B	93.39%

Continued on next page –

– continued from previous page

SRA Sample	Run name	Gleason score	Pstage	Uniquely mapped reads
SRS554929	SRR1164824	437	pT2C	93.42%
SRS554930	SRR1164825	347	pT2C	90.71%
SRS554931	SRR1164826	437	pT2A	93.48%
SRS554932	SRR1164827	347	pT2C	88.83%
SRS554933	SRR1164828	347	pT2A	94.77%
SRS554934	SRR1164829	347	pT2C	94.67%
SRS554935	SRR1164830	336	pT2A	95.70%
SRS554936	SRR1164831	347	pT2A	95.32%
SRS554937	SRR1164832	347	pT2B	93.33%
SRS554938	SRR1164833	347	pT2C	91.23%
SRS554939	SRR1164834	437	pT2	94.51%
SRS554940	SRR1164835	347	pT2C	93.33%
SRS554941	SRR1164836	347	pT2A	95.19%
SRS554942	SRR1164837	336	pT2A	94.06%
SRS554943	SRR1164838	336	pT1C	88.71%
SRS554944	SRR1164839	336	pT1C	91.35%
SRS554945	SRR1164840	325	NA	89.48%
SRS554946	SRR1164841	549	pT1C	85.01%
SRS554947	SRR1164842	549	pT1C	87.16%

Continued on next page –

– continued from previous page

SRA Sample	Run name	Gleason score	Pstage	Uniquely mapped reads
SRS554948	SRR1164843	347	pT2B	84.12%
SRS554949	SRR1164844	347	pT2B	85.67%
SRS554950	SRR1164845	437	pT2A	82.92%
SRS554951	SRR1164846	437	pT2A	88.31%
SRS554952	SRR1164847	437	pT1C	83.87%
SRS554953	SRR1164848	437	pT1C	87.38%
SRS554954	SRR1164849	347	pT2A	83.11%
SRS554955	SRR1164850	347	pT2A	85.12%
SRS554956	SRR1164851	347	pT2B	91.79%
SRS554957	SRR1164852	347	pT2B	92.70%
SRS554958	SRR1164853	347	pT1C	89.31%
SRS554959	SRR1164854	347	pT2B	92.03%
SRS554960	SRR1164855	336	pT1C	91.27%
SRS554961	SRR1164856	347	pT2A	81.47%
SRS554962	SRR1164857	459	pT2B	87.48%
SRS554963	SRR1164858	448	pT3B	84.01%
SRS554964	SRR1164859	437	pT2A	88.89%
SRS554965	SRR1164860	549	pT2B	94.46%
SRS554966	SRR1164861	347	pT1C	88.21%

Continued on next page –

– continued from previous page

SRA Sample	Run name	Gleason score	Pstage	Uniquely mapped reads
SRS554967	SRR1164862	347	pT2A	89.46%
SRS554968	SRR1164863	437	pT3B	88.21%
SRS554969	SRR1164864	437	pT3B	93.94%
SRS554970	SRR1164865	437	pT2A	94.90%
SRS554971	SRR1164866	347	pT2A	85.55%
SRS554972	SRR1164867	448	pT3B	93.78%
SRS554973	SRR1164868	347	pT2A	91.84%
SRS554974	SRR1164869	347	pT1C	92.50%
SRS554975	SRR1164870	437	pT2A	92.78%
SRS554976	SRR1164871	437	pT1C	90.55%
SRS554977	SRR1164872	437	pT3B	95.48%
SRS554978	SRR1164873	347	pT2C	89.17%
SRS554979	SRR1164874	437	pT3A	89.97%
SRS554980	SRR1164875	437	pT2C	94.34%
SRS554981	SRR1164876	448	pT3B	88.91%
SRS554982	SRR1164877	459	pT3B	92.32%
SRS554983	SRR1164878	448	pT3B	94.09%
SRS554984	SRR1164879	347	pT3B	95.13%
SRS554985	SRR1164880	448	pT2C	93.98%

Continued on next page –

– continued from previous page

SRA Sample	Run name	Gleason score	Pstage	Uniquely mapped reads
SRS554986	SRR1164881	347	pT2C	91.91%
SRS554987	SRR1164882	437	pT2	94.01%
SRS554988	SRR1164883	448	pT2C	95.41%
SRS554989	SRR1164884	347	pT2C	91.62%
SRS554990	SRR1164885	347	pT2C	95.20%
SRS554991	SRR1164886	347	pT2C	90.21%
SRS554992	SRR1164887	347	pT4	91.75%
SRS554993	SRR1164888	437	pT2C	89.09%
SRS554994	SRR1164889	437	pT2C	87.34%
SRS554995	SRR1164890	347	pT2C	86.01%
SRS554996	SRR1164891	448	pT2A	90.54%
SRS554997	SRR1164892	347	pT3A	94.43%

A.2 List of samples used in the Laterality method

Table A.2: The location of tumor and the Gleason scores for the 499 prostate cancer patients samples.

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-2A-A8VO	4	5	9	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-2A-A8VT	3	3	6	Bilateral
TCGA-2A-A8VV	4	4	8	Right
TCGA-2A-A8VX	4	3	7	Bilateral
TCGA-2A-A8W1	4	5	9	Left
TCGA-2A-A8W3	3	4	7	Left
TCGA-2A-AAYF	3	3	6	Right
TCGA-2A-AAYO	3	3	6	Bilateral
TCGA-2A-AAYU	4	4	8	Left
TCGA-4L-AA1F	4	3	7	Bilateral
TCGA-CH-5737	3	3	6	Bilateral
TCGA-CH-5738	3	4	7	Bilateral
TCGA-CH-5739	3	4	7	Left
TCGA-CH-5740	4	5	9	Bilateral
TCGA-CH-5741	3	4	7	Bilateral
TCGA-CH-5743	4	3	7	Bilateral
TCGA-CH-5744	3	4	7	Bilateral
TCGA-CH-5745	3	4	7	[Not Available]
TCGA-CH-5746	4	3	7	Bilateral
TCGA-CH-5748	3	4	7	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-CH-5750	5	5	10	Bilateral
TCGA-CH-5751	4	4	8	Bilateral
TCGA-CH-5752	4	5	9	Bilateral
TCGA-CH-5753	4	5	9	Bilateral
TCGA-CH-5754	5	4	9	Bilateral
TCGA-CH-5761	4	3	7	Left
TCGA-CH-5762	3	4	7	Bilateral
TCGA-CH-5763	4	3	7	Bilateral
TCGA-CH-5764	3	4	7	Bilateral
TCGA-CH-5765	4	3	7	Bilateral
TCGA-CH-5766	4	3	7	Bilateral
TCGA-CH-5767	2	4	6	Bilateral
TCGA-CH-5768	5	4	9	Bilateral
TCGA-CH-5769	4	3	7	Bilateral
TCGA-CH-5771	4	5	9	Bilateral
TCGA-CH-5772	4	3	7	Bilateral
TCGA-CH-5788	3	4	7	Bilateral
TCGA-CH-5789	3	4	7	Bilateral
TCGA-CH-5790	3	4	7	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-CH-5791	4	5	9	Bilateral
TCGA-CH-5792	3	4	7	Bilateral
TCGA-CH-5794	4	3	7	Bilateral
TCGA-EJ-5494	5	3	8	Bilateral
TCGA-EJ-5495	3	4	7	Bilateral
TCGA-EJ-5496	3	4	7	Bilateral
TCGA-EJ-5497	4	3	7	Bilateral
TCGA-EJ-5498	4	3	7	Bilateral
TCGA-EJ-5499	4	3	7	Bilateral
TCGA-EJ-5501	4	3	7	Bilateral
TCGA-EJ-5502	5	3	8	Bilateral
TCGA-EJ-5503	4	3	7	Bilateral
TCGA-EJ-5504	4	3	7	Bilateral
TCGA-EJ-5505	5	3	8	Bilateral
TCGA-EJ-5506	4	5	9	Bilateral
TCGA-EJ-5507	3	4	7	Bilateral
TCGA-EJ-5508	3	4	7	Bilateral
TCGA-EJ-5509	4	3	7	Bilateral
TCGA-EJ-5510	3	4	7	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-EJ-5511	3	4	7	Bilateral
TCGA-EJ-5512	4	5	9	Bilateral
TCGA-EJ-5514	3	4	7	Bilateral
TCGA-EJ-5515	3	4	7	Right
TCGA-EJ-5516	3	3	6	Bilateral
TCGA-EJ-5517	4	5	9	Bilateral
TCGA-EJ-5518	4	4	8	Bilateral
TCGA-EJ-5519	3	4	7	Bilateral
TCGA-EJ-5521	3	4	7	Right
TCGA-EJ-5522	4	5	9	Bilateral
TCGA-EJ-5524	4	5	9	Bilateral
TCGA-EJ-5525	4	4	8	Bilateral
TCGA-EJ-5526	4	3	7	Bilateral
TCGA-EJ-5527	3	4	7	Bilateral
TCGA-EJ-5530	4	3	7	Bilateral
TCGA-EJ-5531	3	4	7	Bilateral
TCGA-EJ-5532	3	4	7	Bilateral
TCGA-EJ-5542	4	3	7	Bilateral
TCGA-EJ-7115	3	4	7	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-EJ-7123	3	4	7	Bilateral
TCGA-EJ-7125	3	4	7	Bilateral
TCGA-EJ-7218	4	3	7	Bilateral
TCGA-EJ-7312	4	3	7	Bilateral
TCGA-EJ-7314	4	3	7	Bilateral
TCGA-EJ-7315	3	4	7	Bilateral
TCGA-EJ-7317	4	3	7	Bilateral
TCGA-EJ-7318	3	3	6	Bilateral
TCGA-EJ-7321	3	4	7	Bilateral
TCGA-EJ-7325	3	4	7	Bilateral
TCGA-EJ-7327	4	3	7	Bilateral
TCGA-EJ-7328	4	3	7	Bilateral
TCGA-EJ-7330	3	4	7	Bilateral
TCGA-EJ-7331	3	4	7	Bilateral
TCGA-EJ-7781	4	4	8	Bilateral
TCGA-EJ-7782	4	3	7	Bilateral
TCGA-EJ-7783	4	3	7	Bilateral
TCGA-EJ-7784	3	4	7	Bilateral
TCGA-EJ-7785	3	4	7	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-EJ-7786	3	4	7	Bilateral
TCGA-EJ-7788	4	3	7	Bilateral
TCGA-EJ-7789	3	4	7	Bilateral
TCGA-EJ-7791	3	4	7	Bilateral
TCGA-EJ-7792	3	4	7	Bilateral
TCGA-EJ-7793	3	4	7	Bilateral
TCGA-EJ-7794	3	4	7	Bilateral
TCGA-EJ-7797	4	4	8	Bilateral
TCGA-EJ-8468	4	5	9	Left
TCGA-EJ-8469	4	3	7	Bilateral
TCGA-EJ-8470	4	4	8	Bilateral
TCGA-EJ-8472	4	4	8	[Not Available]
TCGA-EJ-8474	4	4	8	Bilateral
TCGA-EJ-A46B	3	5	8	Bilateral
TCGA-EJ-A46D	4	4	8	Bilateral
TCGA-EJ-A46E	4	4	8	Bilateral
TCGA-EJ-A46F	4	4	8	Bilateral
TCGA-EJ-A46G	3	4	7	Bilateral
TCGA-EJ-A46H	3	4	7	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-EJ-A46I	4	5	9	Bilateral
TCGA-EJ-A65B	4	4	8	[Not Available]
TCGA-EJ-A65D	4	3	7	Bilateral
TCGA-EJ-A65E	4	5	9	Bilateral
TCGA-EJ-A65F	3	5	8	Bilateral
TCGA-EJ-A65G	4	5	9	Bilateral
TCGA-EJ-A65J	3	3	6	Bilateral
TCGA-EJ-A65M	4	4	8	Bilateral
TCGA-EJ-A6RA	4	3	7	Bilateral
TCGA-EJ-A6RC	3	4	7	Bilateral
TCGA-EJ-A7NF	3	4	7	Bilateral
TCGA-EJ-A7NG	3	4	7	Bilateral
TCGA-EJ-A7NH	4	4	8	Bilateral
TCGA-EJ-A7NJ	3	4	7	Bilateral
TCGA-EJ-A7NK	4	5	9	Bilateral
TCGA-EJ-A7NM	4	3	7	Bilateral
TCGA-EJ-A7NN	3	4	7	Bilateral
TCGA-EJ-A8FN	4	3	7	Bilateral
TCGA-EJ-A8FO	3	5	8	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-EJ-A8FP	3	4	7	Bilateral
TCGA-EJ-A8FS	5	3	8	Bilateral
TCGA-EJ-A8FU	3	3	6	Bilateral
TCGA-EJ-AB20	3	3	6	Bilateral
TCGA-EJ-AB27	4	3	7	Bilateral
TCGA-FC-7708	4	5	9	Bilateral
TCGA-FC-7961	5	3	8	Bilateral
TCGA-FC-A4JI	4	5	9	Bilateral
TCGA-FC-A5OB	4	3	7	Bilateral
TCGA-FC-A66V	4	3	7	Bilateral
TCGA-FC-A6HD	3	3	6	Bilateral
TCGA-FC-A8O0	3	4	7	Bilateral
TCGA-G9-6329	4	3	7	Bilateral
TCGA-G9-6332	4	3	7	Bilateral
TCGA-G9-6333	3	4	7	Bilateral
TCGA-G9-6336	4	3	7	Bilateral
TCGA-G9-6338	3	4	7	Bilateral
TCGA-G9-6339	3	3	6	Bilateral
TCGA-G9-6342	4	3	7	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-G9-6343	3	3	6	Bilateral
TCGA-G9-6347	3	4	7	Bilateral
TCGA-G9-6348	3	4	7	Bilateral
TCGA-G9-6351	3	4	7	Bilateral
TCGA-G9-6353	3	4	7	Bilateral
TCGA-G9-6354	4	5	9	Right
TCGA-G9-6356	3	4	7	Bilateral
TCGA-G9-6361	4	3	7	Bilateral
TCGA-G9-6362	4	3	7	Bilateral
TCGA-G9-6363	3	4	7	Bilateral
TCGA-G9-6364	3	4	7	Bilateral
TCGA-G9-6365	4	3	7	Bilateral
TCGA-G9-6366	4	5	9	Left
TCGA-G9-6367	4	3	7	Bilateral
TCGA-G9-6369	3	4	7	Bilateral
TCGA-G9-6370	3	3	6	Bilateral
TCGA-G9-6371	4	3	7	Bilateral
TCGA-G9-6373	3	4	7	Bilateral
TCGA-G9-6377	3	4	7	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-G9-6378	4	3	7	Bilateral
TCGA-G9-6379	3	4	7	Bilateral
TCGA-G9-6384	3	4	7	Bilateral
TCGA-G9-6385	4	3	7	Bilateral
TCGA-G9-6494	4	3	7	Bilateral
TCGA-G9-6496	3	4	7	Bilateral
TCGA-G9-6498	4	5	9	Bilateral
TCGA-G9-6499	3	3	6	Bilateral
TCGA-G9-7509	4	4	8	Bilateral
TCGA-G9-7510	3	4	7	Bilateral
TCGA-G9-7519	4	4	8	Bilateral
TCGA-G9-7521	3	4	7	Bilateral
TCGA-G9-7522	5	5	10	Bilateral
TCGA-G9-7523	4	3	7	Bilateral
TCGA-G9-7525	4	4	8	Bilateral
TCGA-G9-A9S0	4	4	8	Bilateral
TCGA-G9-A9S4	4	4	8	Bilateral
TCGA-G9-A9S7	3	4	7	Bilateral
TCGA-H9-7775	3	3	6	Right

Continued on next page –

– continued from previous page

Sample name	**Gleason Prim.**	**Gleason Sec.**	**Gleason Score**	**Laterality**
TCGA-H9-A6BX	3	4	7	Bilateral
TCGA-H9-A6BY	3	3	6	Bilateral
TCGA-HC-7075	3	3	6	Right
TCGA-HC-7077	3	4	7	Bilateral
TCGA-HC-7078	3	4	7	Bilateral
TCGA-HC-7079	3	4	7	Bilateral
TCGA-HC-7080	4	5	9	Right
TCGA-HC-7081	3	3	6	Bilateral
TCGA-HC-7209	3	4	7	Bilateral
TCGA-HC-7210	3	4	7	Right
TCGA-HC-7211	3	4	7	Bilateral
TCGA-HC-7212	4	5	9	Right
TCGA-HC-7213	3	4	7	Bilateral
TCGA-HC-7230	3	4	7	[Not Available]
TCGA-HC-7231	4	5	9	Bilateral
TCGA-HC-7232	4	3	7	Bilateral
TCGA-HC-7233	3	4	7	Left
TCGA-HC-7736	3	4	7	Bilateral
TCGA-HC-7737	3	4	7	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-HC-7738	3	4	7	Bilateral
TCGA-HC-7740	3	4	7	Bilateral
TCGA-HC-7742	4	3	7	Bilateral
TCGA-HC-7744	4	3	7	Bilateral
TCGA-HC-7745	3	4	7	Bilateral
TCGA-HC-7747	3	3	6	Right
TCGA-HC-7748	4	3	7	Right
TCGA-HC-7749	3	4	7	Left
TCGA-HC-7750	3	4	7	Bilateral
TCGA-HC-7752	4	3	7	Bilateral
TCGA-HC-7817	3	4	7	Bilateral
TCGA-HC-7818	4	4	8	Bilateral
TCGA-HC-7819	3	4	7	Bilateral
TCGA-HC-7820	4	4	8	Bilateral
TCGA-HC-7821	3	3	6	Bilateral
TCGA-HC-8213	4	3	7	Bilateral
TCGA-HC-8216	3	4	7	Right
TCGA-HC-8256	3	4	7	Bilateral
TCGA-HC-8257	3	3	6	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-HC-8258	3	3	6	Bilateral
TCGA-HC-8259	3	4	7	Left
TCGA-HC-8260	3	4	7	Bilateral
TCGA-HC-8261	4	4	8	Bilateral
TCGA-HC-8262	4	5	9	Bilateral
TCGA-HC-8264	3	5	8	Bilateral
TCGA-HC-8265	4	4	8	Bilateral
TCGA-HC-8266	4	4	8	Bilateral
TCGA-HC-A48F	5	4	9	Bilateral
TCGA-HC-A4ZV	4	5	9	Bilateral
TCGA-HC-A631	4	5	9	Bilateral
TCGA-HC-A632	3	4	7	Bilateral
TCGA-HC-A6AL	4	3	7	Bilateral
TCGA-HC-A6AN	4	3	7	Bilateral
TCGA-HC-A6AO	3	4	7	Bilateral
TCGA-HC-A6AP	3	4	7	Bilateral
TCGA-HC-A6AQ	3	4	7	Bilateral
TCGA-HC-A6AS	4	3	7	Right
TCGA-HC-A6HX	3	4	7	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-HC-A6HY	3	4	7	Bilateral
TCGA-HC-A76W	3	4	7	Bilateral
TCGA-HC-A76X	5	4	9	Bilateral
TCGA-HC-A8CY	3	4	7	Bilateral
TCGA-HC-A8D0	3	4	7	Left
TCGA-HC-A8D1	5	4	9	Bilateral
TCGA-HC-A9TE	5	4	9	Bilateral
TCGA-HC-A9TH	4	5	9	Bilateral
TCGA-HI-7168	3	4	7	Bilateral
TCGA-HI-7169	3	3	6	Bilateral
TCGA-HI-7170	5	4	9	Bilateral
TCGA-HI-7171	4	3	7	Bilateral
TCGA-J4-8198	3	4	7	Bilateral
TCGA-J4-8200	4	3	7	Bilateral
TCGA-J4-A67K	3	4	7	Bilateral
TCGA-J4-A67L	4	3	7	Left
TCGA-J4-A67M	3	4	7	Bilateral
TCGA-J4-A67N	3	4	7	Bilateral
TCGA-J4-A67O	3	3	6	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-J4-A67Q	4	3	7	Bilateral
TCGA-J4-A67R	3	4	7	Bilateral
TCGA-J4-A67S	3	4	7	Bilateral
TCGA-J4-A67T	4	4	8	Bilateral
TCGA-J4-A6G1	4	4	8	Bilateral
TCGA-J4-A6G3	3	4	7	Bilateral
TCGA-J4-A6M7	4	3	7	Bilateral
TCGA-J4-A83I	3	4	7	Bilateral
TCGA-J4-A83J	3	3	6	Bilateral
TCGA-J4-A83K	3	4	7	Bilateral
TCGA-J4-A83L	4	3	7	Bilateral
TCGA-J4-A83M	3	4	7	Bilateral
TCGA-J4-A83N	3	3	6	Bilateral
TCGA-J4-AATV	4	5	9	Bilateral
TCGA-J4-AATZ	3	3	6	Bilateral
TCGA-J4-AAU2	5	4	9	Bilateral
TCGA-J9-A52B	4	5	9	Bilateral
TCGA-J9-A52C	4	5	9	Bilateral
TCGA-J9-A52D	4	5	9	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-J9-A52E	4	5	9	Bilateral
TCGA-J9-A8CK	4	5	9	Bilateral
TCGA-J9-A8CL	5	4	9	Bilateral
TCGA-J9-A8CM	3	3	6	Bilateral
TCGA-J9-A8CN	4	3	7	Bilateral
TCGA-J9-A8CP	3	4	7	Bilateral
TCGA-KC-A4BL	3	4	7	Bilateral
TCGA-KC-A4BN	4	5	9	Right
TCGA-KC-A4BR	4	5	9	Bilateral
TCGA-KC-A4BV	3	4	7	Bilateral
TCGA-KC-A7F3	4	3	7	Bilateral
TCGA-KC-A7F5	3	4	7	Right
TCGA-KC-A7F6	3	4	7	Bilateral
TCGA-KC-A7FA	3	4	7	Bilateral
TCGA-KC-A7FD	3	4	7	Bilateral
TCGA-KC-A7FE	4	5	9	Bilateral
TCGA-KK-A59V	5	4	9	Bilateral
TCGA-KK-A59X	4	5	9	Bilateral
TCGA-KK-A59Y	4	3	7	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-KK-A59Z	4	5	9	Bilateral
TCGA-KK-A5A1	3	4	7	Bilateral
TCGA-KK-A6DY	4	5	9	Bilateral
TCGA-KK-A6E0	4	5	9	Right
TCGA-KK-A6E1	3	4	7	Bilateral
TCGA-KK-A6E2	3	4	7	Bilateral
TCGA-KK-A6E3	3	4	7	Bilateral
TCGA-KK-A6E4	4	3	7	Right
TCGA-KK-A6E5	4	5	9	Bilateral
TCGA-KK-A6E6	5	4	9	Bilateral
TCGA-KK-A6E7	4	5	9	Bilateral
TCGA-KK-A6E8	4	5	9	Bilateral
TCGA-KK-A7AP	4	3	7	Right
TCGA-KK-A7AQ	4	5	9	Bilateral
TCGA-KK-A7AU	3	4	7	Bilateral
TCGA-KK-A7AV	4	3	7	Bilateral
TCGA-KK-A7AW	4	3	7	Right
TCGA-KK-A7AY	4	3	7	Bilateral
TCGA-KK-A7AZ	4	5	9	Right

Continued on next page –

– continued from previous page

Sample name	**Gleason Prim.**	**Gleason Sec.**	**Gleason Score**	**Laterality**
TCGA-KK-A7B0	4	3	7	Bilateral
TCGA-KK-A7B1	4	5	9	Bilateral
TCGA-KK-A7B2	4	5	9	Left
TCGA-KK-A7B3	4	5	9	Bilateral
TCGA-KK-A7B4	4	3	7	Bilateral
TCGA-KK-A8I4	3	4	7	Left
TCGA-KK-A8I5	4	3	7	Right
TCGA-KK-A8I6	4	5	9	Bilateral
TCGA-KK-A8I7	4	5	9	Bilateral
TCGA-KK-A8I8	4	4	8	Bilateral
TCGA-KK-A8I9	4	5	9	Bilateral
TCGA-KK-A8IA	4	5	9	Bilateral
TCGA-KK-A8IB	4	5	9	Bilateral
TCGA-KK-A8IC	5	4	9	Bilateral
TCGA-KK-A8ID	4	3	7	Bilateral
TCGA-KK-A8IF	4	3	7	Bilateral
TCGA-KK-A8IG	4	3	7	Bilateral
TCGA-KK-A8IH	4	5	9	Bilateral
TCGA-KK-A8II	4	3	7	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-KK-A8IJ	4	5	9	Bilateral
TCGA-KK-A8IK	4	5	9	Bilateral
TCGA-KK-A8IL	4	3	7	Bilateral
TCGA-KK-A8IM	3	4	7	Bilateral
TCGA-M7-A71Y	4	3	7	Bilateral
TCGA-M7-A71Z	3	3	6	Right
TCGA-M7-A720	3	4	7	Bilateral
TCGA-M7-A721	4	4	8	Bilateral
TCGA-M7-A722	4	3	7	Bilateral
TCGA-M7-A723	4	4	8	Bilateral
TCGA-M7-A724	4	3	7	Right
TCGA-M7-A725	4	5	9	Bilateral
TCGA-MG-AAMC	3	4	7	[Not Available]
TCGA-QU-A6IL	3	4	7	Bilateral
TCGA-QU-A6IM	3	4	7	Left
TCGA-QU-A6IN	3	3	6	Bilateral
TCGA-QU-A6IO	3	3	6	Bilateral
TCGA-QU-A6IP	5	3	8	Right
TCGA-SU-A7E7	3	4	7	Right

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-TK-A8OK	4	3	7	Right
TCGA-TP-A8TT	4	3	7	Left
TCGA-TP-A8TV	3	3	6	Bilateral
TCGA-V1-A8MF	3	4	7	Bilateral
TCGA-V1-A8MG	3	3	6	Bilateral
TCGA-V1-A8MK	3	4	7	Bilateral
TCGA-V1-A8ML	4	3	7	Bilateral
TCGA-V1-A8MM	3	4	7	Bilateral
TCGA-V1-A8MU	3	4	7	Bilateral
TCGA-V1-A8WL	3	3	6	Bilateral
TCGA-V1-A8WN	3	3	6	Bilateral
TCGA-V1-A8WS	4	5	9	Bilateral
TCGA-V1-A8WV	4	5	9	Bilateral
TCGA-V1-A8WW	3	4	7	Bilateral
TCGA-V1-A8X3	5	4	9	Bilateral
TCGA-V1-A9O5	5	4	9	Bilateral
TCGA-V1-A9O7	3	5	8	Bilateral
TCGA-V1-A9O9	4	5	9	Bilateral
TCGA-V1-A9OA	3	3	6	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-V1-A9OF	4	4	8	Right
TCGA-V1-A9OH	5	4	9	Bilateral
TCGA-V1-A9OL	3	3	6	Bilateral
TCGA-V1-A9OQ	3	3	6	Bilateral
TCGA-V1-A9OT	3	5	8	Bilateral
TCGA-V1-A9OX	4	3	7	Bilateral
TCGA-V1-A9OY	5	4	9	Bilateral
TCGA-V1-A9Z7	5	4	9	Bilateral
TCGA-V1-A9Z8	4	5	9	Bilateral
TCGA-V1-A9Z9	4	5	9	Bilateral
TCGA-V1-A9ZG	4	5	9	Bilateral
TCGA-V1-A9ZI	4	4	8	Left
TCGA-V1-A9ZK	4	4	8	Bilateral
TCGA-V1-A9ZR	4	4	8	Bilateral
TCGA-VN-A88I	4	4	8	Bilateral
TCGA-VN-A88K	3	4	7	Bilateral
TCGA-VN-A88L	3	4	7	Bilateral
TCGA-VN-A88M	4	3	7	Bilateral
TCGA-VN-A88N	3	4	7	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-VN-A88O	3	4	7	Bilateral
TCGA-VN-A88P	4	4	8	Bilateral
TCGA-VN-A88Q	4	4	8	Bilateral
TCGA-VN-A88R	4	4	8	Bilateral
TCGA-VN-A943	5	3	8	Bilateral
TCGA-VP-A872	3	4	7	Bilateral
TCGA-VP-A875	3	5	8	Bilateral
TCGA-VP-A876	4	5	9	Bilateral
TCGA-VP-A878	4	5	9	Right
TCGA-VP-A879	4	4	8	Right
TCGA-VP-A87B	4	3	7	Bilateral
TCGA-VP-A87C	4	5	9	Bilateral
TCGA-VP-A87D	3	3	6	Bilateral
TCGA-VP-A87E	4	5	9	Bilateral
TCGA-VP-A87H	3	4	7	Bilateral
TCGA-VP-A87J	4	4	8	Bilateral
TCGA-VP-A87K	5	4	9	Bilateral
TCGA-VP-AA1N	4	4	8	Bilateral
TCGA-WW-A8ZI	4	5	9	Right

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-X4-A8KQ	3	4	7	Bilateral
TCGA-X4-A8KS	4	3	7	Bilateral
TCGA-XA-A8JR	3	4	7	Right
TCGA-XJ-A83F	4	3	7	Bilateral
TCGA-XJ-A83G	3	4	7	Bilateral
TCGA-XJ-A83H	5	4	9	[Not Available]
TCGA-XJ-A9DI	4	4	8	Right
TCGA-XJ-A9DK	3	3	6	Right
TCGA-XJ-A9DQ	5	4	9	Bilateral
TCGA-XJ-A9DX	4	4	8	Bilateral
TCGA-XK-AAIR	5	5	10	Bilateral
TCGA-XK-AAIV	5	4	9	Bilateral
TCGA-XK-AAIW	4	4	8	Bilateral
TCGA-XK-AAJ3	4	3	7	Bilateral
TCGA-XK-AAJA	4	3	7	Left
TCGA-XK-AAJP	4	3	7	Bilateral
TCGA-XK-AAJR	4	3	7	Bilateral
TCGA-XK-AAJT	4	3	7	Bilateral
TCGA-XK-AAJU	4	3	7	Bilateral

Continued on next page –

– continued from previous page

Sample name	**Gleason Prim.**	**Gleason Sec.**	**Gleason Score**	**Laterality**
TCGA-XK-AAK1	5	5	10	Bilateral
TCGA-XQ-A8TA	4	5	9	Bilateral
TCGA-XQ-A8TB	3	3	6	Right
TCGA-Y6-A8TL	4	4	8	Bilateral
TCGA-Y6-A9XI	5	4	9	Right
TCGA-YJ-A8SW	4	5	9	Bilateral
TCGA-YL-A8HJ	4	5	9	Bilateral
TCGA-YL-A8HK	4	5	9	Bilateral
TCGA-YL-A8HL	4	5	9	Bilateral
TCGA-YL-A8HM	4	3	7	Bilateral
TCGA-YL-A8HO	4	5	9	Bilateral
TCGA-YL-A8S8	4	5	9	Bilateral
TCGA-YL-A8S9	4	4	8	Bilateral
TCGA-YL-A8SA	4	5	9	Bilateral
TCGA-YL-A8SB	4	5	9	Bilateral
TCGA-YL-A8SC	4	3	7	Bilateral
TCGA-YL-A8SH	4	5	9	Bilateral
TCGA-YL-A8SI	4	5	9	Bilateral
TCGA-YL-A8SJ	4	5	9	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-YL-A8SK	4	4	8	Bilateral
TCGA-YL-A8SL	4	5	9	Bilateral
TCGA-YL-A8SO	4	5	9	Bilateral
TCGA-YL-A8SP	4	5	9	Bilateral
TCGA-YL-A8SQ	4	5	9	Bilateral
TCGA-YL-A8SR	4	5	9	Bilateral
TCGA-YL-A9WH	4	5	9	Bilateral
TCGA-YL-A9WI	4	4	8	Bilateral
TCGA-YL-A9WJ	4	5	9	Bilateral
TCGA-YL-A9WK	4	5	9	Bilateral
TCGA-YL-A9WL	4	5	9	Bilateral
TCGA-YL-A9WX	5	4	9	Bilateral
TCGA-YL-A9WY	4	5	9	Bilateral
TCGA-ZG-A8QW	3	3	6	Bilateral
TCGA-ZG-A8QX	5	4	9	Bilateral
TCGA-ZG-A8QY	4	5	9	Bilateral
TCGA-ZG-A8QZ	5	4	9	Bilateral
TCGA-ZG-A9KY	5	4	9	Bilateral
TCGA-ZG-A9L0	4	5	9	Bilateral

Continued on next page –

– continued from previous page

Sample name	Gleason Prim.	Gleason Sec.	Gleason Score	Laterality
TCGA-ZG-A9L1	4	5	9	Bilateral
TCGA-ZG-A9L2	5	4	9	[Not Available]
TCGA-ZG-A9L4	5	4	9	Bilateral
TCGA-ZG-A9L5	5	4	9	Bilateral
TCGA-ZG-A9L6	4	5	9	Bilateral
TCGA-ZG-A9L9	5	4	9	Bilateral
TCGA-ZG-A9LB	4	5	9	Right
TCGA-ZG-A9LM	5	4	9	Right
TCGA-ZG-A9LN	5	4	9	Bilateral
TCGA-ZG-A9LS	5	4	9	Bilateral
TCGA-ZG-A9LU	4	5	9	Bilateral
TCGA-ZG-A9LY	4	5	9	Bilateral
TCGA-ZG-A9LZ	5	4	9	Bilateral
TCGA-ZG-A9M4	5	4	9	Bilateral
TCGA-ZG-A9MC	5	4	9	Bilateral
TCGA-ZG-A9N3	5	4	9	Bilateral
TCGA-ZG-A9ND	5	4	9	Bilateral
TCGA-ZG-A9NI	3	3	6	Bilateral

A.3 List of publications

- O. Hamzeh and L. Rueda. 2019. ”A Gene-disease-based Machine Learning Approach to Identify Prostate Cancer Biomarkers”. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB ’19). ACM, New York, NY, USA, 633-638.

  DOI: https://doi.org/10.1145/3307339.3343479.

- O. Hamzeh, A. Alkhateeb, L. Rueda., “Predicting Tumor-locations in Prostate Cancer Tissue Using Gene Expression”, 22st International Conference on Research in Computational Molecular Biology (RECOMB), Paris, France, 2018.

- O. Hamzeh, A. Alkhateeb, L. Rueda., “Prediction of Tumor Location in Prostate Cancer Tissue Using Gene Expression“, IEEE International Conference on Biomedical and Health Informatics (BHI), Las Vegas, USA, 2018.

- O. Hamzeh, A. Alkhateeb, L. Rueda. “Predicting Tumor Locations in Prostate Cancer Tissue Using Gene Expression”. 6th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2018), Granada, Spain, 2018.

- H. Ahmad, O. Hamzeh, A. Alkhateeb, L. Rueda., ”An open source machine learning tool for identifying biomarkers in next generation sequencing [version 1; not peer reviewed]”. F1000Research 2017, 6:2189 (poster) (doi: 10.7490/f1000research.1115180.1).

- O. Hamzeh, A. Alkhateeb, L. Rueda. , “Finding Biomarkers Associated with Prostate Cancer Gleason Stages using Next Generation Sequencing and Machine Learning

Techniques", Great Lakes Bioinformatics Conference, Michigan, USA, 2017.

- O. Hamzeh, A. Alkhateeb, L. Rueda. ,"Finding Transcripts Associated with Prostate Cancer Gleason Stages Using Next Generation Sequencing and Machine Learning Techniques." International Conference on Bioinformatics and Biomedical Engineering. Springer, Granada, Spain, 2017.

VITA AUCTORIS

Osama Hamzeh was born in 1974 in Latakia, Syrian Arab Republic. He received his bachelor's degree in computer science from Ajman University College, Ajman, United Arab Emirates in 1997, and his master's in computer science from the University of Sharjah, United Arab Emirates in 2016. His research interests include pattern recognition, machine learning, bioinformatics, data mining, cancer research and next generation sequencing data analysis.