12-31-2019

Cancer risk prediction with whole exome sequencing and machine learning

Abdulrhman Fahad M Aljouie
New Jersey Institute of Technology

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

CANCER RISK PREDICTION WITH WHOLE EXOME SEQUENCING AND MACHINE LEARNING

by
Abdulrhman Fahad M Aljouie

Accurate cancer risk and survival time prediction are important problems in personalized medicine, where disease diagnosis and prognosis are tuned to individuals based on their genetic material. Cancer risk prediction provides an informed decision about making regular screening that helps to detect disease at the early stage and therefore increases the probability of successful treatments. Cancer risk prediction is a challenging problem. Lifestyle, environment, family history, and genetic predisposition are some factors that influence the disease onset. Cancer risk prediction based on predisposing genetic variants has been studied extensively. Most studies have examined the predictive ability of variants in known mutated genes for specific cancers. However, previous studies have not explored the predictive ability of collective genomic variants from whole-exome sequencing data. It is crucial to train a model in one study and predict another related independent study to ensure that the predictive model generalizes to other datasets. Survival time prediction allows patients and physicians to evaluate the treatment feasibility and helps chart health treatment plans. Many studies have concluded that clinicians are inaccurate and often optimistic in predicting patients' survival time; therefore, the need increases for automated survival time prediction from genomic and medical imaging data.

For cancer risk prediction, this dissertation explores the effectiveness of ranking genomic variants in whole-exome sequencing data with univariate features selection

methods on the predictive capability of machine learning classifiers. The dissertation performs cross-study in chronic lymphocytic leukemia, glioma, and kidney cancers that show that the top-ranked variants achieve better accuracy than the whole genomic variants.

For survival time prediction, many studies have devised 3D convolutional neural networks (CNNs) to improve the accuracy of structural magnetic resonance imaging (MRI) volumes to classify glioma patients into survival categories. This dissertation proposes a new multi-path convolutional neural network with SNP and demographic features to predict glioblastoma survival groups with a one-year threshold that improves upon existing machine learning methods. The dissertation also proposes a multi-path neural network system to predict glioblastoma survival categories with a 14-year threshold from a heterogeneous combination of genomic variations, messenger ribonucleic acid (RNA) expressions, 3D post-contrast T1 MRI volumes, and 2D post-contrast T1 MRI modality scans that show the malignancy. In 10-fold cross-validation, the mean 10-fold accuracy of the proposed network with handpicked 2D MRI slices (that manifest the tumor), mRNA expressions, and SNPs slightly improves upon each data source individually.

CANCER RISK PREDICTION WITH WHOLE EXOME SEQUENCING AND MACHINE LEARNING

by
Abdulrhman Fahad M Aljouie

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

December 2019

.

APPROVAL PAGE

CANCER RISK PREDICTION WITH WHOLE EXOME SEQUENCING AND MACHINE LEARNING

Abdulrhman Fahad M Aljouie

Dr. Usman Roshan, Dissertation Advisor — Date
Associate Professor of Computer Science, NJIT

Dr. Ioannis Koutis, Committee Member — Date
Associate Professor of Computer Science, NJIT

Dr. Michael Schatz, Committee Member — Date
Bloomberg Distinguished Associate Professor of Computer Science and Biology, Johns Hopkins University

---

Dr. Zhi Wei, Committee Member — Date
Professor of Computer Science, NJIT

Dr. Chase Q. Wu, Committee Member — Date
Professor of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Abdulrhman Fahad M Aljouie

Degree: Doctor of Philosophy

Date: December 2019

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2019

- Master of Science in Bioinformatics,
  New Jersey Institute of Technology, Newark, NJ, 2015

- Bachelor of Computer,
  King Saud University, Riyadh, Saudi Arabia, 2007

Major: Computer Science

Presentations and Publications:

A. Aljouie and U. Roshan, "Multi-path convolutional neural network for glioblastoma survival group prediction with point mutations and demographic features," in *Workshop Machine Learning and Artificial Intelligence in Bioinformatics and Medical Informatics (MABM) at IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, Nov 2019. (Accepted).

A. Aljouie, M. Schatz, and U. Roshan, "Machine learning based prediction of gliomas with germline mutations obtained from whole exome sequences from TCGA and 1000 Genomes Project," in *Proc. The Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, Oct 2019. (Accepted).

A. Aljouie, L. Zhong, and U. Roshan, "Anchor selection for pairwise whole genome sequence alignment with the maximum scoring subsequence and GPUs," in *Proc. International Conference on Intelligent Biology and Medicine (ICIBM)*, Los Angeles, CA, Jun 2018.

A. Aljouie, N. Patel, and U. Roshan, "Cross-validation and cross-study validation of kidney cancer with machine learning and whole exome sequences from the National Cancer Institute," in *Proc. IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Saint Louis, MI, May 2018.

A. Aljouie, N. Patel, B. Jhadav, and U. Roshan, "Cross-validation and cross-study validation of chronic lymphocytic leukemia with exome sequences and machine learning," *International Journal of Data Mining and Bioinformatics*, 2016.

N. Patel, B. Jhadav, A. Aljouie, and U. Roshan, "Cross-validation and cross-study validation of chronic lymphocytic leukemia with exome sequences and machine learning," In *Proc. IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov 2015.

A. Aljouie and U. Roshan, "Prediction of continuous phenotypes in mouse, fly, and rice genome wide association studies with support vector regression SNPs and ridge regression classifier," In *Proc. 14th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2015.

A. Aljouie, M. Esfandiari, S. Ramakrishnan, and U. Roshan, "Chi8: a GPU program for detecting significant interacting SNPs with the chi-square 8-df test," *BMC Research Notes*, vol. 8, no. 436, 2015.

Dedicated to my family
إلى عائلتي

ACKNOWLEDGMENT

I want to thank my dissertation advisor, Dr. Usman Roshan, for his continuous support, mentorship, exceptional attention to detail in answering my questions, and for always been reachable in difficult times. I also want to thank Dr. Zhi Wei for serving in my PhD dissertation committee, for his guidance and support throughout my time at NJIT, and for his thorough explanation of Next Generation Sequencing when I studied BNFO 620 Genomic Data Analysis course with him. I am grateful to Dr. Michael Schatz, Dr. Chase Wu, and Dr. Ioannis Koutis, for serving in my PhD dissertation committee and their valuable comments and contributions to my research project.

I also want to thank the Saudi Arabian Cultural Mission and King Abdullah International Medical Research Center (KAIMRC) for funding my graduate studies.

I want to thank the Advance Research and Computing Services at NJIT, in particular, David Perel, Kevin Walsh, and Gedaliah Wolosh, who assisted me in the Kong computing cluster issues. I also thank my lab colleagues, Meiyan Xie, and Yunzhie Xue, for all the thoughtful discussions during my time at NJIT and their technical help.

I am always grateful to my parents, Fahad and Norah, for their encouragement and support at all times. Special thanks to my wife, Nahlah, who stand by my side during challenging times.

The results shown in this dissertation are in part based upon data generated by The Cancer Genome Atlas (TCGA) A Research Network: https://www.cancer.gov/tcga. I want to thank the specimen donors and the research groups for helping in creating this data source

TABLE OF CONTENTS

Chapter		Page
1	INTRODUCTION ……...……	1
	1.1 Cancer Risk Prediction ……	1
	1.2 Cancer Survival Time Estimation ……	3
	1.3 Dissertation Contribution and Outline ……	3
2	BACKGROUND ……..…….	5
	2.1 Polygenic Risk Score (PRS) ……..……	7
	2.2 Genome-Wide Association Studies (GWAS) ……..….	7
	2.3 SNP Array ……	9
	2.4 Next Generation Sequencing (NGS) Variants Discovery ……	11
3	CROSS-VALIDATION AND CROSS-STUDY VALIDATION OF CHRONIC LYMPHOCYTIC LEUKAEMIA WITH EXOME SEQUENCES AND MACHINE LEARNING ……..….	13
	3.1 Introduction ...……..……..	13
	3.2 Materials and Methods ...……..…….…...	16
	3.2.1 Whole Exome Sequencing Data and Human Genome Reference …	16
	3.2.2 Next Generation Sequencing Analysis Pipeline ……	17
	3.2.3 Machine Learning Pipeline ...……	19
	3.2.4 High Performance Computing ……...	22
	3.3 Results …..……	22
	3.3.1 Cross-Validation …….……..	22
	3.3.2 Comparison to Cross-Validation on GWAS …….	25
	3.3.3 Cross-Study Validation …….	26

TABLE OF CONTENTS
(Continued)

Chapter **Page**

3.3.4 Biological Significance of Top Tanked SNPs ……... 29

3.4 Discussion ….……... 31

3.5 Conclusion ….…….. 32

4 CROSS-VALIDATION AND CROSS-STUDY VALIDATION OF KIDNEY CANCER WITH MACHINE LEARNING AND WHOLE EXOME SEQUENCES FROM THE NATIONAL CANCER INSTITUTE …….. 33

4.1 Introduction …..…… 33

4.2 Methods …..…….. 34

4.2.1 Data …… 34

4.2.2 Quality Control for Determining SNPs ….…….. 35

4.2.3 SNP Encoding ……. 37

4.2.4 Cross-Validation and Machine Learning …… 37

4.2.5 Cross-Study Validation ….…….. 38

4.3 Results ….…… 38

4.3.1 Cross-Validation …….. 38

4.3.2 Cross-Study Validation …… 41

4.3.3 Ranking of Previously Known Kidney Cancer Genes ……. 39

4.4 Conclusion and Future Work…….. 39

5 MACHINE LEARNING BASED PREDICTION OF GLIOMAS WITH GERMLINE MUTATIONS OBTAINED FROM WHOLE EXOME SEQUENCES FROM TCGA AND 1000 GENOMES PROJECT …….. 44

5.1 Introduction …….. 44

TABLE OF CONTENTS
(Continued)

Chapter **Page**

5.2 Methods …..…….. 46

5.2.1 Data ……..……. 46

5.2.2 Joint Genotyping ……..……. 47

5.2.3 SNPs Encoding ……...……. 48

5.2.4 Missing Genotypes ……..……. 49

5.2.5 Variants Calling Quality Control ...……..……. 50

5.2.6 Soft Filtering ……..……..……. 53

5.2.7 Hard Filtering ...…..……..……. 53

5.2.8 Soft+Hard Filtering ……..……. 53

5.2.9 Feature Scaling …… 53

5.2.10 Chi-Squared Features Selection ……..……. 54

5.2.11 Classifiers ……..……..……. 54

5.2.12 Performance Metrics …… 55

5.3 Results …..……..…….. 57

5.3.1 Cross-Validation ……..…… 57

5.3.2 Cross-Study Validation …… 59

5.3.3 Cancer Significance of Top Ranked SNPs …….. 60

5.3.4 SNP rs10792053 Mapping Quality …….. 64

5.3.5 Alternate Allele Frequency of Top SNPs ……..…… 66

5.4 Conclusion ….…….. 67

TABLE OF CONTENTS
(Continued)

Chapter **Page**

6 CHALLENGES IN PREDICTING GLIOMA SURVIVAL TIME IN MULTI-MODAL DEEP NETWORKS …….. 68

6.1 Introduction ....…….. 68

6.2 Methods ……..…….. 70

6.2.1 Data ……..…… 70

6.2.2 Network Architecture and Training ……..…… 73

6.3 Methods ……..…….. 75

6.3.1 Combined Data with 3D Volumes ……..…….. 75

6.3.2 Combined Data with 2D Slices …...…….. 76

6.3.2 Combined Data with 2D Slices …...…….. 76

6.3.2 Combined Data with 2D Slices …...…….. 76

6.4 Discussion ….…….. 78

6.5 Conclusion ….…….. 78

7 MULTI-PATH CONVOLUTIONAL NEURAL NETWORK FOR GLIOBLASTOMA SURVIVAL GROUP PREDICTION WITH POINT MUTATIONS AND DEMOGRAPHIC FEATURES ……. 80

7.1 Introduction ....…….. 82

7.2 Methods ….…….. 78

7.2.1 Patients Cohort ……..…….. 82

7.2.2 SNPs Calling and Quality Control ....…….. 82

7.2.3 SNPs Encoding ……..…….. 84

7.2.4 Training and Test Sets ……. 84

TABLE OF CONTENTS
(Continued)

Chapter **Page**

7.2.5 Hyperparameter Selection …… 85

7.2.6 Classifiers ……..…….. 85

7.2.2 Evaluation Metrics ……..…..…….. 88

7.3 Results ……..……...…….. 89

7.3.1 Cross-Validation ……..…….…….…….. 89

7.3.2 Test Set Prediction Performance .....…….. 91

7.4 Conclusion ..…….. 93

REFERENCES ……. 95

LIST OF TABLES

Table		Page
2.1	Dominant Genotypic Model 2x2 Contingency Table ……	8
3.1	Number of Correctly Predicted Case and Controls in Three External Datasets ..	29
3.2	Variants Found in Genes Previously Known to be Associated with CLL ……	26
3.3	Details of Top-Ranking Variants on the Full Dataset ……..	30
3.4	Details of Top-Ranking Variants on the GWAS Dataset …….	31
4.1	Kidney Cancer Datasets Used in the Study ……..	35
4.2	Total Numbers of SNPs in Datasets after Three Filtering Methods …….	36
4.3	Top Ten Ranked SNPs Used in the Cross-Study (Soft+Hard Filtering) ……..	42
4.4	Rank of SNPs (PCC) Hard+Soft Filtering in Known Kidney Cancer Genes …..	43
5.1	Samples Population ……..	47
5.2	SNPs Count after Applying Soft+Hard Filtering …….	49
5.3	Hardy-Weinberg Equilibrium Exact Test P-Values …….	51
5.4	Top SNPs for 1000 Genomes Project, GBM and LGG Datasets …….	56
5.5	Top SNPs for 1000 Genomes Project and GBM Datasets ……	56
5.6	Top SNPs for 1000 Genomes Project and LGG Datasets ……	57
5.7	All Datasets Genes Expression and Survival Time Association ……..	61
5.8	1000 Genomes and GBM Genes Expression and Survival Time Association ….	62
5.9	1000 Genomes and LGG Genes Expression and Survival Time Association …..	62
5.10	Top SNPs in All Datasets Genes and Functional Consequences …….	63
5.11	Top SNPs of 1000 Genomes and GBM Genes and Functional Consequences …	63

LIST OF TABLES
(Continued)

Table		Page
5.12	Top SNPs of 1000 Genomes and LGG Genes and Functional Consequences ….	64
5.13	Cases and Controls Top Ranked SNPs Alternate Allele Frequencies ……..	66
6.1	Samples Clinical Characteristics ……..	71
6.2	SNP, mRNA, and T1 MRI Data ……	72
7.1	Cohort Characteristics ……..	82
7.2	TCGA-GBM SNPs Count after Applying Three Filtering Methods …….	83
7.3	Training and Test Sets Characteristics …….	84
7.4	Prediction Accuracy on Test Set with the Optimal Hyperparameters ……..	92

LIST OF FIGURES

Figure		Page
2.1	Example of different human DNA variations, bases on red color represent the change occurred to the original sequence (in green shade background). Underlined bases represent the repeated subsequence in the DNA sequence. Bases with strikethrough represent deleted nucleotides from the original sequence……..	5
2.2	Cancer primary sites correlation; dots denote P < .01……	12
3.1	Whole-exome sequences are short reads of exomes obtained by next generation sequencing……..	15
3.2	Encoding of SNPs and InDels into 0,1 and 2 integers. GATK program identifies homozygous and heterozygous genotypes when there is a mutation or insertion deletion. For individuals where a SNP is not reported but found in a different individual, the study uses a value of 0……	19
3.3	Illustration of cross-validation technique……	23
3.4	Average cross-validation accuracy of support vector machine with top Pearson ranked SNPs and InDels together and separately on 100 50:50 training validation splits. Error bars indicate the standard deviation……	24
3.5	Average cross-validation accuracy of support vector machine with top Pearson ranked SNPs on 100 50:50 training validation splits of the GWAS dataset……	26
3.6	Accuracy of support vector machine with top Pearson ranked SNPs on just the external independent samples. Since this is a validation dataset one cannot uses the labels for any type of model training including ranking of features. Thus, the ranking of SNPs is obtained from the original full dataset…….	28
4.1	KIRP data three filtering average CV accuracy of SVM on top ranked SNPs……	39
4.2	KICH data three filtering average CV accuracy of SVM on top ranked SNPs……	40
4.3	Accuracy of support vector machine on the KICH dataset after trained on top ranked SNPs in the KIRP dataset……	41
5.1	A toy example for encoding a multiallelic site……	48
5.2	Germline SNPs calling pipeline with genome analysis toolkit performed on a cluster to speed up computation……..	50
5.3	Projection of principal component analysis with the first two components……	52

LIST OF FIGURES
(Continued)

Figure		Page
5.4	Projection of principal component analysis with the first two components after excluding the two outlier data points……...	52
5.5	10-fold cross-validation of learning and classifying binary labels…….	58
5.6	10-fold cross-validation of learning and classifying 3-class labels…….	58
5.7	Cross-study validation…….	60
5.8	Alignments of four cases vs four controls at SNP rs10792053 the upper four tracks for cases (LGG, GBM) viewed with IGV……	65
6.1	Multiallelic SNP encoding into a numerical values example…….	73
6.2	Proposed multi-modal deep neural network. One can see three paths each for SNP, gene expression, and images. The study trains the network as one model instead of training the three paths separately……..	74
6.3	Mean 10-fold accuracy of our network across 15 epochs for training and test sets with 3D volumes as the image data……..…	75
6.4	Test accuracy of each of the 10-folds of our network across 15 epochs on all three data sources combined with 3D volumes as the images……..	76
6.5	Mean 10-fold accuracy of our network across 15 epochs for training and test sets with handpicked 2D slices (that manifest the tumor) as the image data)……	77
6.6	Test accuracy of each of the 10-folds of our network across 15 epochs on all three data sources combined with 2D slices as the images……..	77
7.1	The proposed multi-path model architecture with SNP and demographic features.	87
7.2	Cross-validation average balanced accuracy across 10-folds as a function of the number of epoch and learning rate for multiple data inputs: demographic characteristics (age+age groups+gender) only, SNPs only, or SNPs and demographic characteristics combined. Each line color, which is shown in the series color legends, represents input data (learning rate in parentheses)……..	90

LIST OF FIGURES
(Continued)

Figure **Page**

7.3 Cross-validation mean balanced accuracy across 10-folds with linear SVM (with different C regularization values) and random forest (with different number of trees values) and multiple data inputs: demographic characteristics (age+age groups+gender) only, SNPs only, or SNPs and demographic characteristics combined. Each bar color represents a data source……. 91

7.4 Training accuracy on training set (n=244) for combined SNP and demographic features, and each data source individually……. 92

7.5 Test set prediction accuracy for combined SNP and demographic features, and each data source alone……. 93

CHAPTER 1

INTRODUCTION

Deep learning convolutional neural network (CNN) and existing machine learning methods such as support vector machine (SVM) and random forest (RF) successfully applied to a wide range of fields [1-4]. With the high availability of genomic and medical imaging data, the need increases for automated and accurate cancer risk and survival time predictions.

1.1 Cancer Risk Prediction

Cancer is the second leading cause of death in the United States [5]. The dissertation explores chronic lymphocytic leukemia, kidney cancer, and brain cancer risk predictions. Chronic lymphocytic leukemia accounts for 1.2% of all projected new cancer cases and 0.7% of projected cancer deaths in 2018 in the United States [5]. Kidney and renal pelvis account for 3.7% of the expected new cancer cases and 2.4% of estimated cancer deaths in the United States in 2018 [5]. Brain and other nervous system cancer new expected cases is 1.3% of all cancer new incidents, and 2.7% of all cancer deaths in 2018 [5].

Recent advances in deoxyribonucleic acid (DNA) sequencing technologies allowed sequencing massive parallel DNA fragments, which reduced the time and cost to generate human whole-genome and whole-exome sequencing data. A DNA sequence consists of a chain of letters from four nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T). The human genome comprises about three billion base pairs, and only identical twins may have the same or very similar DNA sequence. There are

different types of variation between human genomes. The most common variation type is Single Nucleotide Polymorphism (SNPs), which is a substitution at a specific locus of the genome. When comparing two human genomes, an SNP happens once about every 1000 nucleotides. The other types of variations, which involve one or more base pairs, are insertion, deletion, duplication, translocation, inversion, and copy number. Causes of many genetic differences in humans are vital in explaining heritable disease susceptibility and the presence of specific phenotypic traits. Linkage analysis and genome-wide association studies (GWAS) revealed more than 450 mutations [6], which predispose to glioma [7], colorectal [8], beast [9], ovarian [7], and other cancers types [6].

Areas of increasing interest in personalized medicine that utilizes DNA sequencing data are cancer risk prediction, gene editing, and cancer targeted therapy. Cancer risk prediction is vital to recommend specific regular checkups and tests for individuals with a high risk for a particular disease that could lead to early detection, which could enhance treatment outcomes.

The dissertation proposes the use of univariate ranking of genomic variations by computing Pearson correlation absolute value and chi-squared test statistic between each variant site and cancer status to weed out noisy features and reduce variants set dimensionality. The analysis shows that by decreasing variants' data set dimensionality support vector machine, and random forest classifiers achieved better classification performance.

1.2 Cancer Survival Time Estimation

Predicting glioma survival time helps patients and their clinicians evaluate available treatment plans and make informed choices. Glioblastoma multiforme (GBM) is the most common and aggressive type of brain cancer, with a median survival rate of 15 months [10]. Most advanced cancer patients prefer to know their estimated survival time [11]. However, clinicians' survival time estimates are inaccurate, and often optimistic [11, 12]. Many studies have devised 3D convolutional neural networks (CNNs) to improve the accuracy of structural magnetic resonance imaging (MRI) volumes to classify glioma patients into survival categories [3, 13-15].

This dissertation proposes a multi-path neural network system to predict glioblastoma survival categories from a heterogeneous combination of genomic variations, messenger ribonucleic acid (RNA) expressions, 3D post-contrast T1 MRI volumes, and 2D post-contrast T1 MRI modality scans that show the malignancy. The dissertation also proposes a new multi-path convolutional neural network with demographic features and SNP data to predict glioblastoma survival groups that improved upon SVM and random forest prediction accuracy.

1.3 Dissertation Contribution and Outline

The contribution of this dissertation is four-fold: 1) to investigate the predictive ability of support vector machine model and the effect of ranking SNPs with Pearson's correlation coefficient and chi-squared statistics in normal versus tumor samples in chronic lymphocytic leukemia (CLL) and kidney cancer subtypes, 2) to compare support vector machine and random forests prediction accuracy of germline SNPs in glioma subtypes

cases and healthy controls from the 1000 Genomes Project [16], 3) to propose a multi-path neural network from heterogeneous data sources: SNP, gene expression, 2D magnetic resonance imaging (MRI) scans, and 3D MRI volumes to classify glioma patients into short- versus long-term survival groups, and 4) to propose a new multi-path convolutional neural network for glioblastoma survival group prediction with SNP and demographic features.

In Chapter 2, the dissertation provides a problem description and a literature review. In Chapter 3, the dissertation proposes using SVM and feature selection to predict normal and tumor samples obtained from exome sequences variants in chronic lymphocytic leukemia (CLL). In Chapter 4, the study investigates the effectiveness of ranking SNPs on the predictive ability of SVM in kidney cancer subtypes normal and tumor samples. Chapter 5 compares the prediction accuracy of random forests and SVM in top-ranked SNPs to classify glioma subtypes individuals (cases) and healthy individuals (controls) from the 1000 Genomes Project. Chapter 6 proposes a multi-path neural network of combined neuroimaging, SNP, gene expression data to predict glioblastomas survival groups at the 14-year threshold. In Chapter 7, the dissertation devises a multi-path neural network architecture to predict short- and long-term survival classes in glioblastomas with multi-modal data.

CHAPTER 2

BACKGROUND

Human genomes differ among individuals in the population. The variations in human genomes give a rise to many phenotypic traits and diseases. First-degree relatives have the most similar genomes when compared to the population. Different types of variations occur between individuals' DNA sequences, such as substitution, insertion, deletion, translocation, inversion, and duplication. Figure 2.1 shows a toy example of human DNA variations.

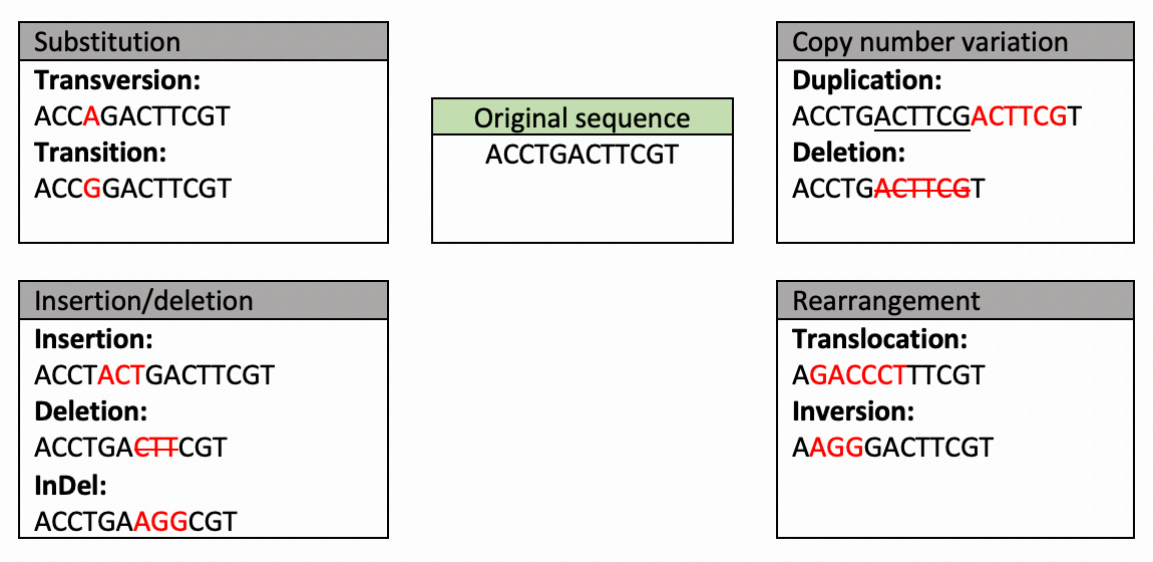


Figure 2.1 Example of different human DNA variations, bases on red color represent the change occurred to the original sequence (in green shade background). Underlined bases represent the repeated subsequence in the DNA sequence. Bases with strikethrough represent deleted nucleotides from the original sequence.

Substitution, also called single nucleotide polymorphism (SNP), involves a one base change in the DNA sequence that can be a transition or transversion. Transition is a type of a SNP where the base change is between purines bases [A, G] or pyrimidines [C,

T] bases. Transversion is single change in the DNA sequence that is between purine and pyrimidines bases. Even though the number of possible transversion are higher, transitions happen more frequently in human genome [17].

A single change in the DNA sequence results in missense, nonsense, or silent mutation. Missense and nonsense mutations alter the protein sequence and are more likely to effect protein function. Silent mutations do not modify amino acid sequence and often have no effect on protein function; however, these mutations can make a phenotypic change such as increasing/decreasing protein synthesis time [18].

Insertion/deletion (InDel) variations, which are the second common variations in human genome [19], are insertion, adding a subsequence to the DNA, or deletion, removing a subsequence from the DNA.

Translocations happen when a part of the DNA sequence is moved from one chromosome to another. Inversions occur when part of the DNA sequence is reverse complemented; for example, in Figure 2.1, the subsequence CCT is first reversed to TCC and then complemented to AGG. The complement for the base A is T and vice versa, and the complement for the base C is G and vice versa.

Copy number variation (CNV) is a type of structural variation where the number of copies in a DNA region varies among the population and involves thousands of nucleotides. There are two types of CNV: duplication and deletion. Duplication is where a one kilobase or more is repeated, and deletion is where one kilobase or more is lost from the DNA sequence.

For cancer risk predictions, there are different genomic-based data that can be used such as Polygenic Risk Scores (PRS) [20-22], DNA variants identified through

Genome-Wide Association Studies (GWAS) [23, 24], genomic variants detected by SNP arrays, and variants discovered from Next Generation Sequencing (NGS) data [25].

2.1 Polygenic Risk Score (PRS)

Polygenic Risk Score (PRS) is a continuous variable that is calculated from an ensemble of known markers for the disease of interest, which are obtained from published (GWAS) findings, one way to construct the PRS feature is to count the number of the known risk alleles present in each sample. Another way is to calculate the risk alleles and assign a weight specifically to each risk allele [26]. Many studies attempt to use PRS to estimate breast cancer risk in high-risk women [27-29]. In [27], the authors found that including PRS from known breast cancer SNPs have improved cancer risk prediction in high-risk women when compared to family history alone [26].

2.2 Genome-Wide Association Studies (GWAS)

The goal of GWAS is to interrogate human genome variation to identify statistically significant variations that differentiate large cohort of cases (individuals with the disease present) from controls (disease-free individuals) [30]. A common measure of the effect size of the association between a given SNP and a particular disease in GWAS is the odds ratio (OR). For example, in a biallelic SNP, which have only two possible bases, for the two allele copies in the DNA there are three unordered possible genotypes *A/A*, *A/a*, or *a/a*, where the letter '*A*' represents the major allele and the letter '*a*' represents the minor allele (less frequent allele). Table 2.1 gives an example of calculating alleles at a particular SNP for case and control groups in a 2X2 contingency table.

Table 2.1 Dominant Genotypic Model 2x2 Contingency Table

	A/A or A/a	***a/a***	**Total**
Disease (cases)	e	f	r1 = e+f
Healthy (controls)	g	h	r2 = g+h
Total	c1 = e+g	c2 = f+h	t = (r1+r2+c1+c2)

The odds ratio under the dominant model is then calculated from Table 2.1 as:

$$OR_{(A/A)\ or\ (A/a)} = \frac{e \times h}{f \times g} \tag{2.1}$$

To compute odds ratio or chi-squared statistic in a given SNP for cases and controls, there are different models to group the genotypes into two classes (2x2) instead of having a 2x3 table for genotypes '*a/a'*, '*A/a*', and '*A/A*'. These models are additive, multiplicative, recessive, and dominant.

To calculate the odds ratio under a dominant model for '*A*', the model assumes that having an '*A*' increases the risk and for recessive model vice versa, one needs to compute the odds of disease given that an individual carries an 'A' genotype and the odds of disease giving that an individual carries an '*a/a*' genotype, then takes the ratio of the two odds. In Equation 2.1, if the OR is greater than one, then the '*A*' genotype increases the risk of the disease. If the OR is less than one, then having a genotype of '*A*' decreases a person's risk of having the disease. However, if the OR is equal to one, then there is no association between the genotype and the disease. The chi-squared test is a standard test used in GWAS for calculating the statistical significance of a genotype, assuming a dominant/recessive model, for a particular disease. From Table 2.1, the chi-squared can be calculated, with a degree of freedom = 1 as:

$$E_{i,j} = \frac{r_i}{t} \times \frac{c_j}{t} \times t \tag{2.2}$$

$$\mathrm{X}^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \tag{2.3}$$

Where $O_{i,j}$ is the observed count in each cell in Table 2.1 for cases and controls, and $E_{i,j}$ is the expected count for each cell under independence assumption.

Many genome-wide association studies (case and control) have singled out SNPs and genes that are individually significant for gliomas [31-33]. Other studies identified several SNPs that are strongly associated with kidney renal clear cell carcinoma (KIRC) [34], cervical kidney renal papillary cell carcinoma (KIRP) [35], and chronic lymphocytic leukemia (CLL) [36].

GWA studies have identified many susceptibility loci for many cancers, but these novel variants cover only a small portion of the genome. Variants called from Next Generation Sequencing or SNPs array data have higher genome coverage, and therefore, there is a need to exploit these collective SNP data to assess its cancer risk predictive ability using machine learning methods.

2.3 SNP Array

An SNP array is a chip-based microarray technology offered primarily by Affymetrix and Illumina companies. Affymetrix genome-wide human SNP array 6.0 has 906,600 probes to genotype SNPs. The array is composed of hundreds of thousands of probes on a glass. Each probe contains multiple fixed short single-strand complement sequences for specific locus in the DNA sequence that binds to specific target sequence fragments (the ones

from the sample) and produces an intensity value for each allele. If two allele intensities have the same values, then the sample is heterozygous, which means the individual carries two different allele at that locus.

The dissertation analyzes Affymetrix arrays from CEL files, which are Affymetrix file format, raw data containing intensity values of the individual probes and locations for the hybridized array, after the array scan finishes [37]. In Chapter 3, the dissertation compares linear SVM classification accuracy in chronic lymphocytic leukemia (CLL) cases and controls with Affymetrix SNP array variants versus variants obtained from whole-exome sequencing (WES). The research obtained 232 samples' CEL files (for case and control samples) from the National Institute of Health (NIH) database of Genotypes and Phenotypes (dbGaP) portal and SNPs that are discovered with next-generation sequencing and genome analysis toolkit (GATK) [38, 39] of the same samples set. The dissertation uses Affymetrix Genotyping Console software to create samples genotype calls.

Affymetrix Genotyping software employs the Birdseed algorithm, which makes a multi-chip analysis to estimate a signal intensity of each SNP's allele, to make a genotype call, the algorithm fits a gaussian mixture model in two-dimensional A-signal vs. B-signal space [40]. The Birdseed algorithm assigns a confidence score for each genotype call between 0 and 1, where 0 is the highest quality, and 1 is the lowest [41]. The dissertation uses the program default contrast quality control threshold ≥ 0.4 for each sample.

2.4 Next Generation Sequencing (NGS) Variants Discovery

NGS technology allows faster and cheaper sequencing of human genomes than Sanger sequencing. The NGS technology sequences millions of short DNA fragments in parallel, and can sequence the whole-genome, or only whole-exome (coding regions). After obtaining the short reads from the sequencing machine, it needs to be mapped to a reference genome using alignment software like Burrows-Wheeler aligner (BWA), Bowtie, or Tophat. Once the short reads are mapped. A BAM/SAM file will be generated and used for downstream variants discovery analysis tools like GATK and SAMtools.

Cancer predictive ability depends on the cancer primary site with some cancers are more likely to be caused by germline mutations. For example, breast cancer is associated with mutations in gene BRCA1 and BRCA2. Therefore, it is expected to achieve higher prediction accuracy in some cancers, and low on others [42].

Many genomic variations are due to ancestry and geographical location of individuals [43]. Studies were able to determine the geographical origin of people based on their genetic makeup with high accuracy [43, 44]. Therefore, to build a model for cancer predication based on genetic variation, it is better to include specific individuals with common ethnicity and race.

Genomic factors are not the sole cause of most cancers. In fact, 90% of cancers are caused by somatic mutations, non-inherited changes in the DNA sequence, that are triggered by combination of contributing factors such as environmental, lifestyle, and genomic predisposition [42].

Recent studies suggest that some cancers share genetic mutation causes [45, 46]. Figure 2.2 shows the most closely correlated cancers and their P-values [46]. Shared

heritability among some cancers makes it theoretically possible to learn on one cancer and predict on related rare cancer. In Chapter 3, the dissertation explores the ability of generalizing predictive model learned on CLL and to predict lymphoma, and head and neck cancers. In Chapter 4 the dissertation preforms a cross-study where it learns a predictive model in one kidney subtype and predicts unseen samples from another kidney subtype.

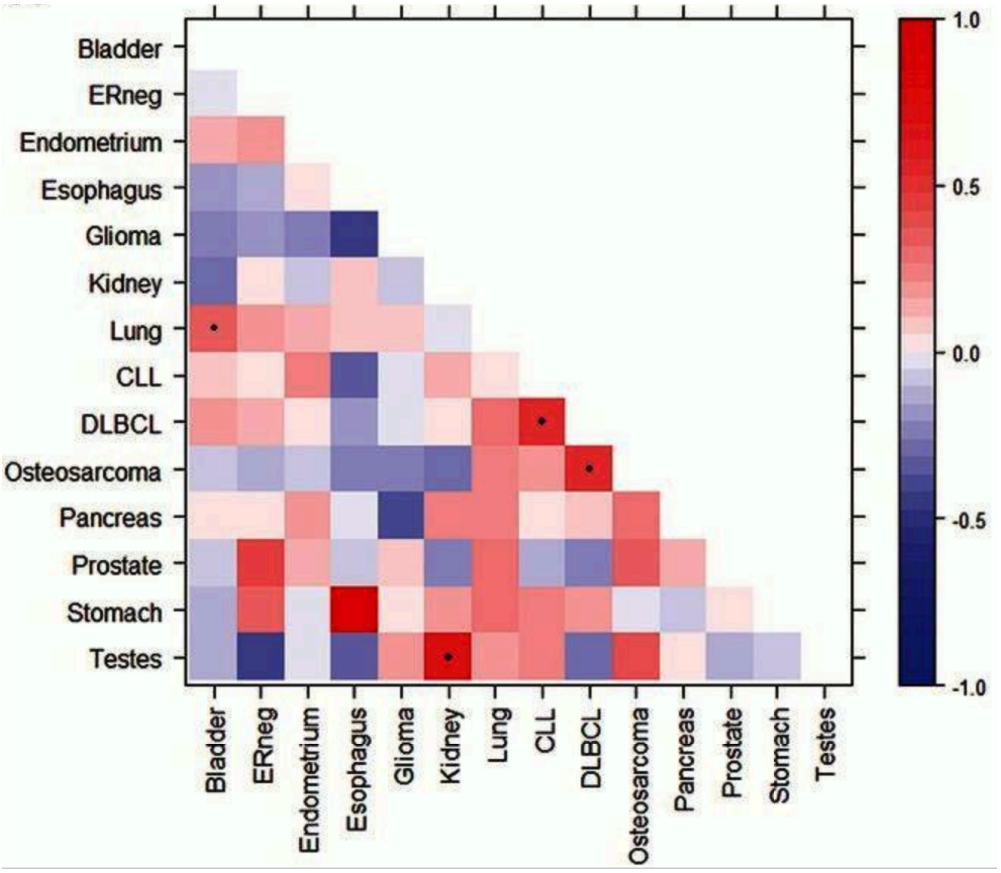


Figure 2.2 Cancer primary sites correlation; dots denote P < .01.

Source: [46]

CHAPTER 3

CROSS-VALIDATION AND CROSS-STUDY VALIDATION OF CHRONIC LYMPHOCYTIC LEUKAEMIA WITH EXOME SEQUENCES AND MACHINE LEARNING

3.1 Introduction

In the last few years, there have been many studies exploring disease risk prediction with machine learning methods and genome-wide association studies (GWAS) [47-56]. This includes various cancers and common diseases [57-61]. Most studies employ a two-fold machine learning approach. First, they identify variants from a set of training individuals that consist of both case and controls. This is usually a set of single nucleotide polymorphisms (SNPs) that pass a significance test, or a number of top-ranked SNPs given by a univariate ranking method. In the second part they learn a model with the reduced set of variants on the training data and predict the case and control of a validation set of individuals.

For diseases of low and moderate frequency, SNPs have been shown to be more accurate than family history under a theoretical model of prediction [62]. However, for diseases with high frequency and heritability family history-based models perform better [62]. Clinical factors with SNPs yield an area under curve (AUC) of 0.8 in a Japanese type 2 diabetes dataset but their SNPs have a marginal contribution of 0.01 to the accuracy [63]. With a large sample size, the highest known AUC of 0.86 and 0.82 for Crohn's disease and ulcerative colitis were reported [64]. There the authors contend this may be a peak or considerably larger sample sizes would be needed for higher AUCs. Bootstrap methods have given AUCs of 0.82 and 0.83 for type 2 diabetes and bipolar disease on the Wellcome Trust Case Control Consortium (2007) datasets, considerably

higher than previous studies. Some studies have also used interacting SNPs in GWAS to boost risk prediction accuracy [65, 66].

Many of these studies are cross-validation studies. They split the original dataset into training and validation several times randomly and for each split predict case and controls in the validation. Recent work has shown that this may not necessarily generalize to data from different studies [67]. Thus, in any risk prediction study it is now essential to include cross-study validation on an independent dataset.

While continuing efforts are made to improve risk prediction accuracy with GWAS datasets, the AUCs are still below clinical risk prediction particularly for cancer. The reasons posed for this failure include lack of rare variants, insufficient sample size, and low coverage (.1% of the genome sequenced) [68-70]. In this study, we detect variants from whole exome data that has a much larger coverage. We seek to determine the cross-validation and cross-study prediction accuracy achieved with variants detected in whole exome data and a machine learning pipeline.

The study obtained a chronic lymphocytic leukemia 140X coverage whole exome dataset [71] [72] of 186 tumor and 169 matched germline controls from the NIH dbGaP [73]. The whole exome dataset is composed of short next generation sequence reads of exomes as shown in Figure 3.1. This is one of the largest datasets available, and is an adult leukemia with an onset median age of 70 [74]. There is currently no known early SNP based detection test for this cancer. Current tests include physical exam, family history, blood count, and other tests given by the National Cancer Institute (see http://www. cancer.gov/cancertopics/pdq/treatment/cll/patient).

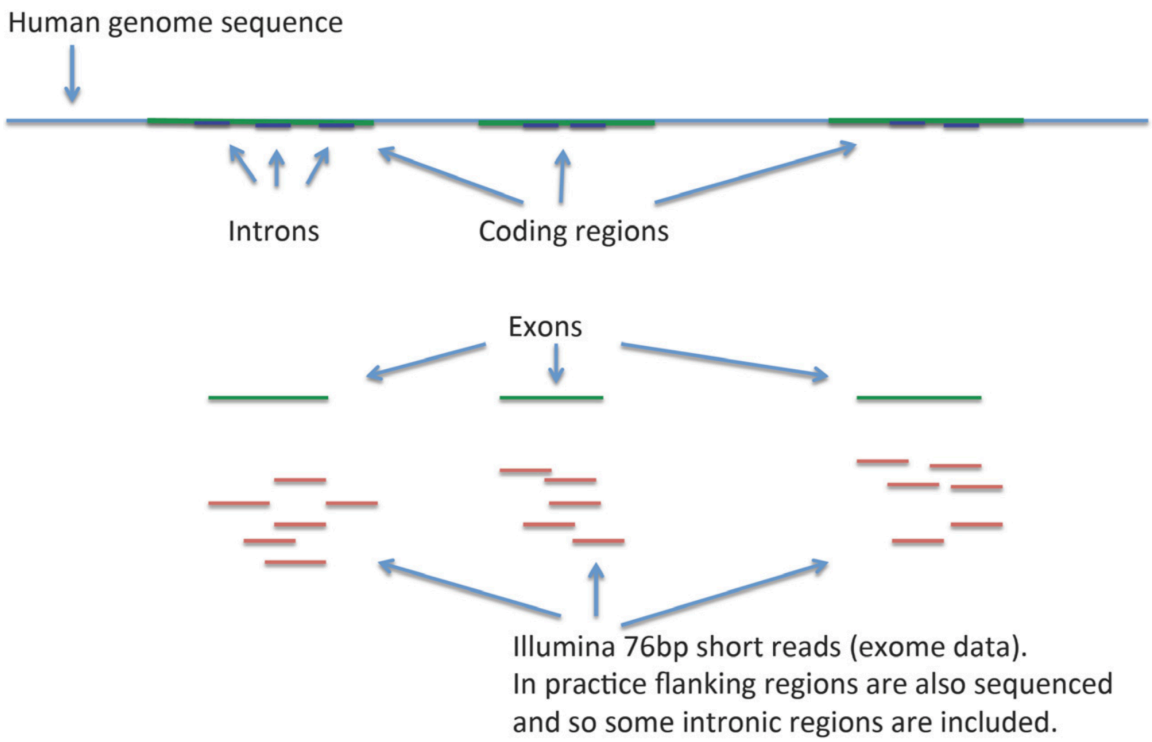


Figure 3.1 Whole-exome sequences are short reads of exomes obtained by next generation sequencing.

Short read exome sequences were mapped to the human genome reference GRCh37 with the popular Burrows-Wheeler aligner (BWA) [75] a short read alignment program. This research then used the genome analysis toolkit (GATK) [38, 39, 76] and the Broad Institute exome capture kit (bundle 2.8 b37) in a rigorous quality control procedure to obtain SNP and InDel variants. Cases and controls that contained excessive missing variants were excluded, and in the end 122,392 SNPs and 2200 InDels across 153 cases and 144 controls were obtained.

To better understand the risk prediction value of these variants, the research performs a cross-validation technique on the total 153 cases and 144 controls by creating random training validation splits. Then the dissertation compares the same cross-validation accuracy to that on an Affymetrix 6.0 panel genome wide association study for

the same subjects to see the improvement given by the exome analysis. The study obtained exome sequences from three different studies from dbGaP for independent external validation (also known as cross-study validation; [67]). The study ranked SNPs in training set with the Pearson correlation coefficient [77] and predicted labels of cases and controls with the support vector machine classifier in an external validation dataset. The research studied the biological significance of top Pearson ranked SNPs in the data.

3.2 Materials and Methods

Rigorous analysis on raw exome sequences was performed. First, sequences were mapped to the human genome and variants obtained. Then variants were encoded into integers to create feature vectors for each case and control sample.

3.2.1 Whole Exome Sequencing Data and Human Genome Reference

Whole exome sequences of 169 chronic lymphocytic leukemia patients [71, 72] was obtained from the NIH dbGaP website [73] with dbGaP study ID phs000435.v2.p1. Each of the 169 patients has matched tumor-normal sequencing data. In addition, exome tumor sequences of 17 patients were obtained from dbGaP after publication of the original study [71, 72]. This gives a total of 186 cases and 169 controls. The ancestry of the patients is not given in the publications or in the dbGaP files except that we know they were obtained from the Dana Farber Cancer Institute in Boston, Massachusetts, USA. The data comprises of 76 base pair (bp) paired-end reads produced by Illumina Genome Analyzer II and Hiseq2000 machines and the Agilent SureSelect capture kit by the Broad Institute [71]. The data was sequenced to obtain mean coverage of approximately 140X.

The research uses the human genome reference sequence version GRCh37.p13 from the Genome Reference Consortium (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/). At the time of doing this experiment, version 38 of the human genome sequence was introduced. However, the mapping process was started well before its release and demanded considerable computational resources. Therefore, this work continued the analysis with version 37.

3.2.2 Next Generation Sequencing Analysis Pipeline

The pipeline includes mapping short reads to the reference genome, post processing of alignment, variant calling, and filtering candidate variants. The total exome data was in approximately 3 Terabytes (TB) and required high performance computing infrastructure to process. Perl and various bioinformatics tools were used to automate the analysis pipeline.

Mapping Reads: As a first step, exome short read sequences of 186 tumor cases and 169 matched germline controls were mapped to the human genome reference GRCh37 with the BWA MEM program [75]. Six cases and 14 controls were excluded due to excessively large dataset size and downloading problems, and reads with mapping quality (MAPQ) below 15 were removed.

The read mapping is a process where short read DNA sequences mapped to a reference genome. There are many different programs available for this task, and each one differs in mapping methodology, accuracy, and speed. This pipeline uses the popular program BWA MEM program (version 0.7a-r405) [75] that implements the Burrows-Wheeler transform. BWA MEM is relatively accurate for its fast processing speed while

mapping against vast reference genome such as humans [78, 79]. Default parameters were used for mapping reads to the human reference genome.

BWA MEM produces its output in a standard format called Sequence Alignment Map (SAM). SAMtools version 0.1.18 [75] were used for further analysis of the SAM output files. Each alignment in SAM format was converted into its binary format (BAM), alignments were sorted with respect to their chromosomal position, and then indexed. SAMtools were used to generate mapping statistics and merging alignments of the same patient across different files. PICARD tool (version 1.8, http://broad institute.github.io/picard/) were used to add read groups, which connects the reads to the patient subject. The pipeline removed duplicates reads introduced by the PCR amplification process to avoid artifacts using the PICARD MarkDuplicates program. Finally, using SAMtools, unmapped reads and the ones with mapping score (given in the MAPQ SAM field) smaller than 15 were removed.

Variants Detection: GATK [38, 39, 76] version 3.2-2 with the Broad Institute exome capture kit (bundle 2.8 b37 available from ftp://ftp.broadinstitute.org/) was used to detect SNPs and InDels in the alignments. These SNPs and InDels are referred to as variants. These variants pass a series of rigorous statistical tests [39, 76]. If a variant does not pass the quality control or no high-quality alignment of a read to the genome was found in that region, then GATK reports a missing value.

The analysis found 38 individuals that contain at least 10% missing values. These samples were removed from the data and variants recomputed for the remaining 153 cases and 144 controls again with GATK and the exome capture kit. The study also removes all variants that have at least one missing value, and so eliminate the need for

imputation. Note that if these variant features with missing values were removed before the 38 individual samples with many missing values, it going to generate just a few variants with limited predictive value.

The variant detection procedure gave a total of 122392 SNPs and 2200 InDels. These variants were then encoded into integers thus obtaining feature vectors for each case and control. Figure 3.2 shows the encoding process to get integer values for variant features.

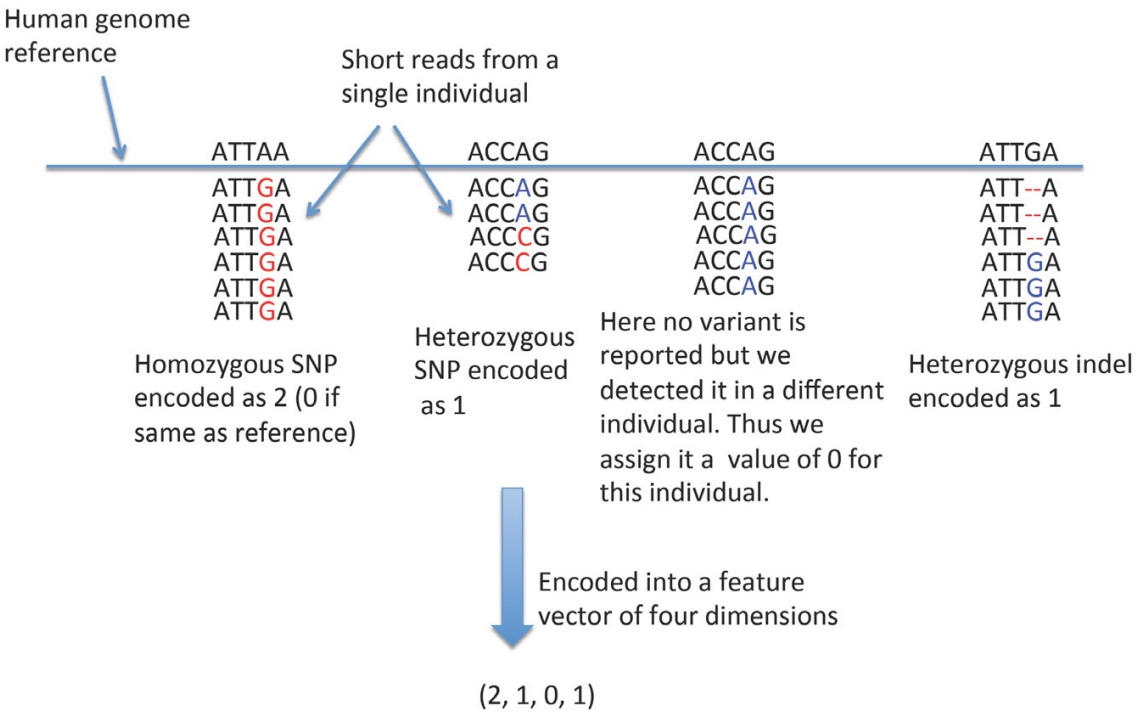


Figure 3.2 Encoding of SNPs and InDels into 0,1 and 2 integers. GATK program [39] identifies homozygous and heterozygous genotypes when there is a mutation or insertion deletion. For individuals where a SNP is not reported but found in a different individual, the study uses a value of 0.

3.2.3 Machine Learning Pipeline

After completing the variant analysis in the previous step, the study proceeds with the machine learning analysis. Machine learning methods are widely used to learn models

from classified data to make predictions on unclassified data. They consider each data item as a vector in the space of dimension given by the number of features. In this study, each data item is a case or control set of exome sequences. By mapping each set to the human genome, variants were obtained, which represent features. Thus, the number of variants determines the number of dimensions in the feature space.

Data Encoding: Since the input to machine learning programs must be feature vectors, each SNP and InDel converted into an integer. The variants reported by GATK are in standard genotype form A/B where both A and B denote the two alleles found in the individual. The GATK output is in VCF file format whose specifications (available from http://samtools.gith ub.io/hts-specs/VCFv4.1.pdf) provide details on the reported genotypes. When A = 0 this denotes the allele in the reference. Other values of 1 through 6 denote alternate alleles and gaps. The study kept the max alternative allele option to six, which is also the default value in GATK. The pipeline performs the encoding 7 A + B to represent all possible outputs.

Each feature vector represents variants from a human individual and is labeled –1 for a case and +1 for a control. The labels +1 and –1 are standard in the machine learning literature [80].

Feature Selection: The research ranks the features with the Pearson correlation coefficient (PCC) [77].

$$PCC = \frac{\sum_i^n (x_{i,j} - x_{j,mean})(y_i - y_{mean})}{\sqrt{\sum_i^n (x_{i,j} - x_{j,mean})^2} \sqrt{\sum_i^n (y_i - y_{mean})^2}} \quad (3.1)$$

where $x_{i,j}$ represents the encoded value of the j^{th} variant in the i^{th} individual and y_i is

the label (+1 for a case and –1 for a control) of the i^{th} individual. The Pearson correlation ranges between +1 and –1 where the extremes denote perfect linear correlation and 0 indicates none. The study ranks the features by the absolute value of the Pearson correlation.

Classifier: The pipeline uses the popular soft margin support vector machine (SVM) method [81] implemented in the SVM-light program [82] to train and classify a given set of feature vectors created with the above encoding. In brief, the SVM finds the optimally separating hyperplane between feature vectors of two classes (case and control in our case) that minimizes the complexity of the classifier plus a regularization parameter C times error on the training data. For all experiments, the pipeline uses the default regularization parameter given by:

$$C = \frac{1}{\sum_i^n x_i^T x_i} \tag{3.2}$$

where n is the number of vectors in the input training (case and control individuals in this study) and x_i is the feature vector of the i^{th} individual [82], in other words, the C is the inverse of the average squared length of feature vectors in the data.

Measure of Accuracy: We define the classification accuracy as 1-BER, where BER is the balanced error rate [83]. The balanced error is the average misclassification rate across each class and ranges between 0 and 1. For example, suppose class case has 10 individuals, and class control has 100. If the pipeline incorrectly predicted 3 cases and 10 controls, then the balanced error is:

$$\frac{\left(\frac{3}{10} + \frac{10}{100}\right)}{2} = 0.2 \tag{3.3}$$

3.2.4 High Performance Computing

The research uses the Kong computing cluster and the condor distributed computing system at NJIT to speed up the computation.

3.3 Results

Next-generation sequencing pipeline and data encoding give feature vectors each representing a case or control sample and each dimension representing an SNP or InDel variant. The study employs a machine learning procedure to understand the predictive value of the variants.

3.3.1 Cross-Validation

Cross-validation is a standard approach to evaluate the accuracy of a classifier from a given dataset [80]. The pipeline randomly shuffles the feature vectors and picks 50% for training and leaves the remaining for validation. On the training, the study ranks the variants with the Pearson correlation coefficient. This step is key to performing feature selection in a cross-validation study. Alternatively, one may perform feature selection on the whole dataset and then split it into 50% training. However, this method is unrealistic because in practice test labels are not available. In the cross-validation study simulates that setting by using a validation dataset in place of the test data. The validation labels are only to evaluate the accuracy of the classifier and should not be used for any model training, including feature selection. Some studies make this mistake (as previously identified; [84], but here the pipeline performs all SNP selection only on the training data.

The pipeline then learns a support vector machine [81] with the SVM-light software [82] and default regularization on the training set with k top-ranked SNPs (see

Figure 3.3). This study considers increments of ten variants up to 100 and increments of 100 up to 1000. Thus, the values of k=10, 20, 30, ..., 100, 200, ...,1000. For each value of k, this experiment predicts the case and control status of the validation samples and record the accuracy. The pipeline repeats this for 100 random splits and graph the average with standard deviations.

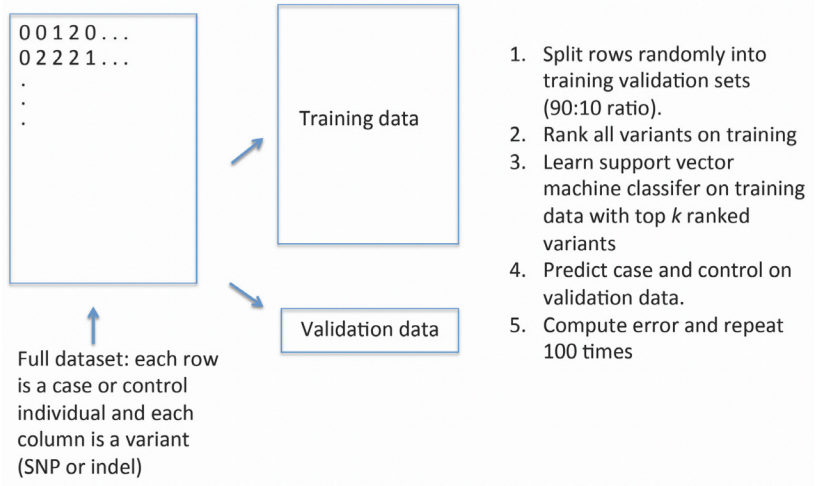


Figure 3.3 Illustration of cross-validation technique.

Figure 3.4 shows the mean cross-validation accuracy of the support vector machine on 50% training data across 100 random splits. It shows that InDels alone have much poorer accuracy than SNPs alone and contribute marginally to the SNPs. This experiment achieved a top accuracy of about 82% with the top 20 SNPs. The accuracy drops after the top 20 SNPs threshold.

Accuracies shown in Figure 3.4 are averaged across 100 training validation splits. In each split, this study first ranks the SNPs and compute prediction on validation with top k ranked ones. There is no one set of 20 SNPs to be identified recall that the accuracies shown in Figure 3.4 are averaged across 100 training validation splits. In each

split, the pipeline first ranks the SNPs and computes prediction on validation with top k ranked ones. Thus, there is no one set of 20 SNPs to be identified here, and this is certainly not the same as the top 20 SNPs from the ranking on the full dataset (although there are some in common with top-ranked ones from different splits). Alternatively, one may consider the intersection of the top 20 SNPs from all 100 split and use them for prediction on an independent external dataset. The drawback here is that not all of the SNPs in the intersection may pass the GATK quality control filtering thresholds. This is why this research ranks SNPs on the full dataset and considers the first top 100 that are found in the external dataset.

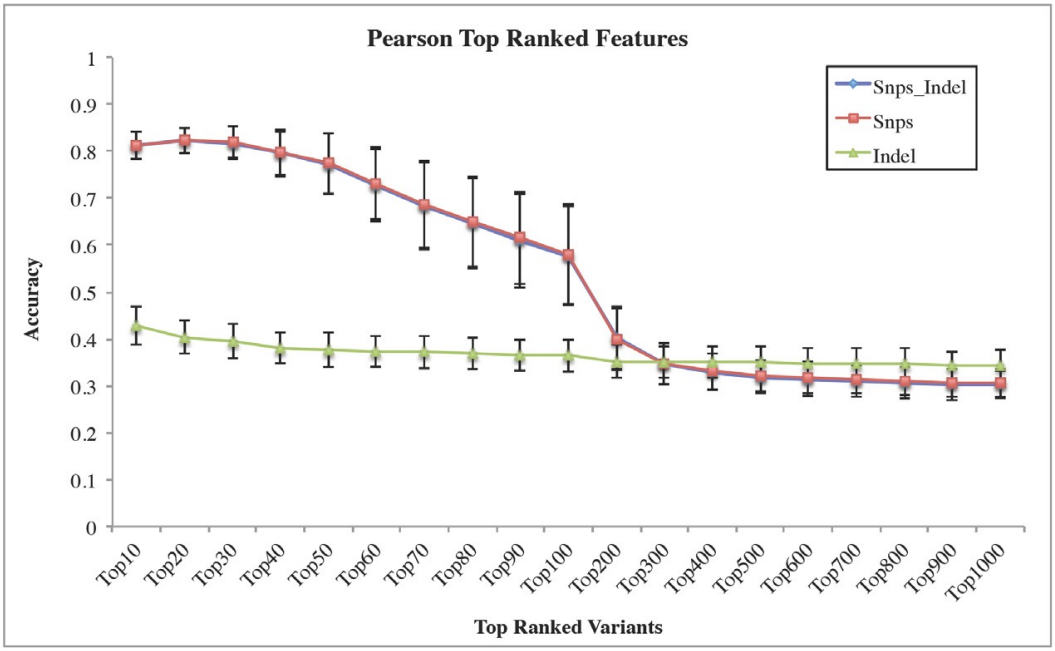


Figure 3.4 Average cross-validation accuracy of support vector machine with top Pearson ranked SNPs and InDels together and separately on 100 50:50 training validation splits. Error bars indicate the standard deviation.

3.3.2 Comparison to Cross-Validation on GWAS

To better understand the cross-validation results from SNPs obtained in the whole-exome sequencing analysis, this study examines a GWAS for the same subjects. This is an Affymetrix 6.0 genome-wide human SNP array of the same disease and subjects that have been obtained from the dbGaP site for the whole exome study. The study first removes SNPs with more than 10% missing entries and excludes samples that do not pass the quality control test with 0.4 threshold in the Affymetrix Genotyping Console. The quality control test measures the differences in contrast distributions for homozygote and heterozygote genotypes in each cel file. Following this, the research ranks the SNPs with the Pearson correlation coefficient. Then creates one hundred random 50:50 train and validation splits and determines the average prediction accuracy of the support vector machine in the same manner as described above for the whole-exome sequencing study.

Figure 3.5 shows that the GWAS SNPs give the highest prediction accuracy of 57% in the top 10 SNPs, but then it gradually decreases. Thus, the SNPs given by the whole exome analysis, which yields higher prediction accuracy, may serve as better markers that are not found in the GWAS. Upon closer examination, one sees that there is no overlap between the top 1000 ranked SNPs from the exome sequencing and GWAS datasets except for the four that have low Pearson correlation values.

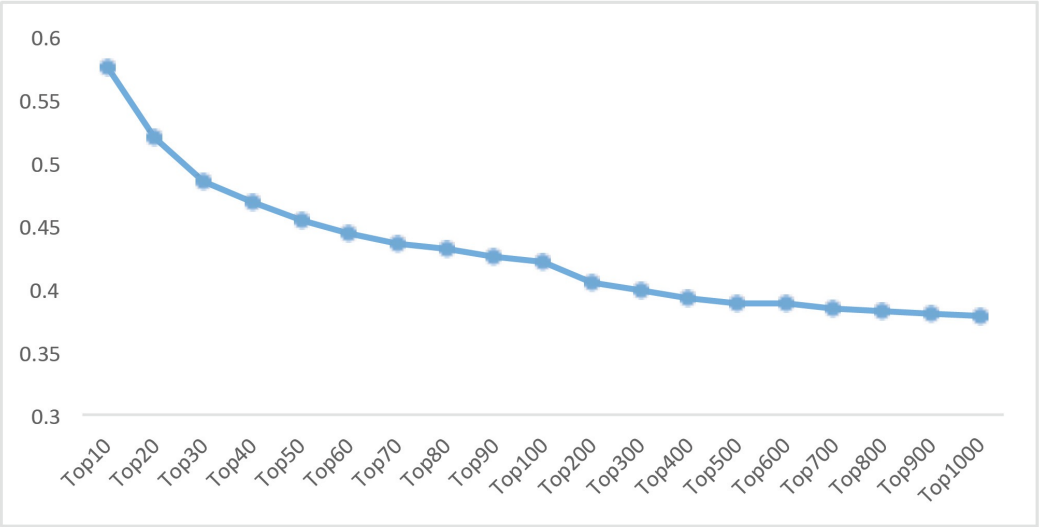


Figure 3.5 Average cross-validation accuracy of support vector machine with top Pearson ranked SNPs on 100 50:50 training validation splits of the GWAS dataset.

3.3.3 Cross-Study Validation

For cross-study validation on an independent dataset the pipeline considers a lymphoma whole exome study that has case subjects for lymphocytic leukemia as well as a few controls. This research considers controls from a head and neck cancer and a breast cancer study.

- Eighteen cases and three controls from a lymphoma whole exome study with dbGaP study ID phs000328.v2.p1 [85]. Reads are 101bp length produced from Illumina HiSeq 2000 machine and have 3.4X coverage. The ancestry or origins of data in this study are unavailable in the publication and the dbGaP site.

- Three controls from neck and head cancer whole exome study with dbGaP study ID phs000328.v2.p1 [86]. Reads are 77bp length produced from Illumina HiSeq 2000 and have 6.9X coverage. Individuals in this study are from the University of Pittsburgh Head and Neck Spore neoplasm virtual repository.

- Seven controls from breast cancer whole exome study with dbGaP study phs000369.v1.p1 ID [87]. Reads are 77bp length produced from Illumina HiSeq 2000 of coverage 5.9X. Individuals in this study have Mexican and Vietnamese ancestry.

In all three datasets, this research follows a similar procedure that used for the chronic

lymphocytic leukemia exome dataset. The pipeline maps the short reads to the human genome with the BWA program and detects variants with GATK using the same software and parameters as for the lymphocytic leukemia dataset.

Since this is a validation dataset, one cannot use the labels to perform any feature selection or model training. Instead, one learns the support vector machine model from the full original dataset. The study refers to that as the training set here. The study does not consider all SNPs from the training dataset to build a model. First, the study obtains the top 1000 Pearson correlation coefficient ranked SNPs in the full training. Many of these SNPs do not pass the GATK quality control tests on some of the external validation samples. One reason for this is the much lower coverage (<10X) of the external datasets. Amongst the ones that were detected, this research considers just the top 100 ranked ones. For each top k ranked ones (for k =10,20,30,...,100), the research learns a support vector machine model on the training and uses it to predict labels of the validation data. As previously discussed, the top k ranked SNPs here are not the same as the top k ranked SNPs in the earlier cross-validation study.

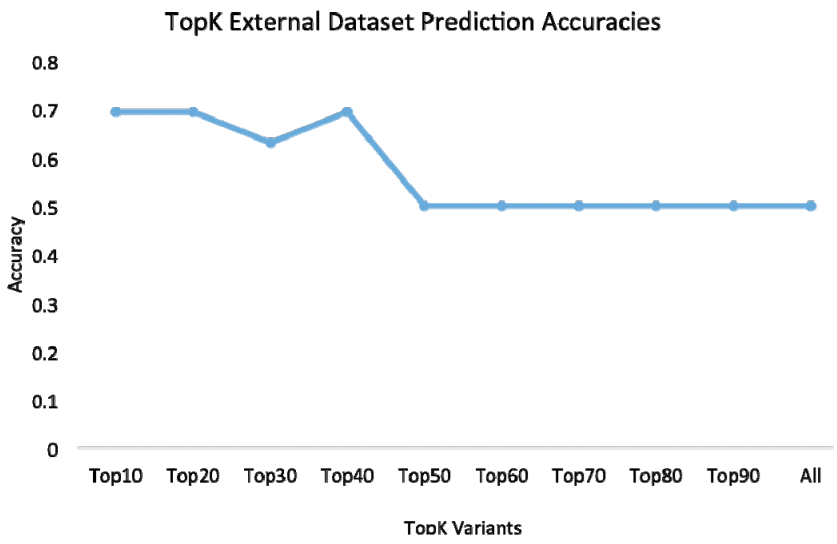


Figure 3.6 Accuracy of support vector machine with top Pearson ranked SNPs on just the external independent samples. Since this is a validation dataset one cannot uses the labels for any type of model training including ranking of features. Thus, the ranking of SNPs is obtained from the original full dataset.

Figure 3.6 shows that only the top-ranked SNPs give prediction accuracy above 0.5. The study examines the number of cases and controls predicted correctly by the top 20 ranked SNPs in Table 3.1. Note that the imbalanced accuracy from the table is 64.5%. However, this research uses the balanced accuracy that accounts for different sizes of each class, and that value, which is plotted in Figure 3.6, is 69.4%. Table 3.1 shows that the controls for the head and neck cancer are correctly predicted. In the lymphoma dataset also, all controls are correctly classified, but more than half cases are incorrectly classified as controls.

Table 3.1 Number of Correctly Predicted Case and Controls in Three External Datasets

Study	Cases	Controls	Correct cases	Correct controls
Lymphoma	18	3	7	3
Head and neck cancer	0	3	0	3
Breast cancer	0	7	0	7

3.3.4 Biological Significance of Top Tanked SNPs

The study considers the top 200 ranked SNPs in the Pearson correlation ranking of all SNPs in the full dataset. Those variants were fed to the popular ANNOVAR program [88] to determine genes and genomic regions they lie on.

The study founds SNPs in genes SF3B1 and MYD88 both of which were reported as significant genes in the original study of the dataset [71]. It also founds SNPs in genes STRN4 and HLA-DRB5 both of which have been shown to be previously associated with this disease in genome wide association studies [89-92] . Table 3.2 provides additional details of the SNPs in these genes. All four are exonic but don't necessarily rank high in Pearson correlation coefficient.

Table 3.2 Variants Found in Genes Previously Known to be Associated with CLL

Pearson	Chr	Pos	Rank	Region	Gene	Ref	Alt	Type
0.19	19	47230736	93	Exonic	STRN4	G	T	Hom
0.19	3	38182641	98	Exonic	MYD88	T	C	Hom
0.19	2	198266834	98	Exonic	SF3B1	T	C	Hom
0.17	6	32497985	159	Exonic	HLA-DRB5	A	G	Hom

Note: The first column gives the Pearson correlation coefficient value, followed by chromosome number, position in chromosome, SNP rank given by the Pearson correlation coefficient, genomic region, gene, reference nucleotide, alternate nucleotide, and the type.

The research also provides the SNPs information from the top three high-ranking genes in Table 3.3. There it shows that the Pearson correlation of the top ranked SNPs is considerably higher than the SNPs in known genes identified above. While their direct association with lymphocytic leukemia is unknown, they are well implicated in many different cancers. The highest rank is the Aminoacyl tRNA synthetases (AARS) gene that is known to be associated with various cancers [93]. Following this is the valyl-tRNA synthetase (VARS) gene that is also known to be associated with cancer [94]. The WD repeat domain 89 (WDR89) is associated with many cancers as given by the Human Protein Atlas (http://www.proteinatlas.org/ENSG00000140006-WDR89/cancer) and The Cancer Network Galaxy (http://tcng.hgc.jp/index.html?t=gene id=112840).

Table 3.3 Details of Top-Ranking Variants on the Full Dataset

Pearson	Chr	Pos	Rank	Region	Gene	Ref	Alt	Type
0.72	16	70305806	1	Exone	AARS	G	A	Hom
0.71	16	70305809	2	Exone	AARS	G	A	Hom
0.59	16	70305812	3	Exone	AARS	C	A	Hom
0.36	6	31749930	5	Exone	VARS	C	G	Hom
0.33	14	64066352	9	Exone	WDR89	T	A	Hom

Table 3.4 lists top ranking SNPs from the GWAS with previous association to this disease and that lie on known genes. Some of these genes are previously linked to leukemia. For example, EML1 [95], KDM4C [96], NEBL [97], BNC2 [98], and ANO10 [99] are all known to be associated with leukemia while RGS20 and ZNF25 are known to be expressed in leukemia. However, none of the top 1000 ranked SNPs in the GWAS overlap with the ones from the exome study except for four that lie far down in the

rankings.

Table 3.4 Details of Top-Ranking Variants on the GWAS Dataset

dbSNP ID	Pearson	Chr	Pos	Rank	Gene
rs1951574	0.33	14	100346664	4	EML1
rs1905359	0.33	8	54851272	5	RGS20
rs2792228	0.32	9	6976680	6	KDM4C
rs11011415	0.31	10	38264389	7	ZNF25
rs3900922	0.31	10	21287528	8	NEBL
rs3739714	0.31	9	16435848	9	BNC2
rs9844641	0.31	3	43476335	10	ANO10

3.4 Discussion

In addition to the results shown here two variations were explored in the machine learning pipeline to see if they would increase prediction accuracy. First, the study looked at a naive encoding where it converts homozygous alleles to 0 and 2 and the heterozygous to 1. This marginally lowered the accuracy. Second, the research considered the chi-square ranking of SNPs instead of Pearson correlation and this also marginally lowered the accuracy.

One main challenge in this study is the size of the training set that is considerably smaller than sample sizes (of several thousand) used in GWAS based risk prediction studies. The primary source of data is the NIH dbGaP repository and so the sample sizes are limited to the data accumulated there.

Another challenge is the quality and coverage of data in dbGaP. For the three external studies, the research aimed to predict case and control of many samples. Yet for several of the downloaded datasets coverage was insufficient and the analysis founds the top-ranked variants only in a few samples.

Finally, differences in ancestry can affect risk prediction [100-102]. In this case the pipeline learned a model from data obtained in patients at the Dana Farber Cancer Institute in Boston, Massachusetts. In the three external datasets one is of Mexican and Vietnamese ancestry whose genetics are likely to be different from patients at the Dana Farber Institute.

3.5 Conclusion

Starting from raw exome sequences this study obtained a model for predicting chronic lymphocytic leukemia after a rigorous next generation sequencing and machine learning pipeline. The analysis evaluated the model in cross-validation studies as well as on three independent external datasets as part of cross-study validation. In cross-validation, the pipeline achieves a mean prediction of 82% compared to 57% obtained on an Affymetrix 6.0 panel genome wide association study. In the external cross-study validation, the pipeline obtains 70% accuracy with a model learned entirely from the original dataset. Finally, the study shows biological significance of top-ranking SNPs in the dataset. The research shows that even with a small sample size we can obtain moderate to high accuracy with exome sequences and is thus encouraging for future work.

CHAPTER 4

CROSS-VALIDATION AND CROSS-STUDY VALIDATION OF KIDNEY CANCER WITH MACHINE LEARNING AND WHOLE EXOME SEQUENCES FROM THE NATIONAL CANCER INSTITUTE

4.1 Introduction

Cancer risk prediction from one's DNA is of considerable interest in modern medicine [103, 104]. One way to achieve this is to determine mutations by comparing DNA in tumor cells to healthy ones. Such mutations are called somatic and could potentially be used for early detection and prevention of cancer [105-107].

The majority of efforts on predicting cancer are focused on using SNPs obtained from genome-wide association studies and from whole exome sequences [108-112]. However, there are also dangerous pitfalls associated with SNP-based cancer risk prediction [113]. The most common one is lack of validation on an independent dataset, also known as cross-study validation [67]. Most studies focus on the cross-validation accuracy, which is obtained by splitting a given dataset randomly into training and validation several times and obtaining the average accuracy on the validation. In a cross-study validation we want to see how well SNPs determined on data for one disease from a specific study generalizes to the same disease or a related one from a different study.

This research explores the accuracy of predicting kidney cancer case and controls with somatic mutations across two different whole exome sequence datasets obtained from the National Cancer Institute Genomic Data Commons database [114]. This research considers datasets of renal papillary cell carcinoma and chromophobe renal cell carcinoma. The data are pre-aligned short read sequences from which the pipeline determines variants. Three quality control methods of variant detection were examined

and the most rigorous one gives the most parsimonious model with the highest accuracy.

The important result is the cross-study validation between the two datasets, and this experiment achieves an accuracy of 66.2% when predicting chromophobe individuals after learning a model of ten SNPs from the renal papillary dataset. The work here suggests that it can predict kidney chromophobe carcinoma with high quality SNPs obtained from a kidney papillary carcinoma dataset. The following sections describe the methods in detail followed by experimental results.

4.2 Methods

This section describes the data along with the quality control protocol used. Then it describes the machine learning pipeline.

4.2.1 Data

There are several kidney cancer whole exome datasets at the National Cancer Institute (NCI) Genomic Data Commons (GDC) portal from across three different projects: The Cancer Genome Atlas (TCGA), TARGET, and Foundation Medicine Adult Cancer Clinical Dataset (FM-AD). Authorization to the TCGA project only was obtained, and two of the TCGA three datasets were downloaded.

- Kidney Renal Papillary Cell Carcinoma (KIRP): total of 291 individuals.
- Kidney Chromophobe (KICH): total of 113 individuals.

Both datasets contain individuals of European, African, and Asian ancestry, and have older patients between cancer stages I and III. For each individual in each dataset exome sequences of the affect cell and a healthy cell (from the same person) are made available.

To avoid mutations that occur from ancestry differences, only individuals of European ancestry (which is also the majority ancestry) were considered. Due to time constraints and checksum/download errors, only some male subjects from each study were downloaded. Table 4.1 gives the number of case and controls that were obtained for each study.

Table 4.1 Kidney Cancer Datasets Used in the Study

Dataset	Cases	Controls
TCGA-KIRP	110	110
TCGA-KICH	34	34

Each case and control file that was downloaded is a pre-aligned exome sequence mapped to the human genome reference (build 38, version GRCh38.d1.vd1) with the BWA program [75]. Thus, from the GDC portal BAM files were obtained [75] for each individual's tumor and healthy exome sequences. These are binary files of the SAM format that show the alignment of each short read to the reference genome.

The NCI GDC portal also contains files with already detected variants for each individual. However, those variants were obtained by comparing each individual's healthy exome sequences to their tumor ones. In this analysis, a collective analysis of all the individuals at the same time to determine variants were performed, so it can detect missing values as explained below.

4.2.2 Quality Control for Determining SNPs

The pipeline combines all the case and controls and performs a collective variant calling with the popular Genome Analysis Toolkit (GATK) software [38, 39, 76]. In the

collective analysis, the study was able to identify SNPs that are not reported across samples. For example, if a SNP does not pass quality control it is not reported and is thus a missing value. The study explores three filtering methods of obtaining SNPs with the popular Genome Analysis Toolkit (GATK) software [38, 39, 76]. By default, any reads with a MAPQ quality score (which is a measure of the alignment quality) below 25 is eliminated in the analysis:

- Soft filtering: This is the GATK Variant Quality Score Recalibration, which uses machine learning to identify good variants from bad ones.
- Hard filtering: Any SNP with a genotype quality score below 30 and a depth below 5 is ignored. The genotype quality score is a statistical quantity the gives us the accuracy of the SNP and the depth is the minimum of reads that contain the SNP. These are default values used in the GATK program.
- Soft and hard filtering: Both of the above are applied.

After each filtering method, the pipeline removes any SNP that is missing (not reported by GATK) in at least one sample, thus eliminating the need for imputation. After filtering, the number of SNPs that are obtained is given in Table 4.2. The same table also shows the number of SNPs common in the two studies, this set is used for the cross-study validation.

Table 4.2 Total Numbers of SNPs in Datasets after Three Filtering Methods

	Filtering Method		
Dataset	**Soft**	**Hard**	**Soft + Hard**
TCGA-KIRP	264858	131141	109700
TCGA-KICH	246290	135937	111394
Intersection of KIRP and KICH	131157	44426	36029

4.2.3 SNP Encoding

Once a dataset of SNPs is obtained after the quality control described previously, the pipeline performs encoding. The GATK program outputs variants in the VCF format [115], which encodes the reference allele as 0 and alternate alleles (including gap) from 1 onwards. For example, a genotype of 0/0 means the individual is homozygous in the reference allele, 0/2 means heterozygous in the second alternate allele, and 1/1 means homozygous in the second alternate allele. Therefore, it can be encoded to unique numbers with the simple formula 4A+B for a SNP encoded as A/B.

To evaluate the predictive capability of SNPs, the research performs cross-validation and a cross-study validation experiments. In the cross-validation, the analysis splits a given dataset into training and test and evaluates the error of predicting the test. A high accuracy does not necessarily mean the SNPs would generalize to other datasets or related diseases. Thus, a cross-study validation was performed to determine generalization to another dataset.

4.2.4 Cross-Validation and Machine Learning

A step-wise cross-validation procedure is as follow:

1. First, the analysis performs 10-fold cross-validation, where data roughly divided into ten equal parts. The first 90% set is the training data, and the remaining 10% is the test data.

2. The pipeline ranks the SNPs according to the Pearson correlation coefficient [80] as implemented in the Python scikit-learn machine learning library [116]. The Pearson correlation coefficient is the sample correlation coefficient that measures the covariance between two variables divided by their variances to normalize. A value close to 1 or -1 indicates a linear correlation whereas 0 means the variables are uncorrelated [80].

3. The pipeline considers the top k ranked SNPs for increasing values of k and trains

a support vector machine (SVM) [81], a fast linear classifier with known powerful generalization capabilities, also with the linear SVM in the Python scikit-learn library [116]. Then cross-validate the regularization parameter C of the SVM by cross-validating on the training set only.

4. With the trained model, the pipeline predicts cases and controls of the individuals in the test dataset and determines the error (since their true case and control status is known).

5. Steps 1 through 4 are repeated ten times, and then the average error is calculated.

4.2.5 Cross-Study Validation

The research aims to determine the error of predicting case and controls across two independently obtained studies. It measures the error of a predicting case and control in the KICH dataset, which contain individuals with renal chromophobe carcinoma, with a model trained on the KIRP dataset, which are renal papillary carcinoma individuals.

4.3. Results

Here the results of The Cancer Genome Atlas Kidney Chromophobe (TCGA-KICH) and The Cancer Genome Atlas Cervical Kidney renal papillary cell carcinoma (TCGA-KIRP) datasets cross-validation are outlined.

4.3.1 Cross-Validation

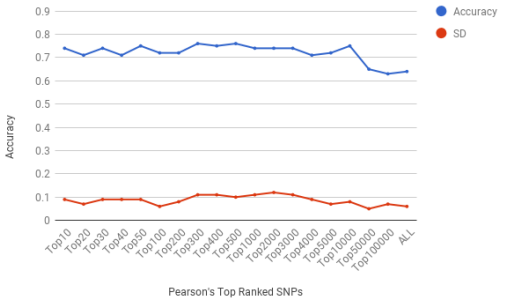
Figure 4.1 shows the average accuracy of the support vector across 10-fold cross-validation of the KIRP dataset. The research makes several interesting observations consistent with previous findings.

- Top ranked SNPs with the Pearson correlation coefficient give a higher accuracy than lower ranked ones and all SNPs. This is consistent with previous findings on predicting cancer and disease risk with genomic SNPs [50, 109, 110].

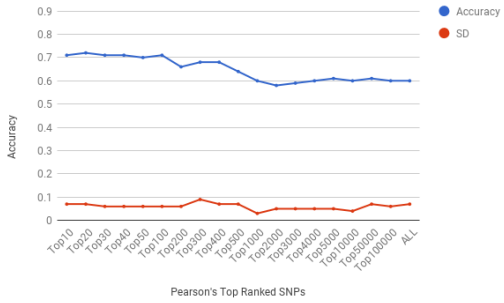
- The soft filtering gives a slightly higher accuracy (reaching 0.76 with 500 SNPs) and fluctuating curve compared to hard and combined filtering.

- The hard and combined filtering achieve their top accuracies of 0.72 and 0.71 with just top 20 and 10 ranked SNPs respectively.

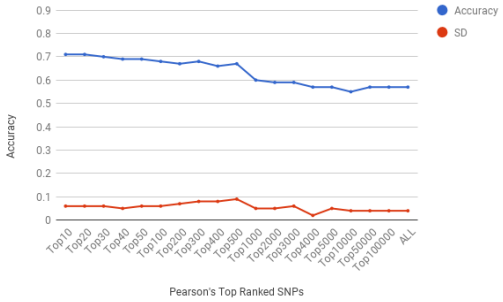
- The combined filtering gives us the most parsimonious model; it achieves its highest accuracy with the fewest number of SNPs (10).



(a) Soft filtering

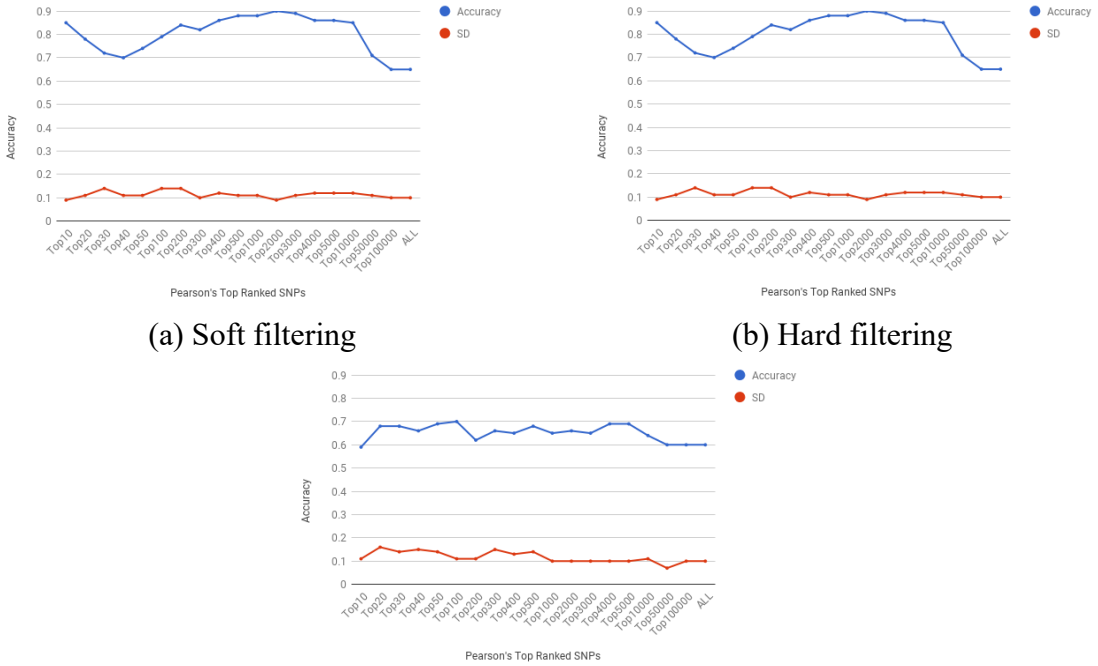


(b) Hard filtering



(c) Soft and hard filtering combined

Figure 4.1 KIRP data three filtering average CV accuracy of SVM on top ranked SNPs.



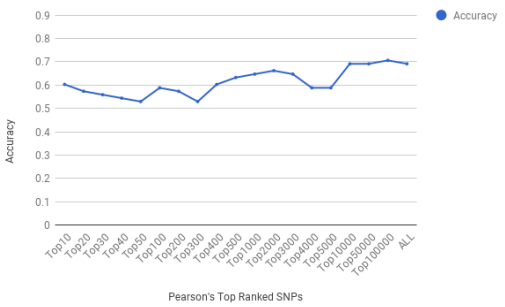
(a) Soft filtering

(b) Hard filtering

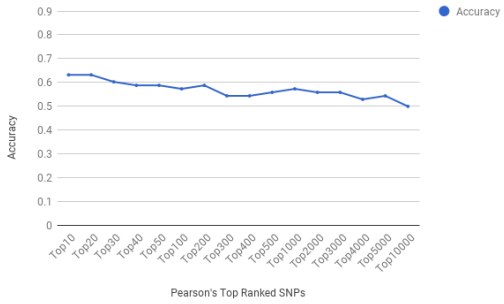
(c) Soft and hard filtering combined

Figure 4.2 KICH data three filtering average CV accuracy of SVM on top ranked SNPs.

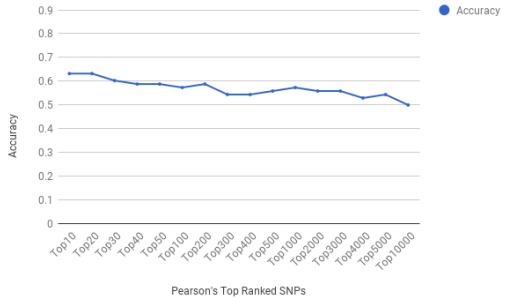
Figure 4.2 depicts the average accuracy of the support vector across 10-fold cross-validation of the KICH dataset. This dataset is less than one third the size of the KIRP dataset, and so it shows different trends. Due to its small sample size, the accuracy fluctuates in all three filtering and peaks equally with a few and many SNPs. This dataset is used primarily as an independent set.



(a) Soft filtering



(b) Hard filtering



(c) Soft and hard filtering combined

Figure 4.3 Accuracy of support vector machine on the KICH dataset after trained on top ranked SNPs in the KIRP dataset.

4.3.2 Cross-Study Validation

For the cross-study validation, the pipeline learns a support vector machine model on the top-ranked SNPs in the KIRP dataset and predicts individuals in the KICH dataset. Figure 4.3 illustrates that the most parsimonious model is given by the soft and hard filtering. There, the accuracy obtained is 0.66 with just 10 SNPs. In comparison the hard filtering peaks at 0.63 with 10 SNPs and soft peaks at 0.7 with 100,000 SNPs.

Table 4.3 Top Ten Ranked SNPs Used in the Cross-Study (Soft+Hard Filtering)

SNP	Ref	Alt	Pearson	Gene/Region	Chromosome
1	A	G,C	0.35	ANO2	12
2	A	C	0.07	ADAMTS9	3
3	G	A	0.07	Non-coding	
4	A	C	0.07	GORAB	1
5	T	C	0.06	NR2C2	3
6	G	A	0.06	SELP	1
7	A	T	0.06	Non-coding	
8	C	T	0.06	LOC100421093	6
9	A	C	0.06	C9orf71	9
10	G	A	0.06	FBXL4	6

While the focus is on the cross-study prediction accuracy, Table 4.3 shows the top 10 ranked SNPs in the cross- study validation. These SNPs are present in both studies, but the ranking is performed on just the KIRP (training) dataset. Most of the SNPs are in coding regions except for two. The SNP in the ANO2 gene has the highest Pearson correlation whereas the others are lower by a large margin. The same SNP is also highly ranked in both of the datasets separately.

The ANO2 gene belongs to the family of anoctamins that are known to be expressed in gastrointestinal stromal tumors and neck and head carcinoma [117]. This gene is known to have a functional role in calcium activated chloride currents [118] but it is unclear how that relates kidney cancer. The ANO1 gene that comes from the same family, however, is known to be expressed in pancreatic cancer [119].

4.3.3 Ranking of Previously Known Kidney Cancer Genes

Table 4.4 shows the ranking of SNPs present in genes previously known to be associated

with kidney cancer [120]. It shows the rankings as well as the Pearson correlation coefficients in the KIRP and KICH datasets separately and the intersection of their SNPs (as in the cross-study). The MET gene is the highest ranked in the KIRP study in this dataset, and is also a drug target for clinical treatment of renal papillary carcinoma [121]. It shows that these genes have low Pearson correlation coefficients indicating that while they are associated with kidney cancer from previous studies their predictive value is limited here.

Table 4.4 Rank of SNPs (PCC) Hard+Soft Filtering in Known Kidney Cancer Genes

Gene	KIRP	KICH	KIRP and KICH
VHL	28296 (5.39e-18)	58117 (1.026e-17)	15831 (5.39e-18)
FH	107511 (0)	6437 (0.07)	35682 (0)
FLCN	15889 (1.88e-17)	2777 (0.096)	13308 (7.26e-18)
MET	1975 (0.026)	20909 (3.86e-17)	799 (0.026)
TSC1	5732 (0.009)	4327 (0.088)	7430 (2.05e-17)
TSC2	16060 (1.85e-17)	4295 (0.088)	7816 (1.85e-17)

4.4 Conclusion and Future Work

The pipeline performs an initial cross-validation and cross-study validation across two kidney cancer datasets obtained from the NCI GDC database. The results show that it can predict kidney chromophobe carcinoma case and controls with 66% accuracy with SNPs learned from a kidney papillary cell carcinoma dataset. More samples from the existing datasets and other datasets from the NCI GDC database is needed to confirm the predictive ability of SNPs in kidney cancer.

CHAPTER 5

MACHINE LEARNING BASED PREDICTION OF GLIOMAS WITH GERMLINE MUTATIONS OBTAINED FROM WHOLE EXOME SEQUENCES FROM TCGA AND 1000 GENOMES PROJECT

5.1 Introduction

Estimating susceptibility to cancer from germline variants is important for recommending regular screening that helps in early cancer detection and enhances patient chances of successful treatment. Linkage analysis studies show that gliomas may cluster within families [122-125]. Also, many genome-wide association studies have identified germline genomic loci that increase glioma risk [7, 126, 127].

This work looks into the collective germline SNPs predictive ability for brain cancer predisposition. The research preforms a Genome Analysis Toolkit (GATK) joint germline SNPs discovery workflow for TCGA Glioblastoma Multiforme (GBM) and Lower-Grade Glioma (LGG) white individual cases and 1000 Genomes Project white individual controls. The SNPs that failed GATK Variant Quality Score Recalibration soft filtering or hard filtering (genotype quality ≤ 20, depth ≤ 5, or missing genotype) quality control were discarded from further machine learning analysis.

On the training set, SNPs with zero variance were excluded and each SNP is scaled so that it remains between zero and one. Then, the best K SNPs were selected based on chi-squared test value. For cross-validation, 1000 Genomes Project, GBM, and LGG samples and their common SNPs were combined. The data were split into 10-fold (90% for training and 10% for testing) and a predictive model was learned with SVM and random forest classifiers. In each training fold, SVM C hyperparameters and the number of trees to grow for RF were cross-validated with 3-fold for each top K selected SNPs.

Then the predictive ability with average balanced accuracy across all folds were measured. For cross-study, the research ran linear SVM on best K selected SNPs on 50% randomly selected samples from 1000 Genomes Project and GBM and predicted LGG and the remaining half of the 1000 Genomes Project samples.

To confirm that all samples came from the same population, principal component analysis (PCA) was performed on the entire dataset (before feature selection), and the first two principal components were projected. Figure 5.3 shows that the two datasets (case and controls) are related. This step is done to confirm that the cases and controls are not separable to limit the effect of ethnicity differences on cases and control classification. SNPs departure from Hardy-Weinberg Equilibrium (HWE) can be a sign of genotyping error or population stratification. Top SNPs in controls that violate HWE are removed from further machine learning analysis. Plink software [128] was used to perform HWE with exact test since using chi-squared test is not suitable for multi-allelic sites.

The research shows that it can predict GBM and LGG white individual cases and 1000 Genomes Project white individual controls with 90% mean balanced accuracy of 10-fold cross-validation (CV) when learning in best 10 germline variants selected by chi-squared value with support vector machine (SVM) and random forest (RF). In cross-study, learning with GBM+controls and predicting LGG achieved 89% balanced accuracy, and 88% balanced accuracy the other way around.

The most contribution to the accuracy comes from SNP rs10792053. When this SNP was removed, cross-validation mean balanced accuracy drops to 54% with top 10 SNPs, and 50% in cross-study. The research looked into the original alignments of SNP

rs10792053 in cases and controls samples with the Integrative Genomics Viewer (IGV). In both cases and controls, reads coverage and mapping quality at this locus were high.

5.2 Methods

5.2.1 Data

For case individuals, white normal samples (germline) whole-exome sequencing (WES) data pre-aligned to Genome Reference Consortium Human Build 38 (GRCh38) in binary alignment map (BAM) format were obtained from The Cancer Genome Atlas (TCGA) through National Cancer Institute's Genomic Data Commons (GDC) portal for two brain cancer studies (males: 477, females: 331, mean age: 52.08). For control individuals, Europeans samples WES pre-aligned to GRCh38 in CRAM format were downloaded from 1000 Genomes Project phase 3 (males: 250, females: 297). This analysis considered only white individuals, to reduce race differences effect on phenotype occurrence. It then performed a variant calling workflow followed by a machine learning pipeline on these samples. Table 5.1 summarizes cohort studies used in this analysis. Table 5.2 shows the number of SNPs for 1000 Genomes Project, GBM, and LGG as well as common SNPs after applying soft+hard filtering.

Table 5.1 Samples Population

Population (sub-population)	Count
1K Genomes Project (CEU)	102
1K Genomes Project (FIN)	105
1K Genomes Project (GBR)	102
1K Genomes Project (IBS)	108
1K Genomes Project (TSI)	112
1K Genomes Europeans (all)	**529**
GBM white (not Hispanic)	274
GBM white (Hispanic)	5
GBM white (not reported)	58
GBM white (all)	**337**
LGG white (not Hispanic)	421
LGG white (Hispanic)	27
LGG white (not reported)	23
LGG white (all)	**471**

5.2.2 Joint Genotyping

For germline variant discovery, the Genome Analysis Toolkit (GATK) version 4 [39] were used. GATK HaplotypeCaller variant calling walker produces an intermediate Genomic Variant Call Format (GVCF) file for each sample. The intermediate GVCF files of all samples were pooled together for genotyping by passing it to GATK genotypeGVCFs to obtain a VCF file for samples cohort. Passing samples GVCFs with the whole-exome regions is computationally intensive, to speed up the variants calling workflow each chromosome is divided into roughly 10 equal intervals in a scatter and gather fashion and were executed simultaneously on a cluster. Figure 5.2 illustrates the

joint variant discovery workflow. After obtaining the final VCF file, quality control measures were applied to reduce sequencing artifacts and false-positive genotypes.

5.2.3 SNPs Encoding

The output of the GATK GenotypeGVCFs tool is in a VCF format. In the header, it has the reference base (REF), one of A, C, G, T, N bases, and alternate non-reference alleles (ALT) base(s). It is possible but not common to have a multiallelic site (two or more ALT bases). All permutations of genotypes were considered as input features to learn a predictive model. An SNP encoding to a numerical value is an essential pre-processing step to machine learning. Each SNP is encoded as follows:

$$4 \times A + B \tag{5.1}$$

where A and B are the two alleles (copies) for a given sample at a particular locus of the genome.

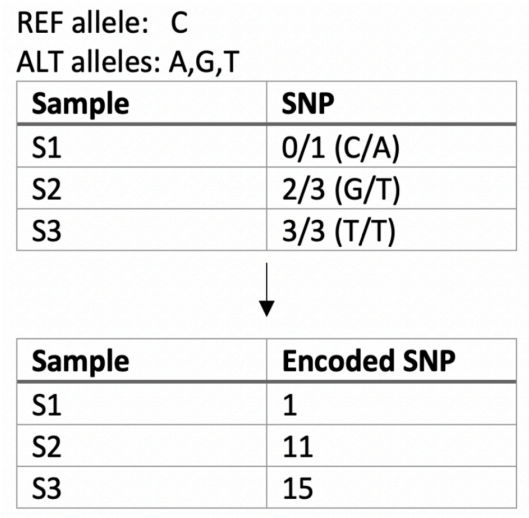


Figure 5.1 A toy example for encoding a multiallelic site.

Table 5.2 SNPs Count after Applying Soft+Hard Filtering

	Number of SNPs
1000 Genomes Project	184690
GBM	297106
LGG	485115
Common SNPs	118439

5.2.4 Missing Genotypes

In GATK, a genotype with low supporting reads is encoded as "./." to denote no variant call was made at that site for a given sample. Imputation is a method that is commonly used in GWA studies to increase the number of genotypes in the as- sociation analysis. Imputation algorithms predict ungenotyped loci in individuals that were genotyped on a subset of loci of SNPs chip to boost SNPs array coverage utilizing haplotype information across samples and HapMap data as an imputation reference panel [129-131]. This study excluded column features that have a missing genotype in any sample from further analysis. Thus, this eliminated the need for imputation.

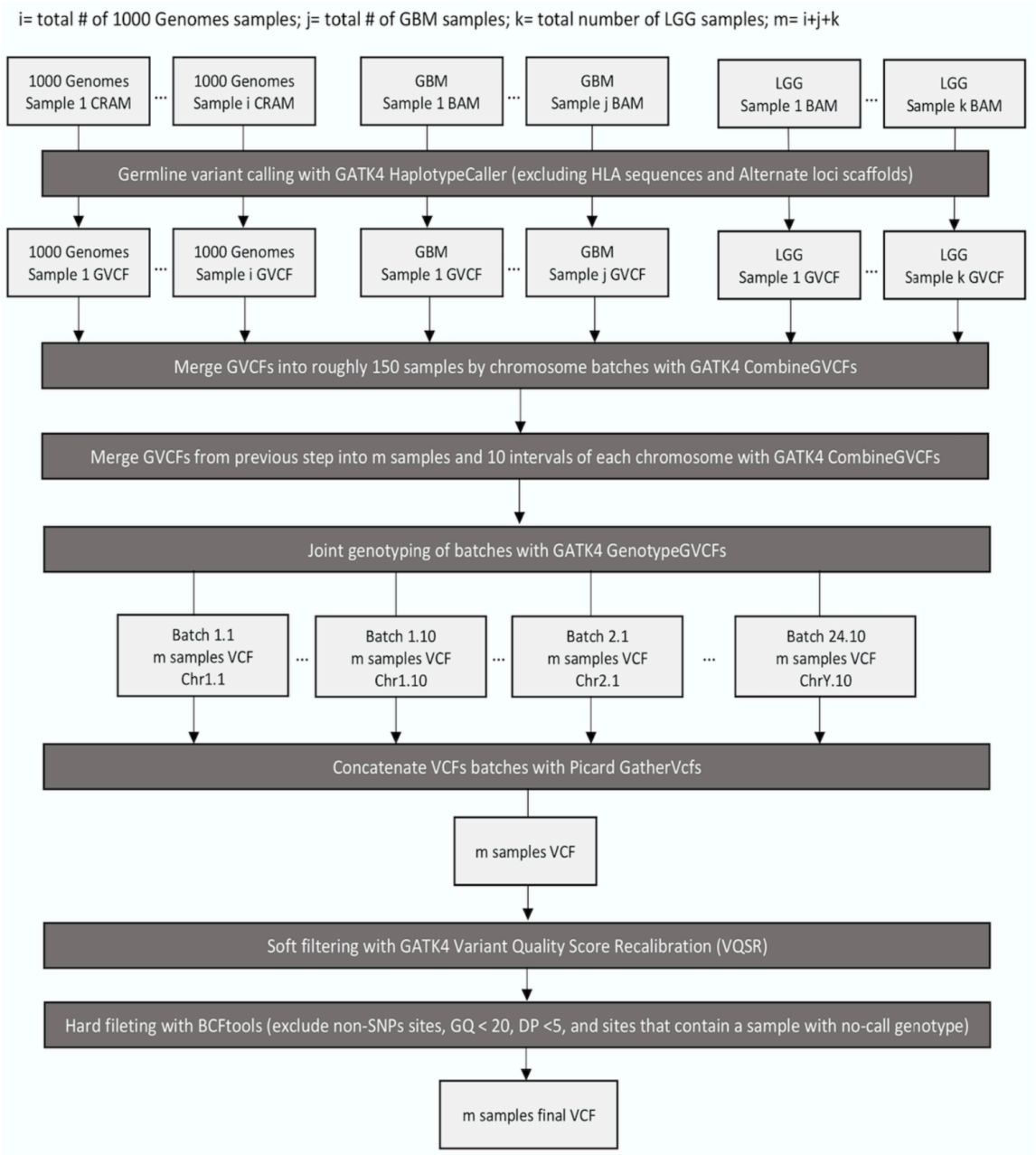


Figure 5.2 Germline SNPs calling pipeline with genome analysis toolkit performed on a cluster to speed up computation.

5.2.5 Variants Calling Quality Control

GATK HaplotypeCaller by default excludes sites with mapping quality (MAPQ) ≤ 20. This analysis used two layers of quality controls: SNPs soft filtering followed by hard filtering to minimize false-positive SNPs. To confirm that the samples came from the

same population, the research ran principal components analysis (PCA) on the whole dataset before SNPs selection. In Figure 5.3, the projection of the first two components shows that the samples are related. Two outlier samples were removed and PCA projections of the first two components were replotted in Figure 5.4, and case and control individuals do not form distinct clusters. Plink version (1.9) [128] were used to test for departure from Hardy Weinberg equilibrium with an exact test in control samples. SNPs that deviate from HWE were excluded. Only top SNPs in HWE are included in the analysis, Table 5.3 shows the exact test p-values of top 10 SNPs in control individuals from 1000 Genomes Project dataset.

Table 5.3 Hardy-Weinberg Equilibrium Exact Test P-Values

SNP	Observed Het	Expected Het	P-Value
rs80356578	0.06049	0.05866	1
rs150707706	0.03592	0.03527	1
rs143139551	0.03214	0.03162	1
rs145172249	0.04159	0.04072	1
rs148782546	0.02268	0.02243	1
rs10792053	0.2042	0.2069	0.6774
rs144518683	0.02268	0.2243	1
rs140561687	0.03025	0.02979	1
rs138772802	0.03403	0.03345	1
rs147042091	0.02836	0.02795	1

Note: Hardy-Weinberg equilibrium exact test P-values on top selected ten SNPs in control individuals from the 1000 Genomes Project.

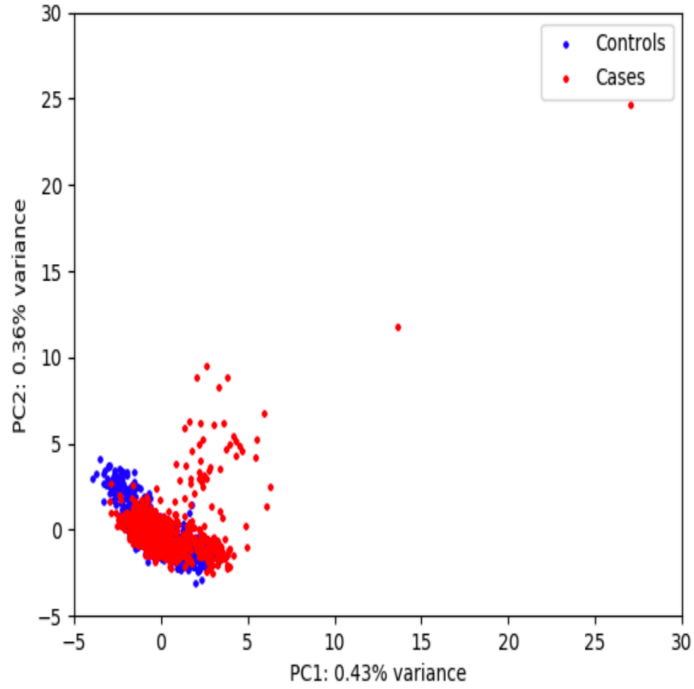


Figure 5.3 Projection of principal component analysis with the first two components.

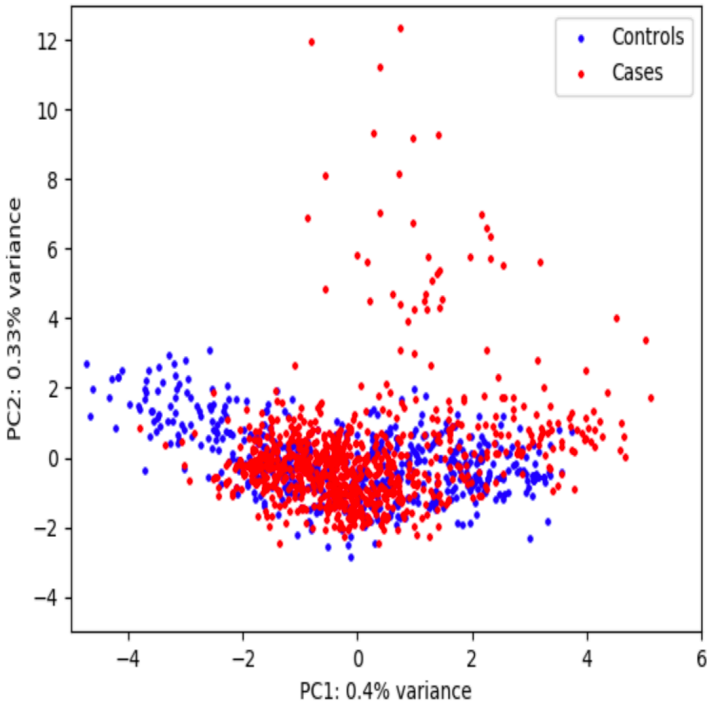


Figure 5.4 Projection of principal component analysis with the first two components after excluding the two outlier data points.

5.2.6 Soft Filtering

The soft filtering method assigns a probability for each variant call with GATK variant quality recalibration score (VQSR) that uses machine learning by training on external databases with known variant sites, and then it assigns a probability score to each variant in the cohort. The truth sensitivity filter for VQSR were set to a 99.0% threshold. The following VCF annotations with VQSR to build a recalibration model were used: InbreedingCoeff, QD, MQ, MQRankSum, ReadPosRankSum, FS, SOR. variants that failed soft filtering are removed from further analysis.

5.2.7 Hard Filtering

At the sample level, variant sites that have genotype quality (GQ) > 20 and depth (DP) > 5 for all samples are considered. DP is the number of reads to support the genotyping, and GQ is a confidence score between 0 and 99, the higher the more confident the program in its assigned genotype. BCFtools (version 1.3) [132] were used for hard filtering and to extract VCF fields into table format.

5.2.8 Soft+Hard Filtering

Only SNPs that passed both soft filtering and hard filtering are considered for further machine learning analysis.

5.2.9 Feature Scaling

Features with zero variance in training split were removed. The remaining features were linearly transformed based on the training subset using Min-Max normalization to keep the data between zero and one while preserving distance. Scikit-learn [116] minMaxScaler were used and the implementation is as follows:

$$x'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \tag{5.2}$$

were x'_{ij} is the current scaled value for the i^{th} individual in the j^{th} SNP, min (x_j) and max (x_j) are the minimum and maximum values for j^{th} SNP, and $\max(x_j) - \min(x_j)$ is the range of the j^{th} SNP. We applied the exact same transformation to validation data where we determined SNPs min and max from training data only.

5.2.10 Chi-Squared Features Selection

Top SNPs are selected based on the chi-squared statistic between each SNP and the label. In the chi-squared test, a higher value is an indicator of dependence between the SNP and the label. SNPs were ranked based on their chi-squared value using the scikit-learn chi2 function.

$$\mathrm{X}^2 = \sum_{i}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{5.3}$$

where n is the number of classes, O_i is the sum of SNP alleles encoding for the i^{th}class. ,$f = \sum_i^n O_i$, and ,$E_i = \frac{1}{n} \times f$.

Table 5.4 shows the top chi-squared ranked 1000 Genomes Project+GBM+LGG common SNPs.

5.2.11 Classifiers

Support vector machine (SVM) with linear kernel [81] and random forest (RF) [133] classifiers were performed using scikit-learn package [116].

Support vector machine: SVM finds a hyperplane that maximizes the distance between classes:

$$\min_{w,w_0} \frac{\|w\|^2}{2} + C \max(0, 1 - y_i(w^T x_i + w_0)) \tag{5.3}$$

where x_i is the genotype vector of the i^{th} individual, y_i is the label, w is the weight vector, C is a regularization parameter, max $(0, 1 - y_i(w^T x_i + w_0))$ is the hinge loss and the sign of $(w^T x_i + w_0)$ assigns the input x into class -1 or $+1$. We cross-validated the hyperparameter C with 3-fold cross-validation from the set (0.1, 1).

Random forest: An ensemble method that builds decision trees by selecting random samples with replacement to construct each tree and randomly generating a subset of features to choose from for each candidate split, the one with the highest Gini impurity or entropy, then it takes the majority vote of trees predictions to output a class prediction. We used the default parameters for the quality measure of the split, and 3-fold cross-validation from the set (100, 1000) for the number of trees to construct.

5.2.12 Performance Metrics

Since classes are imbalanced in the studies included in our analysis, it is inappropriate to use accuracy as a measure of classifiers performance. We used balanced accuracy for performance evaluation. Balanced accuracy is the average of true positive rate and true negative rate.

$$Balanced\ accuracy = \frac{\left(\frac{true\ positive}{positive} + \frac{true\ negative}{negative}\right)}{2} \tag{5.3}$$

Table 5.4 Top SNPs for 1000 Genomes Project, GBM and LGG Datasets

Alt allele frequency			
1K Genomes	**GBM+LGG**	**SNP rs ID**	**Chi2 score**
0.0302	0.0068	rs80356578	21.84
0.0180	0.0019	rs150707706	20.15
0.0161	0.0006	rs143139551	22.67
0.0208	0.0006	rs145172249	30.26
0.0113	0	rs148782546	18.33
0.1096	0.4963	rs10792053	50.63
0.0113	0	rs144518683	18.33
0.0151	0	rs140561687	24.44
0.0170	0	rs138772802	27.49
0.0142	0.0006	rs147042091	19.65

Table 5.5 Top SNPs for 1000 Genomes Project and GBM Datasets

Alt allele frequency			
1K Genomes	**GBM**	**SNP rs ID**	**Chi2 score**
0.0047	0.0237	rs140717526	12.28
0.0038	0.0341	rs782010133	12.78
0.0076	0.0312	rs779492064	13.69
0	0.0134	rs202040378	14.13
0.0009	0.0148	rs146032550	12.51
0.0019	0.0386	rs759512484	34.27
0.0009	0.0163	rs76672487	14.05
0.0009	0.0148	rs148088117	12.51
0.1096	0.4926	rs10792053	42.67
0.0019	0.0341	rs768904765	24.25

Table 5.6 Top SNPs for 1000 Genomes Project and LGG Datasets

Alt allele frequency			
1K Genomes	**LGG**	**SNP rs ID**	**Chi2 score**
0.0302	0.0074	rs80356578	13.30
0.0076	0.0308	rs12721607	14.53
0.0009	0.0159	rs35723440	13.97
0.0208	0	rs145172249	19.59
0.0076	0.0297	rs2232449	13.60
0.0236	0.0032	rs61734485	14.88
0.0085	0.0329	rs2069548	14.84
0.1096	0.4989	rs10792053	45.18
0.0151	0	rs140561687	14.25
0.0170	0	rs138772802	16.03

5.3 Results

5.3.1 Cross-Validation

With chi-squared statistic best ten SNPs, linear SVM and RF achieved 90% mean balanced accuracy of 10-fold cross-validation when predicting the 1000 Genomes Project controls and GBM+LGG cases. The predictive ability deteriorates when all SNPs were considered to 65% and 54% for SVM and RF, respectively. Figure 5.6 shows the results for predicting three-classes of 1000 Genomes, GBM, and LGG with linear SVM, one-vs-one. The mean balanced accuracy attained is 68% on top 10 SNPs, however, the accuracy drops to 46% with all SNPs. The accuracy declines as more SNPs are added in both binary and three-class classification of glioma subtypes individuals and control individuals.

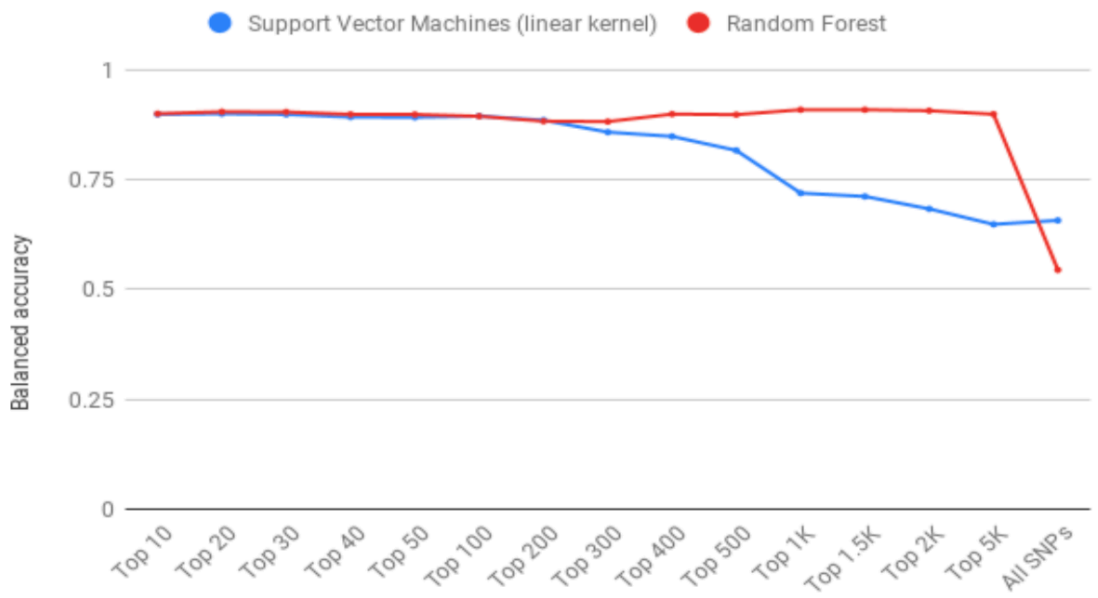


Figure 5.5 10-fold cross-validation of learning and classifying binary labels.

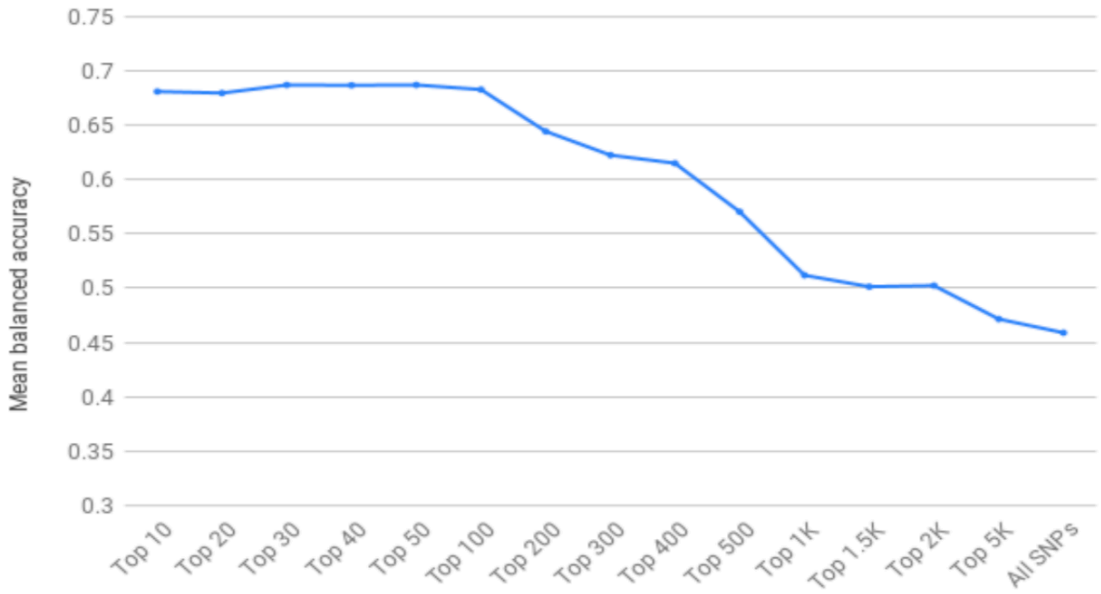


Figure 5.6 10-fold cross-validation of learning and classifying 3-class labels.

5.3.2 Cross-Study Validation

To test the generalization of the model, the research trained the data on GBM and randomly selected 50% of 1000 Genomes samples and predicted the labels of the unseen LGG dataset and the remaining 50% of 1000 Genomes samples. Top 10 ranked SNPs obtained the highest balanced accuracy of 89%, again the advantage of ranking the SNPs with chi-squared is shown, the accuracy drops when more SNPs are included to learn a model. The worst accuracy of 63% was attained by considering all SNPs. The research also tested the accuracy the other way around, where it learned in LGG and 50% randomly selected samples from 1000 Genomes and predict the labels of GBM and the remaining 50% samples of 1000 Genomes. It observed the same thing, where ranking the SNPs with chi-squared boost the balanced accuracy from 60% with all SNPs to 88% with only 10 SNPs. As expected, ranking SNPs by their dependence on labels improved the balanced accuracy greatly on all cross-validation and cross-study validation experiments.

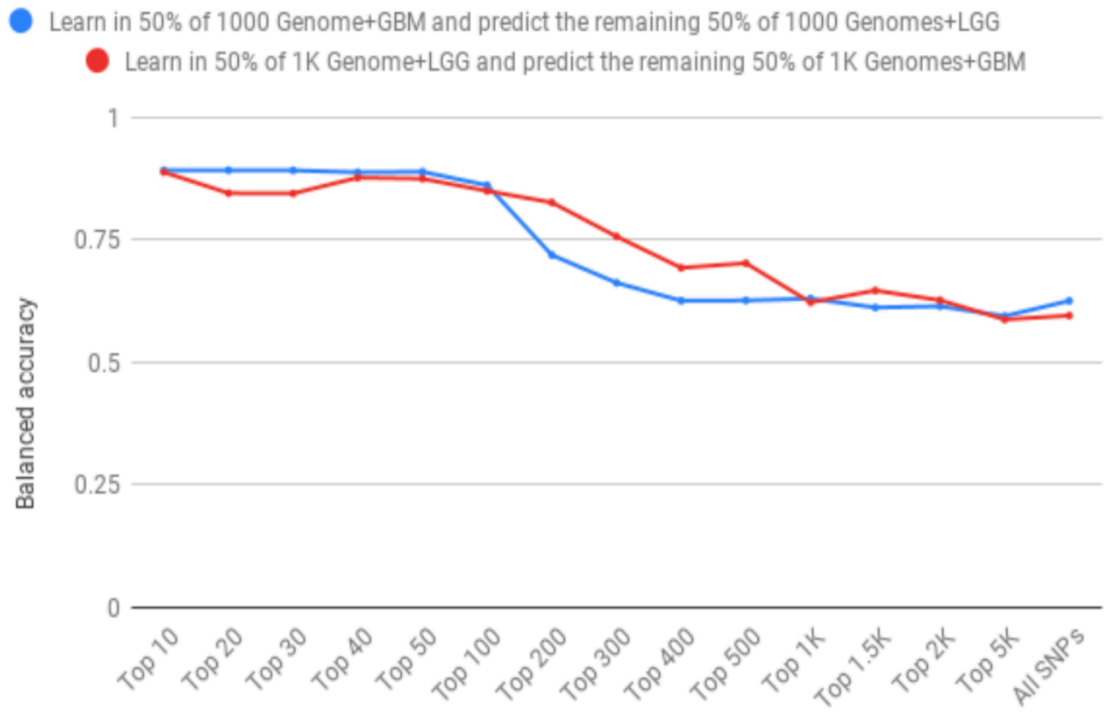


Figure 5.7 Cross-study validation.

5.3.3 Cancer Significance of Top Ranked SNPs

A point mutation could be nonsynonymous (missense, or nonsense) or synonymous (silent). Missense mutations, which is a change in a single nucleotide that substitutes amino acid encoding and influences protein function [134, 135], are heavily investigated in cancer research because it can alter protein function. Synonymous mutations are often called silent mutations due to their inability to change the amino acid sequence, therefore, these mutations usually are disregarded in cancer research [135]. However, synonymous variants can affect protein folding, and thus it plays a role in cancer [136]. In this work, we investigated both synonymous and nonsynonymous variants. SNPs rs76672487 (in gene ABCC2) and rs2069548 (in gene TG) are cancer-related genes according to The Human Atlas Protein. SNP rs76672487 ranked fifth on the selected SNPs by chi-squared

from the GBM+1000 Genomes dataset, while SNP rs2069548 ranked fourth on GBM+1000 Genomes top SNPs. From the 1000 Genomes+GBM+LGG datasets' top ten ranked SNPs six genes are reported by The Human Atlas Protein to be prognostic markers for survival in glioma, liver, renal, cervical, urothelial, pancreatic, and endometrial cancers based on gene expression FPKM values.

Table 5.7 All Datasets Genes Expression and Survival Time Association

Gene	Survival prognostic marker in cancer
OTOF	No
EAF2	Prognostic marker.
ALPK1	Prognostic marker.
LOC108783645, HFE	No
PTPRJ	Prognostic marker.
OR9G1	No
P4HA3	Prognostic marker.
ATF7IP	Prognostic marker.
PLBD1	Prognostic marker.
KCNC2	No

Note: top ranked SNPs genes expression and survival time association in the 1000 Genomes Project, LGG, and GBM datasets based on The Human Atlas Protein database.

Tables 5.7 through 5.9 show top-ranked genes in the 1000 Genomes+GBM+LGG, 1000 Genomes+GBM, and 1000 Genomes+LGG datasets that are prognostic for survival time in cancer. Five genes of the top-ranked in the 1000 Genomes+GBM+LGG dataset are expressed in all cancers according to The Human Atlas Protein. KCNC2 gene is expressed in breast and prostate cancers. P4HA3 gene is expressed in pancreatic, breast, renal, glioma, and lung cancers. Genes OR9G1 and OTOF are not expressed in cancer. Tables 5.10 through 5.12 show the genes that are expressed in cancer in top-ranked SNPs

in the 1000 Genomes+GBM+LGG, 1000 Genomes+GBM, and 1000 Genomes+LGG datasets.

Table 5.8 1000 Genomes and GBM Genes Expression and Survival Time Association

Gene	Survival prognostic marker in cancer
SARS	Prognostic marker
CA14	No
LHX9	No
DGKG	No
OSMR	Prognostic marker.
DMXL1	Prognostic marker.
ABCC2	Prognostic marker.
OR56B4	No
OR9G1	No
ZNF641	Prognostic marker.

Note: top ten ranked SNPs genes expression and survival time association based on The Human Protein Atlas database.

Table 5.9 1000 Genomes and LGG Genes Expression and Survival Time Association

Gene	Survival prognostic marker in cancer significance (P<0.001)
OTOF	No
NR1I2	No
IGSF10	No
LOC108783645, HFE	No
MICAL1, ZBTB24	Prognostic marker.
CA1	No
TG	No
OR9G1	No
ATF7IP	Prognostic marker.
PLBD1	Prognostic marker.

Note: Top ten ranked SNPs in the 1000 Genomes Project and LGG and their genes expression and survival time association as reported by The Human Protein Atlas database.

Table 5.10 Top SNPs in All Datasets Genes and Functional Consequences

SNP rs ID	Gene	Functional consequence	Cancer mRNA expression
rs80356578	OTOF	synonymous	Not detected
rs150707706	EAF2	missense	Expressed in all
rs143139551	ALPK1	missense	Expressed in all
rs145172249	HFE	intron variant	Expressed in all
rs148782546	PTPRJ	synonymous	Expressed in all
rs10792053	OR9G1	synonymous	Not detected
rs144518683	P4HA3	synonymous	Mixed
rs140561687	ATF7IP	missense	Expressed in all
rs138772802	PLBD1	intron	Expressed in all
rs147042091	KCNC2	missense	Group enriched

Note: Top ten SNPs and their genes functional consequences in the 1000 Genomes project, GBM, and LGG datasets as reported by The Human Protein Atlas database.

Table 5.11 Top SNPs of 1000 Genomes and GBM Genes and Functional Consequences

SNP rs ID	Gene	Functional consequence	Cancer mRNA expression
rs140717526	SARS	missense	Expressed in all
rs782010133	CA14	missense	Group enriched
rs779492064	LHX9	intron	Mixed
rs202040378	DGKG	intron	Tissue enhanced
rs146032550	OSMR	synonymous	Expressed in all
rs759512484	DMXL1	missense	Expressed in all
rs76672487	ABCC2	intron	Tissue enhanced
rs148088117	OR56B4	missense	Not detected
rs10792053	OR9G1	synonymous	Not detected
rs768904765	ZNF641	intron	Expressed in all

Table 5.12 Top SNPs of 1000 Genomes and LGG Genes and Functional Consequences

rs ID	Gene	Functional consequence	Cancer mRNA expression
rs80356578	OTOF	synonymous	Not detected
rs12721607	NR1I2	missense	Group enriched
rs35723440	IGSF10	synonymous	Mixed
rs145172249	HFE	intron	Expressed in all
rs2232449	ZBTB24	synonymous	Expressed in all
rs61734485	CA1	missense	Group enriched
rs2069548	TG	missense	Tissue enriched
rs10792053	OR9G1	synonymous	Not detected
rs140561687	ATF7IP	missense	Expressed in all
rs138772802	PLBD1	intron	Expressed in all

5.3.4 SNP rs10792053 Mapping Quality

To confirm that there is no issue with reads mapping quality or coverage, eight individuals from both cases and controls alignments were inspected with the Integrative Genomics Viewer (IGV) in the original reads mapping at locus 11:56701017 and its adjacent loci. Figure 5.8 shows alignments with IGV of four samples from cases vs four from controls against the GRCh38 reference genome. The red arrow in Figure 5.8 points SNP rs10792053 position. The tangerine color in the tracks at the position refers to allele C and the green refers to reference allele A.

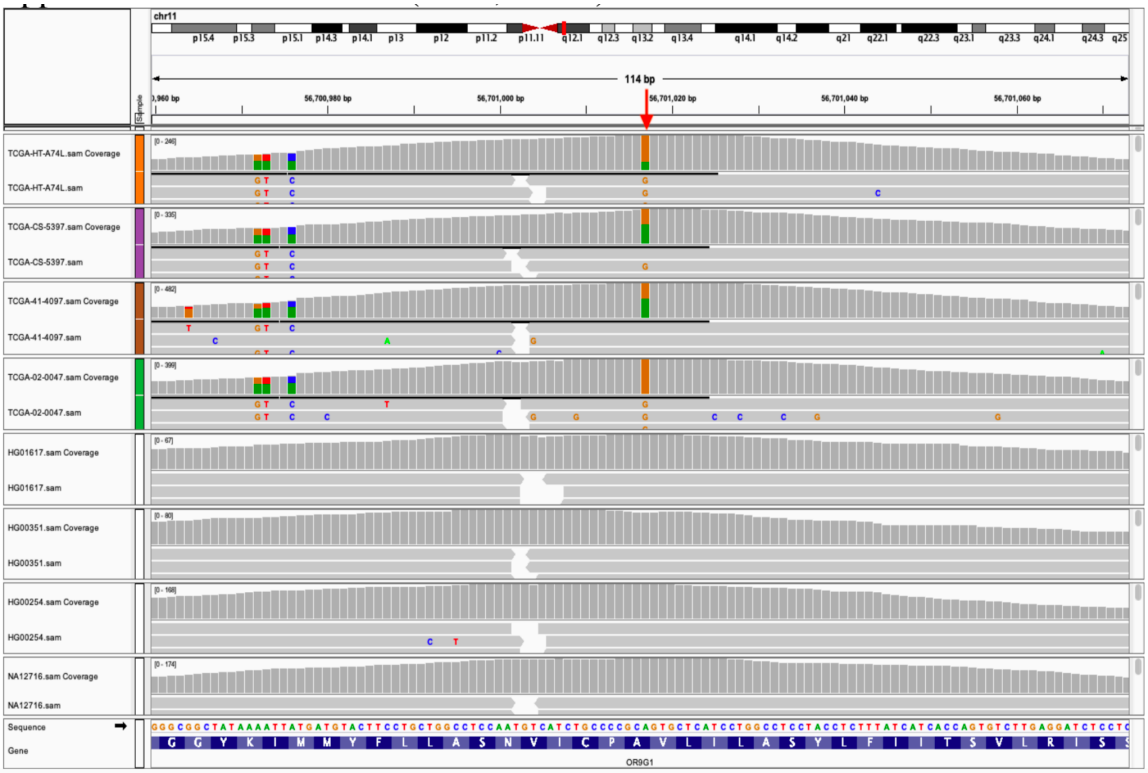


Figure 5.8 Alignments of four cases vs four controls at SNP rs10792053 the upper four tracks for cases (LGG, GBM) viewed with IGV.

In IGV, if both allele copies in the sample is homozygous reference, then it is shown in gray. Three of the four cases viewed are heterozygous and the remaining one is homozygous alternate allele. All controls in Figure 5.8 are homozygous reference. In IGV the mapping quality threshold were set to 1 since GATK HaplotypeCaller discards reads with a mapping quality of 0. The original alignments of both cases and controls have high coverage at this location. Although GATK HaplotypeCaller reassembles alignments at active regions and discards original alignments, the final VCF is consistent with what is observed in original alignments. For SNP rs10792053, the average depth across all cases is 407.15 and across all controls is 63.58. These average depths are after running the GATK germline variant discovery workflow. The research tested for Hardy Weinberg equilibrium exact test in controls individual and the p-value is 0.677, which

confirms that this SNP is in HWE, however, it is out of HWE in cases.

5.3.5 Alternate Allele Frequency of Top SNPs

Table 5.13 shows the alternate allele frequency of dbSNP 1000 Genomes Project Europeans samples, GBM, LGG and 1000 Genomes Project samples that are considered in this study, which is slightly larger than 1000 Genomes sample size in dbSNPs since this study downloaded samples from 1000 Genomes Project phase 3.

Table 5.13 Cases and Controls Top Ranked SNPs Alternate Allele Frequencies

rs ID	dbSNP (EUR)	Controls	Cases
rs80356578	A=0.029	0.0302	0.0068
rs150707706	C=0.019	0.0179	0.0018
rs143139551	A=0.017	0.0160	0.0006
rs145172249	C=0.019	0.0207	0.0006
rs148782546	T=0.012	0.0113	0
rs10792053	G=0.116	0.1096	0.4962
rs144518683	C=0.012	0.0113	0
rs140561687	T=0.016	0.0151	0
rs138772802	C=0.017	0.0170	0
rs147042091	C=0.013	0.0141	0.0006

Note: 1000 Genomes Project (Controls) and GBM+LGG (Cases). The third and fourth columns contain the alternate allele frequencies in this research (white individuals). The second column shows the reported allele frequencies by dbSNP (European samples) database.

For example, SNP rs80356578 sample size in dbSNP is 503 and the sample size for our 1K Genomes is 526. Our alternate allele frequency is close to what is reported by dbSNP for 1000 Genomes Project Europeans samples.

5.4 Conclusion

This chapter shows that it can predict glioma cases with few germline SNPs selected based on the chi-squared statistics with 90% mean balanced accuracy in cross-validated TCGA-GBM and TCGA-LGG white individual cases and 1000 Genomes Project Europeans controls whole-exome sequences with linear SVM and random forest classifiers. The chapter also shows that in cross-study linear SVM achieves 89% predictive accuracy when learning with GBM and 1000 Genomes Project controls and predicting LGG and 88% contrariwise on the top-ranked germline SNPs. However, most of the accuracy comes from SNP rs10792053, a replication study is needed to verify its discriminative power in glioma.

CHAPTER 6

CHALLENGES IN PREDICTING GLIOMA SURVIVAL TIME IN MULTI-MODAL DEEP NETWORKS

6.1 Introduction

Predicting glioma survival time helps patients and their clinicians evaluate available treatment plans and make informed choices. Glioblastoma Multiforme (GBM) is the most common and lethal glioma type in adults [137]. In GBM, less than 5% of patients reach five years survival threshold after diagnosis with a median survival time of 15 months [138]. Most advanced cancer patients prefer to know their estimated prognostic information [11]. However, clinicians' survival time estimates are inaccurate and often optimistic [11, 12].

Many studies have devised 3D convolutional neural networks (CNNs) to improve the accuracy of structural MRI scans to classify glioma patients into survival categories [3, 13-15]. This work looks into a heterogeneous combination of somatic and germline genetic single variations, messenger RNA expressions, and post-contrast T1 MRI modality data that show the malignancy. Whole exome sequencing data (WES) were obtained from The Cancer Genome Atlas (TCGA) portal (https://www.cancer.gov/tcga), messenger RNA, and post-contrast axial T1 MRI sequences from The Cancer Imaging Archive (TCIA [139]) for all European ancestry individuals with GBM. In this analysis, only samples for which all three data types are available were included, which gives a total of 126 samples. Each sample is assigned a label of zero if the survival time is below14 months and a label of one otherwise (to obtain a balanced set), thus converting the survival time prediction problem into a classification one.

This work designed a multi-path neural network that takes as input all three data sources and evaluates its accuracy in a 10-fold cross-validation experiment. Genome Analysis Toolkit (GATK4) pipeline was performed to obtain single mutations with exhaustive site-level and sample-level quality controls to eliminate sequencers artifacts and false-positive SNPs. Both and multiallelic loci were included, and the two allele copies of each SNP were converted into a numerical format using an in-house python script. Then, SNPs were ranked on each training split to select the best 100 SNPs to use as predictive markers. The mRNA expression information for the TCGA-GBM was obtained from the Broad Institute TCGA Genome Data Analysis Center Firehose after Robust Multi-array Analysis (RMA) normalization.

MRI sequences in Digital Imaging and Communications in Medicine (DICOM) format were downloaded from TCIA. From the 3D axial T1 MRI sequences. Both 3D volumes and 2D slices were explored. For 3D scans, the DICOM images were converted to the Neuroimaging Informatics Technology Initiative (NIfTI) format. The non-brain tissue was extracted with FSL BET, and the images were registered to T1 MRI MNI152 reference with FSL FLIRT. A model was trained with 3D U- Net [140] separately as well as simultaneously with SNPs and mRNA data.

For 2D slices, one slice that shows the tumor for each sample was manually selected. Then these 2-D image slices were used to train a 2D CNN with ResNet18 [141] encoder and measured the accuracy of predicting test samples in 10-fold cross-validation. This study compared the accuracy of predicting survival time with SNPs, mRNA expressions, and MRI scans separately as well as when combining the three data sources. For SNPs and mRNA expressions, separate multi-layer neural nets were used, and for

images, 2D and 3D convolutional neural networks were explored.

A slight improvement within combined model with 2D images is observed over the individual data sources but considerable variation in test accuracy across different train test folds. This work conjecture that this may be due to the small training set of 126 individuals. By synthetically augmenting the data with a generative model, this research may improve sample representation and consequently model accuracy.

6.2 Methods

6.2.1 Data

Data is composed of TCGA-GBM European ancestry individuals (https://www.cancer.gov/tcga) that have all of the following data: 1) Survival time (days from diagnosis to death), also, right censoring to increase the dataset size was performed, where samples for which days to the last follow-up are above the 14 months threshold were included, 2) WES data, 3) mRNA expressions information, and 4) post-contrast T1 axial MRI sequence. Samples that do not meet the inclusion criteria were excluded. The total number of samples included in the analysis is 126. Table 6.1 shows the clinical characteristic of these samples.

Table 6.1 Samples Clinical Characteristics

Clinical Characteristics	TCGA-GBM
Ancestry (European)	126
Ethnicity (not reported/not Hispanic)	25/101
Gender (male/female)	76/51
Average age	60.38 ± 13.37
Vital status (dead/alive)	123/3
Average survival (days)	483.44 ± 431.95
# of samples in each class (short-term/long-term)	63/63

SNPs: TCGA-GBM 126 European ancestry individuals pre-aligned WES for each sample that met the inclusion criteria were obtained from the TCGA (https://www.cancer.gov/tcga) through the NCI Genomic Data Commons (GDC) data portal (https://gdc.cancer.gov/). GATK (version 4) HaplotypeCaller [38, 39, 76] was performed on each sample. All samples were then pooled together for joint genotyping utilizing a computing cluster in a scatter and gather approach on each chromosome to expedite variant discovery process. To filter out low-quality SNPs, the GATK variant quality recalibration score (VQSR) was performed, which uses a machine learning trained on external datasets to assign a quality score to each site-level variant. The truth sensitivity of 99% as a threshold is used for VQSR. Those SNPs that passed VQSR are further interrogated on sample- level genotype quality (GQ) and depth (DP). SNPs that passed VQSR at the site-level and GQ > 20 and DP ≥ 5 at the sample-level are included for further analysis. This work performed the widely used multi-allelic encoding of SNPs shown in Figure 6.1.

The chi-squared statistic [142] between each SNP and the binary class label was calculated, and SNPs were ranked based on the test statistics. The higher the statistic, the

more important the SNP in its predictive ability. We included top-ranked 100 SNPs for further analysis.

mRNA expressions: Gene expression information was downloaded for the samples that were normalized with Robust Multi-array Analysis (RMA) from the Broad Institute TCGA Genome Data Analysis Center Firehose [143].

3D MRI scans: Axial T1 MRI sequences in DICOM format were obtained from The Cancer Imaging Archive (TCIA). DICOM images were converted to NIfTI format with dcmtonii software, and non-brain tissue was removed with FSL BET [144] with option -B (an option that leads to overall better performance in skull-stripping [145]). Then images were aligned to T1 axial MNI152 reference with FSL FLIRT.

2D MRI slices: For each subject, an image slice was manually selected that best shows the tumor and its surrounding tumor enhancing-area. Table 6.2 shows the vector and matrix dimensions for the three data sources that were used in this analysis.

Table 6.2 SNP, mRNA, and T1 MRI Data

Dataset	Vector (matrix) dimension
SNPs (passed filtering)	79980
mRNA expressions	12042
3D post-contrast T1 MRI scans	(182, 218, 182)
2D post-contrast T1 MRI slices	(256,256)

REF allele: C
ALT alleles: A,G,T

Sample	SNP
S1	0/1 (C/A)
S2	2/3 (G/T)
S3	3/3 (T/T)

↓

Sample	Encoded SNP
S1	1
S2	11
S3	15

Figure 6.1 Multiallelic SNP encoding into a numerical values example.

6.2.2 Network Architecture and Training

Three separate neural nets were constructed using PyTorch [146] to evaluate the predictive power of each data source alone. the output of these nets was concatenated to evaluate combining SNPs, mRNA expressions, and images predictive ability.

For SNPs, a neural net with two hidden layers was constructed. Relu activation function, 0.01 learning rate, batch size of five, and 30 epochs with Stochastic Gradient Descent (SGD) [147] and Nesterov Momentum update. For mRNA expressions, the parameters were set to exactly what was used for SNPs but with three hidden layers (1000, 100, 10). For 2D T1 MRI sequence slices, ResNet18 convolutional neural network [141] was used which has 18 hidden layers and has 18 output nodes. Because the ResNet18 input size shape is (244, 244), all slice images were resized to (256, 256) dimensions and randomly center cropped (224, 224), the cropped images were used as an input for the ResNet18 convolutional neural network. The following parameters with ResNet18 were used: learning rate of 0.01, batch size of 6, 15 epochs. For 3D volumes,

the 3D U-Net [140] where employed, and the original images were padded with zero to fit the network input dimensions (192, 224, 192). The same hyperparameters that were used to train the 2D slices in ResNet18 were used for the 3D U-net.

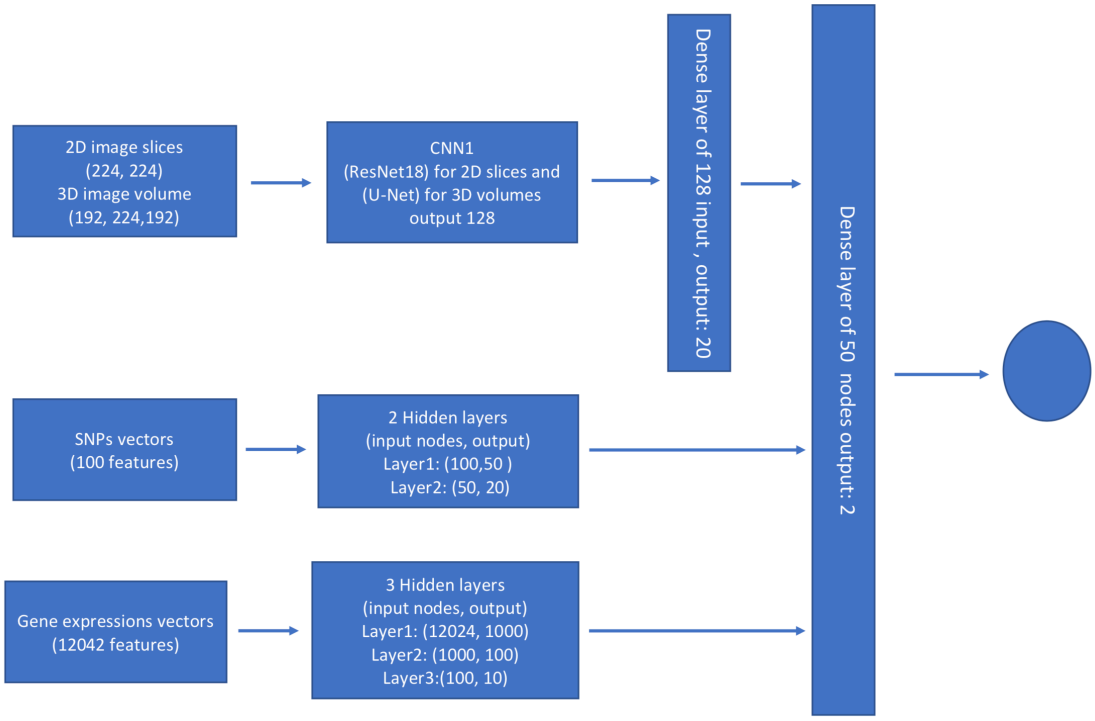


Figure 6.2 Proposed multi-modal deep neural network. One can see three paths each for SNP, gene expression, and images. The study trains the network as one model instead of training the three paths separately.

To combine each of SNPs and mRNA dataset with MRI slices, a one more dense layer was added to the end of ResNet18. After Relu activation, the output was concatenated to the network's output. Then, it was fed into a dense layer with 50 input nodes and two output nodes. For combining the three data sources, all the three outputs of each network were concatenated. Figure 6.2 shows the network architecture for combined data sources.

6.3 Results

Here the 10-fold cross-validation results on all three data sources combined as well as individual data sources with both 2D and 3D images is reported. The accuracy is evaluated as the sum of correct predictions over the total number of the test set. Survival threshold at 14 months was intentionally selected so that the data is balanced: the number of samples in both classes is equal.

6.3.1 Combined Data with 3D Volumes

Figure 6.3 shows the mean accuracy of our model across 10-folds and 15 epochs for each of the three data sources separately and the combined data model. The model with combined data sources can achieve a 100% accuracy on the individual and combined data models. In the test, however, the accuracies are lower. The combined data model does not perform better than the individual ones. In fact, here the gene expression data source gives the best test accuracy of 62.4% at epoch 13.

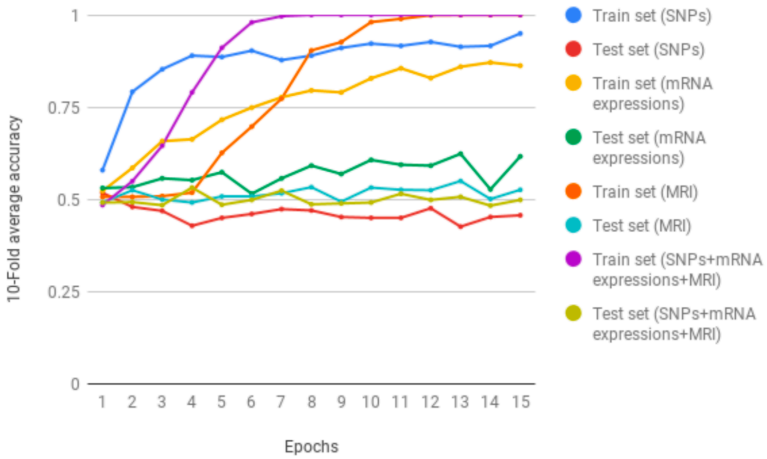


Figure 6.3 Mean 10-fold accuracy of our network across 15 epochs for training and test sets with 3D volumes as the image data.

Figure 6.4 shows the test accuracy of each of the ten folds of the combined model. In some folds, the test accuracy goes to 75%, whereas in others as low as 25%. This suggests that some folds have a diverse enough training set that captures the distribution of test data points, whereas, in other folds, the training and test image datasets are very different.

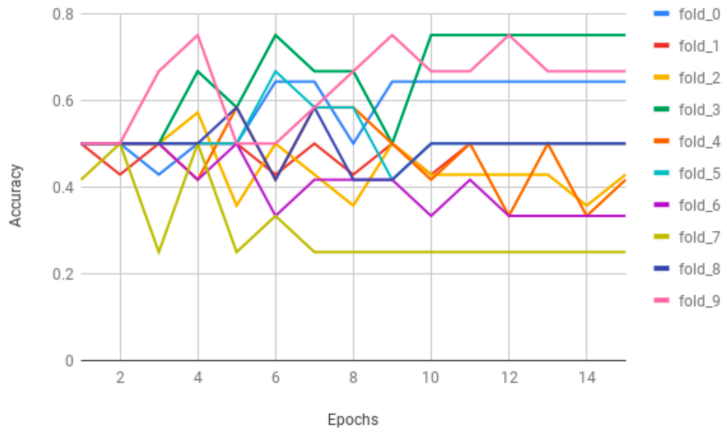


Figure 6.4 Test accuracy of each of the 10-folds of our network across 15 epochs on all three data sources combined with 3D volumes as the images.

6.3.2 Combined Data with 2D Slices

Figure 6.5 shows the mean 10-fold accuracy on training and test sets across 15 epochs of the model with 2D slices as images. Here one can see an overall better test accuracy with the combined model but by a small margin. At epoch 11, the combined data gives 63% accuracy, whereas the gene expression alone gives 62.4% at epoch 13. This difference, however, is not statistically significant.

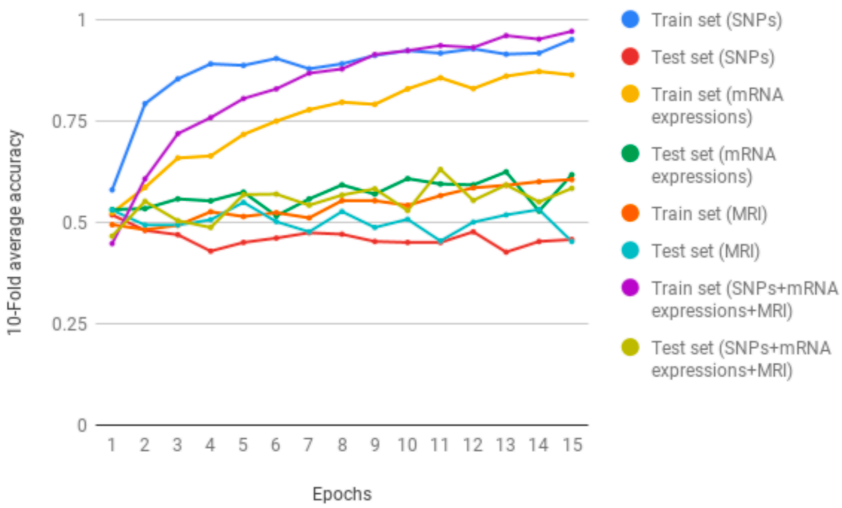


Figure 6.5 Mean 10-fold accuracy of our network across 15 epochs for training and test sets with handpicked 2D slices (that manifest the tumor) as the image data).

The test accuracy of each of the ten folds on the combined data, one can see a considerable variation, as shown in Figure 6.4. Again, this shows that in some folds, the train and test distributions are likely to be the same, while in others, the distribution is very different, thus making it hard to classify.

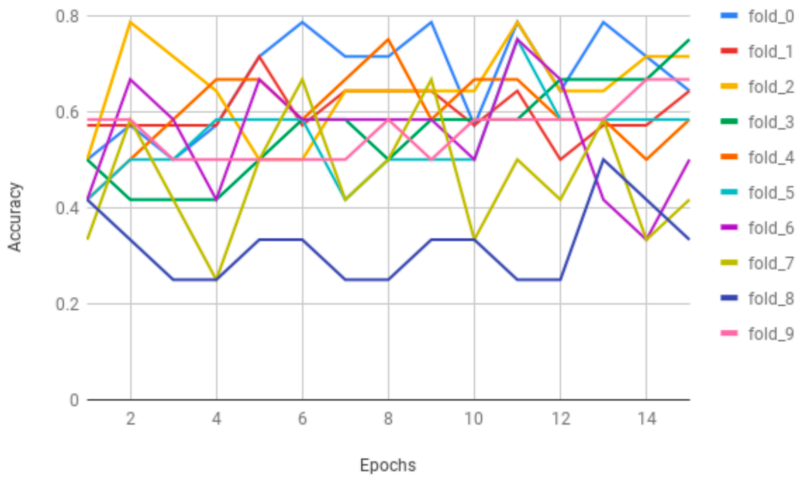


Figure 6.6 Test accuracy of each of the 10-folds of our network across 15 epochs on all three data sources combined with 2D slices as the images.

6.4 Discussion

Combining data into three sources yields a slight improvement that is not statistically significant. Clearly, by simply combining more data, one cannot expect to predict survival time accurately, but this study also possibly needs to enlarge the training set size. The variation in test accuracy across the folds suggests model instability, which is attributed to insufficient data. One possible avenue to solve this is to generate artificial samples for all three sources with a generative model like a generative adversarial network [148].

Another thing in the results is the 100% accuracy in training on the combined data in both 2D and 3D. Could the model be overfitting? A dropout [149], which is a popular and powerful method to reduce overfitting, was added. It reduces the training accuracy all the way down to in the 50-60% range and does not improve test accuracy. This suggests we may need a richer model with dropout since even fitting training samples becomes very hard with this method.

Finally, one can see that the 2D combined model performs better than the 3D. A 3D model, in general, requires much more data than a 2D, which is one likely reason for the 3D model's poorer performance. The 3D U-Net was fine-tuned, but it did not improve accuracy. Again, the research conjecture that additional data points via generative modeling may increase accuracy.

6.5 Conclusion

Integrating genomic and neuroimages in a multi-path neural network slightly improve glioma survival time prediction at the 14-month threshold. One can see instability in test

accuracy in the model, and this study conjecture that a larger sample size produced via a generative model may improve stability and overall accuracy.

CHAPTER 7

MULTI-PATH CONVOLUTIONAL NEURAL NETWORK FOR GLIOBLASTOMA SURVIVAL GROUP PREDICTION WITH POINT MUTATIONS AND DEMOGRAPHIC FEATURES

7.1 Introduction

Glioblastoma multiforme (GBM) is the most common and aggressive type of brain cancer, with a median survival rate of 15 months [10]. Untreated patients with GBM have a median survival time of 3 months [150]. It is well-established that age is a strong independent predictor of survival time in gliomas [151-153]. Several studies have found that gender is significantly correlated [154-156]. A study that analyzed 6586 GBM patients shows that age and gender, among other seven features, are independent survival prognostic factors [157]. Other studies investigated the role of Single Nucleotide Polymorphisms (SNPs) on GBM overall survival outcomes [158, 159]. One study found that GBM patients who carry both TERT mutations and homozygous C-allele mutation for SNP rs2853669 have shorter survival time versus patients with wild-type allele [160]. There is accumulating evidence in the literature that GBM patients with IDH1 somatic mutation have significantly higher overall survival time compared to patients who carry a wild-type allele [161-163].

This work hypothesizes that combining tumor sample's SNP, age, and gender data increases the predictive power of GBM survival outcome. The research proposes a multi-path neural network to predict short (< one-year) and long (≥ one-year) survival groups. The predictive ability of combined SNPs, demographic features (age, age groups, and gender) versus each data source alone was assessed, and the proposed method was compared to support vector machine (SVM) with linear kernel, and random forest

classifiers.

This study downloaded The Cancer Genome Atlas Glioblastoma Multiforme (TCGA-GBM) of 272 white individuals demographics (age, gender), survival (days from diagnosis to death), and tumor samples' pre-aligned whole-exome sequencing data from National Cancer Institute's Genomic Data Commons (GDC) portal. To obtain SNP data from sequence alignment files, a variant calling workflow was performed with Genome Analysis Toolkit (GATK, version 3.8) [39, 76] followed by two-layers quality controls: 1) variant quality score recalibration (VQSR), and 2) hard filtering (depth < 5, genotype quality < 20). SNPs that have any missing value were excluded from further analysis.

The pipeline randomly held out 10% of the whole data set, 5% from each class to create a balanced subset, and kept it as a test set. The other 90% of data was used for training and hyperparameters tuning by employing 10-fold cross-validation. The pipeline then fit a model with the 90% of data that is kept for training with best-performing hyperparameters and predict the test data set. The accuracy of SNPs alone, age and gender alone, and combined SNPs, age, and gender were reported. The research then compares the performance of the proposed method to SVM and random forest.

On the test dataset, the best classification performance is reached by feeding SNP and demographic features into the proposed multi-path convolutional neural network. There we achieved an accuracy of 67%, where linear SVM and random forest attained an accuracy of 60% and 46%. When considering demographic features alone, the linear SVM has 60% accuracy, our method has an accuracy of 60%, and random forest reaches 53% prediction accuracy.

7.2 Methods

7.2.1 Patients Cohort

TCGA-GBM data were obtained for all white individuals that have tumor sample's binary alignment map (BAM) files, survival information (days from cancer index to death), and demographic features (age and gender) from NIH's GDC portal. A total of 272 patients met the inclusion criteria. The pipeline converted age, and gender into numerical values, and also created an age group binary feature with 70 years threshold since GBM patients with age ≥70 have significantly lower survival time [164]. Table 7.1 shows GBM patients' characteristics.

Table 7.1 Cohort Characteristics

	n=272
Short-/long-term survival	128/144
Average age	61.14 (±12.83)
Age ≥70	71
Male/female	177/95

7.2.2. SNPs Calling and Quality Control

Variants calling were performed with tumor samples only, and GATK HaplotypeCaller (version 3.8) [39, 76] was used. GATK scans samples' genomes to identify regions with variability that exceed a defined threshold. From these regions, it builds an assembly directed graph with a reference genome as a template. It uses the most likely graph paths, the ones that have higher read data, to list candidate haplotypes. The candidate haplotype sequences are aligned against the reference genome with the Smith-Waterman algorithm to produce a CIGAR string. GATK determines the likelihood of haplotype by aligning

every read against each haplotype with the PairHMM algorithm, which gives a likelihood for each haplotype given read data. From read data likelihoods, the program assigns allele likelihoods (possible genotypes). Finally, GATK uses Bayes' Theorem to assign genotypes for each sample from the list of possible genotypes.

All subjects' samples were pooled together for variant discovery. To speed up the variants calling stage, each chromosome was cut into roughly ten equal chunks and executed at the same time on a cluster in a scatter-gather approach. In the final variant call set, the GATK variant quality score recalibration (VQSR) algorithm was performed, which uses machine learning to filter out low-quality variants. After applying VQSR filtering (soft filtering). The truth sensitivity filter for VQSR was set to a "99.0%" threshold. We used the following annotations with VQSR to build a recalibration model: InbreedingCoeff, QD, MQ, MQRankSum, ReadPosRankSum, FS, SOR. This study also filtered out variants that have a depth (number of supporting reads) ≤ 5 or genotyping quality ≤ 20. Also, non-SNPs variants and sites that have any missing value were removed. The final output contains a matrix of SNPs and samples. Each SNP column is in the form A/B where A and B are the two alleles copies. Table 7.2 shows the number of SNPs after applying each filtering method.

Table 7.2 TCGA-GBM SNPs Count after Applying Three Filtering Methods

Filtering method	Number of SNPs
Soft filtering (VQSR)	304302
Hard filtering	155673
Soft+hard filtering	107777

7.2.3 SNPs Encoding

To encode an SNP into a numerical format to perform machine learning tasks, the formula: $4 \times A + B$ were used, where A and B are the two alleles copies for a given individual sample. This study multiplies A by 4 to consider all permutations in a multiallelic site (the maximum alternate alleles for an SNP is 3). For example, if an individual is homozygous at the third alternate allele for a particular SNP, then this specific SNP encoding is 15. SNPs were sorted in increasing order according to their genomic position.

7.2.4 Training and Test Sets

Separate training and test data sets were created to ensure the validity of the results. In the original TCGA-GBM dataset of 272 samples, the pipeline shuffled the data and randomly selected 5% from each class, to get a balanced subset, and kept this 10% balanced dataset for model testing. The remaining 90% was used for hyperparameters tuning, by employing 10-fold cross-validation, and to fit a model to predict the unseen test dataset with the best performing hyperparameters. Table 7.3 displays patients' characteristics in training and test data sets.

Table 7.3 Training and Test Sets Characteristics

	Training set n=244	**Test set n=28**
Survival < 1 year	114	14
Survival ≥ 1 year	130	14
Average age	61.1 (±12.5)	61.4 (±14.6)
Age ≥70	62	9
Male/female	157/87	20/8

7.2.5 Hyperparameter Selection

Classifiers hyperparameters, such as the SVM C regularization value, need to be set before model training begins, and thus are not optimized during the learning stage. To choose the best learning rate and the number of epochs hyperparameters for the proposed neural network, all possible pairs in the Cartesian product of the two sets were evaluated: learning rate = (0.001, 0.01, 1) and the number of epochs = (1,2,3, ..., 20) using 10-fold cross-validation in the 90% of the original dataset (number of samples= 244) that are kept for training. This research also employed the same method, with the same data in each fold, to select the best regularization C hyperparameter from the set C= (0.01, 0.1, 1) for linear SVM, as well as the number of trees to grow for random forest from the set (10, 100, 1000). The pipeline then fits a model on the whole training dataset with the best performing hyperparameters and uses the model built to predict the unseen 10% of the original data that was reserved as a test dataset.

7.2.6 Classifiers

Convolutional Neural Network (CNN): CNNs typically are stacked layers of convolution operations with pooling (downsampling of original data for training efficiency) and batch normalization layers in between convolutional layers. The convolution runs on sliding windows of a specified size and fixed step size, to control the moving dot product over training data. A non-linear and differentiable activation function, such as a rectified linear unit (Relu), is then applied to the flattened output.

Multi-path Model: The research proposes a new neural network system, where it feeds the network two inputs: 1) SNPs data, 2) demographic data (age, age groups, and gender). Since SNP data were sorted in an increasing order based on its genomic position, the

SNPs were passed through a series of 1D convolutions, with different kernel sizes and a step size of one, Relu activation function, 1D average pooling, batch normalization layers. Simultaneously, the pipeline fed the three demographic features into two hidden-layers and merge the two paths and train the weights together through three fully connected layers. Then the network passes the weights into a sigmoid function that outputs a value between zero and one. If the output is ≥ 0.5, the network assigns it to class one, and class zero otherwise. The network trained the model with stochastic gradient descent with a momentum that was set to 0.9. The pipeline used 10-fold cross-validation to select the number of epochs and learning rate (lr) value. The batch size was set to 128. Figure 7.1 shows the multi-path model's architecture, all input and output shapes, and convolutions kernel and average pooling sizes. The network is implemented using Keras library [165].

Figure 7.1 The proposed multi-path model architecture with SNP and demographic features.

Single-path Model: The research compares fitting a combined SNP and demographic features with our multi-path model to fitting a single-path 1D convolutional neural net with SNPs only and with three demographic features alone neural network.

Support Vector Machine (SVM): SVM with a linear kernel was used. Briefly, SVM finds a hyperplane that maximizes the distance between the two classes' data points that are closest to the margin (support vectors). In its soft-margin version, SVM allows misclassification of noisy data points and introduces a trade-off hyperparameter C that needs to be tuned. As C approaches infinity, the classifier gets closer to the hard-margin solution. The pipeline uses 10-fold cross-validation to select the best performing C in the training dataset. The research compared combining SNP and demographic features to fitting an SVM model with each data source individually. For SVM and random forest experiments, the scikit-learn library [116] was used.

Random Forest: Random forest is an ensemble method that constructs many decision trees by choosing random samples with replacement to build each tree and randomly generates a subset of features to select from for each candidate split, usually the one with the highest Gini impurity or entropy, then it takes the majority vote of all trees predictions to output a class prediction. The default parameters for the quality measure of the split were used. The pipeline employed 10-fold cross-validation to select the optimal hyperparameter for the number of trees to construct. The pipeline fits a model with combined SNP and demographic features, SNP alone, and age+age group+gender individually.

7.2.8 Evaluation Metrics

Accuracy, which is the number of correctly classified samples over the number of all predicted samples, was used to measure classifiers' prediction power in the test data set. However, in training and validation data sets, we used the balanced accuracy, which is the average of true positive rate and true negative rate, since it has imbalanced class distribution.

7.3 Results

7.3.1 Cross-Validation

In the training set, the pipeline performed 10-fold cross-validation to select the best number of epochs and learning rates for single- and multi-path neural network system. Figure 7.2 shows the mean balanced accuracy attained with different learning rates and the number of epochs across the ten folds. The best mean balanced accuracy of 63% (±0.08) across ten folds is realized when we fed both SNP and demographic features into our multi-path model with 0.01 as the learning rate. The mean balanced accuracy slightly drops after it reaches its peak at the 13th epoch. With SNP data alone, the best learning rate was 0.001 with nine epochs, where the single-path convolutional neural network attained 54% (±0.12) mean balanced accuracy. With the demographic features alone, the single-path neural network reached its highest mean balanced accuracy of 59% (±0.12) at epoch 14 with a learning rate of 0.1.

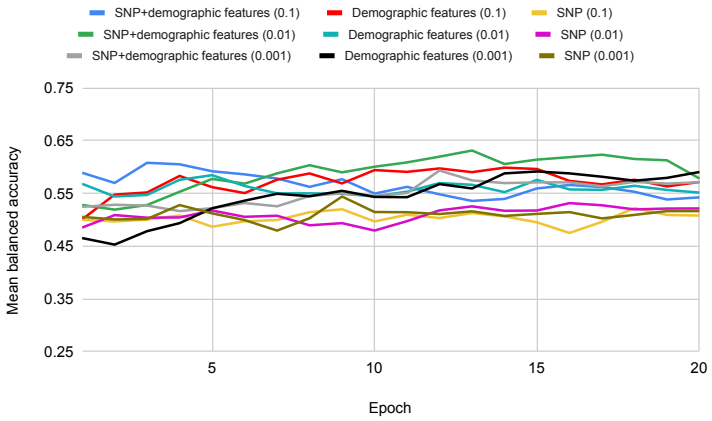


Figure 7.2 Cross-validation average balanced accuracy across 10-folds as a function of the number of epoch and learning rate for multiple data inputs: demographic characteristics (age+age groups+gender) only, SNPs only, or SNPs and demographic characteristics combined. Each line color, which is shown in the series color legends, represents input data (learning rate in parentheses).

SVM C regularization hyperparameter and the number of trees to grow for random forest classifiers were tuned. Figure 7.3 shows that the SVM achieved its best results when C= (1, 0.1), where both values are equally the best in combined SNP and demographic features, SNP alone, and demographic features alone. When learning with demographic features alone, SVM attained 61% (±0.08) mean balanced accuracy. SVM achieved 56% (±0.11) mean balanced accuracy with SNPs data alone, and the mean balanced accuracy drops to 50% (±0.10) when combining SNP and demographic features.

For random forest, setting the number of trees to 10 yielded a better performance for SNPs alone with 50% (±0.12) mean balanced accuracy and demographic features alone 52% (±0.08) mean balanced accuracy. In combined SNPs and demographics, with the optimal number of trees of 100 that was selected with cross-validation in the training set, achieved 49% (±0.11) mean balanced accuracy. Figure 7.3 shows the average 10-fold

cross-validation with different hyperparameters for SVM and random forest.

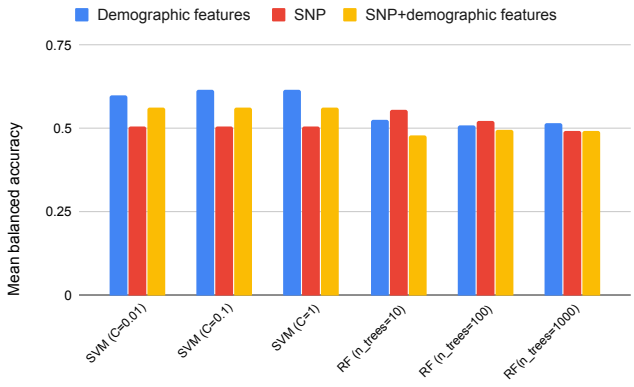


Figure 7.3 Cross-validation mean balanced accuracy across 10-folds with linear SVM (with different C regularization values) and random forest (with different number of trees values) and multiple data inputs: demographic characteristics (age+age groups+gender) only, SNPs only, or SNPs and demographic characteristics combined. Each bar color represents a data source.

7.3.2 Test Set Prediction Performance

After cross-validating, the optimal hyperparameters for each classifier with each data source. The pipeline fits a model on the full training and validation sets and predicts an independent and balanced test set. Table 7.4 shows the accuracies attained by the proposed model, SVM, and random forest accuracies with and without combining SNP and demographic features. The proposed multi-path model, with combined SNP and demographic features (age, age group, and gender), achieved the highest classification accuracy of 67%, when learning with the optimal hyperparameters that were selected with the 10-fold cross-validation: learning rate of 0.01, and 13 epochs.

Table 7.3 Prediction Accuracy on Test Set with the Optimal Hyperparameters

	SNP and demographic	SNP	Demographic
Our method	0.67 lr =0.01 epoch =13	0.60 lr =0.001 epoch =9	0.60 lr =0.1 epoch =14
SVM	0.60 C =1	0.57) C =1	0.60 C =1
Random forest	0.46 # of trees =10	0.50 # of trees =10	0.53 # of trees =100

Combined SNP and Demographic Features: When combining SNP and demographic features, the proposed multi-path model achieved an accuracy of 67%, which outperformed both SVM (60%) and random forest (47%) accuracies. Furthermore, passing SNP, age, age groups, and gender yielded a nicer training curve that is stable across training epochs. Figure 7.4 compares the training balanced accuracy of the combined SNP and demographic features with SNP data alone and demographics individually.

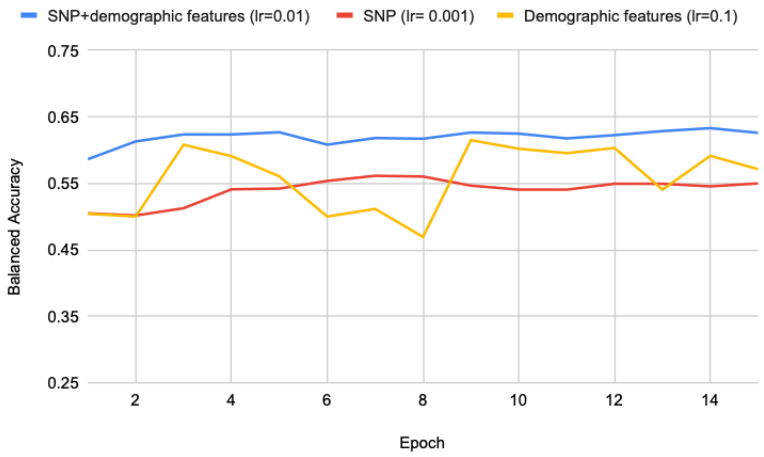


Figure 7.4 Training accuracy on training set (n=244) for combined SNP and demographic features, and each data source individually.

SNP and Demographic Features: In the test set, fitting a model with SNPs individually or age+age groups+gender alone had lower accuracy than combining SNPs and demographic features. With SNP data only, the proposed single-path CNN had an accuracy of 60% with a learning rate of 0.001 and 9 epochs. SVM achieved an accuracy of 57% with C=1, and random forest accuracy is 50% with 100 trees. Figure 7.5 displays the proposed model prediction accuracy with different data sources on the test set. With demographic features alone, SVM and the proposed single-path neural network performed equally with 60% accuracy. Random forest attained 50% accuracy. Table 7.4 compares the accuracy achieved by the proposed CNN, SVM, and random forest with combined SNP and demographic features and with each data source individually.

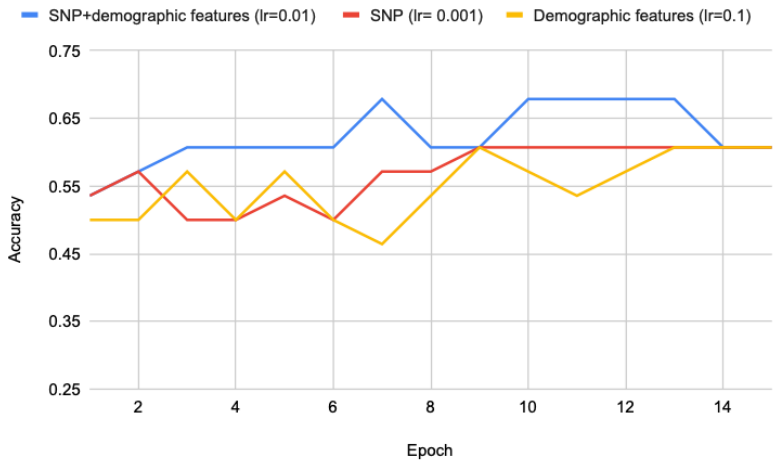


Figure 7.5 Test set prediction accuracy for combined SNP and demographic features, and each data source alone.

7.4 Conclusion

This chapter proposes a new multi-path convolutional neural network for combined SNP and age, age group, and gender that improved upon SVM and random forest in terms of model accuracy in cross-validation and an independent test set. The research shows that

using combined SNP and demographic features in a multi-path network attains a better classification performance than each data source individually and stabilized the learning process.

REFERENCES

[1] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognition Letters,* vol. 107, pp. 3-16, May 2018.

[2] S. O. Olatunji, "Improved email spam detection model based on support vector machines,"*Neural Computing and Applications,* vol. 31, no. 3, pp. 691-699, March 2019.

[3] M. Soltaninejad, L. Zhang, T. Lambrou, G. Yang, N. Allinson, and X. Ye, "MRI Brain Tumor Segmentation and Patient Survival Prediction Using Random Forests and Fully Convolutional Networks," in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Cham, Springer International Publishing, 2018, pp. 204-215.

[4] Y. Suter *et al.*, "Deep Learning Versus Classical Regression for Brain Tumor Patient Survival Prediction," Cham, 2019: Springer International Publishing, in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, pp. 429-440.

[5] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics," *Cancer Journal for Clinicians,* vol. 68, no. 1, pp. 7-30, Jan 2018.

[6] A. Sud, B. Kinnersley, and R. S. Houlston, "Genome-wide association studies of cancer: current insights and future perspectives," *Nature Reviews Cancer,* vol. 17, no. 11, pp. 692-704, Nov 2017.

[7] T. Rice *et al.*, "Understanding inherited genetic risk of adult glioma - a review," *Neuro-Oncology Practice,* vol. 3, no. 1, pp. 10-16, Mar 2016.

[8] S. von Holst *et al.*, "Linkage analysis revealed risk loci on 6p21 and 18p11.2-q11.2 in familial colon and rectal cancer, respectively," *European Journal of Human Genetics,* vol. 27, no. 8, pp. 1286-1295, Aug 2019.

[9] F. J. Couch *et al.*, "Inherited mutations in 17 breast cancer susceptibility genes among a large triple-negative breast cancer cohort unselected for family history of breast cancer," *Journal of Clinical Oncology,* vol. 33, no. 4, pp. 304-11, Feb 2015.

[10] P.-Y. Kao, T. Ngo, A. Zhang, J. W. Chen, and B. S. Manjunath, "Brain Tumor Segmentation and Tractographic Feature Extraction from Structural MR Images for Overall Survival Prediction," in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Cham, Springer International Publishing, 2018, pp 128-141.

[11] B. Gwilliam *et al.*, "Prognosticating in patients with advanced cancer--observational study comparing the accuracy of clinicians' and patients' estimates of survival," *Annals of Oncology,* vol. 24, no. 2, pp. 482-8, Feb 2013.

[12] S. Stiel *et al.*, "Evaluation and comparison of two prognostic scores and the physicians' estimate of survival in terminally ill patients," *Support Care Cancer,* vol. 18, no. 1, pp. 43-9, Jan 2010.

[13] U. Baid *et al.*, "Deep Learning Radiomics Algorithm for Gliomas (DRAG) Model: A Novel Approach Using 3D UNET Based Deep Convolutional Neural Network for Predicting Survival in Gliomas," 2018.

[14] L. Chato and S. Latifi, "Machine Learning and Deep Learning Techniques to Predict Overall Survival of Brain Tumor Patients using MRI Images," in *IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, 23-25 Oct. 2017, pp. 9-14.

[15] W. Zong, J. Lee, C. Liu, J. Snyder, and N. Wen, "Abstract 3351: Overall survival prediction of glioblastoma patients combining clinical factors with texture features extracted from 3-D convolutional neural networks," *Cancer Research,* vol. 79, no. 13 Supplement, p. 3351, 2019, doi: 10.1158/1538-7445.AM2019-3351.

[16] C. The Genomes Project *et al.*, "A global reference for human genetic variation," *Nature,* Article vol. 526, p. 68, 09/30/online 2015, doi: 10.1038/nature15393 https://www.nature.com/articles/nature15393#supplementary-information.

[17] K. Neininger, T. Marschall, and V. Helms, "SNP and indel frequencies at transcription start sites and at canonical and alternative translation initiation sites in the human genome," *PLoS One,* vol. 14, no. 4, p. e0214816, 2019, doi: 10.1371/journal.pone.0214816.

[18] C. Kimchi-Sarfaty *et al.*, "A "silent" polymorphism in the MDR1 gene changes substrate specificity," *Science,* vol. 315, no. 5811, pp. 525-8, Jan 26 2007, doi: 10.1126/science.1135308.

[19] J. M. Mullaney, R. E. Mills, W. S. Pittard, and S. E. Devine, "Small insertions and deletions (INDELs) in human genomes," *Hum Mol Genet,* vol. 19, no. R2, pp. R131-R136, 2010, doi: 10.1093/hmg/ddq400.

[20] F. Dudbridge, "Power and predictive accuracy of polygenic risk scores," *PLoS Genet,* vol. 9, no. 3, p. e1003348, Mar 2013, doi: 10.1371/journal.pgen.1003348.

[21] M. J. Machiela, C. Y. Chen, C. Chen, S. J. Chanock, D. J. Hunter, and P. Kraft, "Evaluation of polygenic risk scores for predicting breast and prostate cancer risk," *Genet Epidemiol,* vol. 35, no. 6, pp. 506-514, Sep 2011, doi: 10.1002/gepi.20600.

[22] M. Hartman, C. Suo, W. Y. Lim, H. Miao, Y. Y. Teo, and K. S. Chia, "Ability to predict breast cancer in Asian women using a polygenic susceptibility model," *Breast Cancer Res Treat,* vol. 127, no. 3, pp. 805-12, Jun 2011, doi: 10.1007/s10549-010-1279-z.

[23] H. M. Wang *et al.*, "A new method for post Genome-Wide Association Study (GWAS) analysis of colorectal cancer in Taiwan," *Gene,* vol. 518, no. 1, pp. 107-13, Apr 10 2013, doi: 10.1016/j.gene.2012.11.067.

[24] E. Capriotti and R. B. Altman, "A new disease-specific machine learning approach for the prediction of cancer-causing missense variants," *Genomics,* vol. 98, no. 4, pp. 310-7, Oct 2011, doi: 10.1016/j.ygeno.2011.06.010.

[25] R. Kamps *et al.*, "Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification," *Int J Mol Sci,* vol. 18, no. 2, Jan 31 2017, doi: 10.3390/ijms18020308.

[26] S. W. Kong *et al.*, "Summarizing polygenic risks for complex diseases in a clinical whole-genome report," *Genet Med,* vol. 17, no. 7, pp. 536-44, Jul 2015, doi: 10.1038/gim.2014.143.

[27] H. Li *et al.*, "Breast cancer risk prediction using a polygenic risk score in the familial setting: a prospective study from the Breast Cancer Family Registry and kConFab," *Genet Med,* vol. 19, no. 1, pp. 30-35, Jan 2017, doi: 10.1038/gim.2016.43.

[28] T. A. Muranen *et al.*, "Polygenic risk score is associated with increased disease risk in 52 Finnish breast cancer families," *Breast Cancer Res Treat,* vol. 158, no. 3, pp. 463-9, Aug 2016, doi: 10.1007/s10549-016-3897-6.

[29] C. M. Vachon *et al.*, "A polygenic risk score for breast cancer in women receiving tamoxifen or raloxifene on NSABP P-1 and P-2," *Breast Cancer Res Treat,* vol. 149, no. 2, pp. 517-23, Jan 2015, doi: 10.1007/s10549-014-3175-4.

[30] T. LaFramboise, "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances," *Nucleic Acids Res,* vol. 37, no. 13, pp. 4181-4193, 2009, doi: 10.1093/nar/gkp552.

[31] K. M. Egan *et al.*, "Cancer susceptibility variants and the risk of adult glioma in a US case-control study," *J Neurooncol,* vol. 104, no. 2, pp. 535-42, Sep 2011, doi: 10.1007/s11060-010-0506-0.

[32] P. Rajaraman *et al.*, "Genome-wide association study of glioma and meta-analysis," *Hum Genet,* vol. 131, no. 12, pp. 1877-88, Dec 2012, doi: 10.1007/s00439-012-1212-0.

[33] Z. Wang *et al.*, "Further Confirmation of Germline Glioma Risk Variant rs78378222 in TP53 and Its Implication in Tumor Tissues via Integrative Analysis of TCGA Data," *Hum Mutat,* vol. 36, no. 7, pp. 684-8, Jul 2015, doi: 10.1002/humu.22799.

[34] Q. Cao *et al.*, "Genetic variants in RKIP are associated with clear cell renal cell carcinoma risk in a Chinese population," *PLoS One,* vol. 9, no. 10, p. e109285, 2014, doi: 10.1371/journal.pone.0109285.

[35] K. L. Huang *et al.*, "Pathogenic Germline Variants in 10,389 Adult Cancers," *Cell,* vol. 173, no. 2, pp. 355-370.e14, Apr 5 2018, doi: 10.1016/j.cell.2018.03.039.

[36] D. Crowther-Swanepoel and R. S. Houlston, "Genetic variation and risk of chronic lymphocytic leukaemia," *Semin Cancer Biol,* vol. 20, no. 6, pp. 363-9, Dec 2010, doi: 10.1016/j.semcancer.2010.08.006.

[37] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "affy--analysis of Affymetrix GeneChip data at the probe level," *Bioinformatics,* vol. 20, no. 3, pp. 307-15, Feb 12 2004, doi: 10.1093/bioinformatics/btg405.

[38] G. A. Van der Auwera *et al.*, "From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline," *Curr Protoc Bioinformatics,* vol. 43, pp. 11.10.1-33, 2013, doi: 10.1002/0471250953.bi1110s43.

[39] A. McKenna *et al.*, "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res,* vol. 20, no. 9, pp. 1297-1303, 2010, doi: 10.1101/gr.107524.110.

[40] M. de Andrade *et al.*, "Evaluating the influence of quality control decisions and software algorithms on SNP calling for the affymetrix 6.0 SNP array platform," *Hum Hered,* vol. 71, no. 4, pp. 221-33, 2011, doi: 10.1159/000328843.

[41] J. M. Korn *et al.*, "Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs," *Nat Genet,* vol. 40, no. 10, pp. 1253-60, Oct 2008, doi: 10.1038/ng.237.

[42] B. J. Kim and S. H. Kim, "Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method," *Proc Natl Acad Sci U S A,* vol. 115, no. 6, pp. 1322-1327, Feb 6 2018, doi: 10.1073/pnas.1717960115.

[43] J. Novembre *et al.*, "Genes mirror geography within Europe," *Nature,* vol. 456, no. 7218, pp. 98-101, 2008/11/01 2008, doi: 10.1038/nature07331.

[44] P. Kersbergen, K. van Duijn, A. D. Kloosterman, J. T. den Dunnen, M. Kayser, and P. de Knijff, "Developing a set of ancestry-sensitive DNA markers reflecting

continental origins of humans," *BMC Genetics,* vol. 10, no. 1, p. 69, 2009/10/27 2009, doi: 10.1186/1471-2156-10-69.

[45] S. P. Kar *et al.*, "Genome-Wide Meta-Analyses of Breast, Ovarian, and Prostate Cancer Association Studies Identify Multiple New Susceptibility Loci Shared by at Least Two Cancer Types," *Cancer Discov,* vol. 6, no. 9, pp. 1052-67, Sep 2016, doi: 10.1158/2159-8290.Cd-15-1227.

[46] J. N. Sampson *et al.*, "Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types," *J Natl Cancer Inst,* vol. 107, no. 12, p. djv279, Dec 2015, doi: 10.1093/jnci/djv279.

[47] G. Abraham, A. Kowalczyk, J. Zobel, and M. Inouye, "Performance and Robustness of Penalized and Unpenalized Methods for Genetic Prediction of Complex Human Disease," *Genetic Epidemiology,* vol. 37, no. 2, pp. 184-195, 2013/02/01 2013, doi: 10.1002/gepi.21698.

[48] N. Chatterjee, B. Wheeler, J. Sampson, P. Hartge, S. J. Chanock, and J.-H. Park, "Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies," *Nature Genetics,* vol. 45, p. 400, 03/03/online 2013, doi: 10.1038/ng.2579 https://www.nature.com/articles/ng.2579#supplementary-information.

[49] J. Kruppa, A. Ziegler, and I. R. Konig, "Risk estimation and risk prediction using machine-learning methods," *Hum Genet,* vol. 131, no. 10, pp. 1639-54, Oct 2012, doi: 10.1007/s00439-012-1194-y.

[50] U. Roshan, S. Chikkagoudar, Z. Wei, K. Wang, and H. Hakonarson, "Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest," *Nucleic Acids Res,* vol. 39, no. 9, p. e62, May 2011, doi: 10.1093/nar/gkr064.

[51] M. Sandhu, A. Wood, and E. Young, "Genomic risk prediction," *Lancet,* vol. 376, no. 9750, pp. 1366-7, Oct 23 2010, doi: 10.1016/s0140-6736(10)61921-6.

[52] C. Kooperberg, M. LeBlanc, and V. Obenchain, "Risk prediction using genome-wide association studies," *Genet Epidemiol,* vol. 34, no. 7, pp. 643-52, Nov 2010, doi: 10.1002/gepi.20509.

[53] D. M. Evans, P. M. Visscher, and N. R. Wray, "Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk," *Hum Mol Genet,* vol. 18, no. 18, pp. 3525-31, Sep 15 2009, doi: 10.1093/hmg/ddp295.

[54] A. C. Janssens and C. M. van Duijn, "Genome-based prediction of common diseases: advances and prospects," *Hum Mol Genet,* vol. 17, no. R2, pp. R166-73, Oct 15 2008, doi: 10.1093/hmg/ddn250.

[55] N. R. Wray, M. E. Goddard, and P. M. Visscher, "Prediction of individual genetic risk to disease from genome-wide association studies," *Genome Res,* vol. 17, no. 10, pp. 1520-1528, 2007, doi: 10.1101/gr.6665407.

[56] N. R. Wray, M. E. Goddard, and P. M. Visscher, "Prediction of individual genetic risk of complex disease," *Curr Opin Genet Dev,* vol. 18, no. 3, pp. 257-63, Jun 2008, doi: 10.1016/j.gde.2008.07.006.

[57] P. Kraft and D. J. Hunter, "Genetic Risk Prediction — Are We There Yet?," *New England Journal of Medicine,* vol. 360, no. 17, pp. 1701-1703, 2009/04/23 2009, doi: 10.1056/NEJMp0810107.

[58] M. H. Gail, "Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk," *J Natl Cancer Inst,* vol. 100, no. 14, pp. 1037-41, Jul 16 2008, doi: 10.1093/jnci/djn180.

[59] A. C. Morrison *et al.*, "Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study," *Am J Epidemiol,* vol. 166, no. 1, pp. 28-35, Jul 1 2007, doi: 10.1093/aje/kwm060.

[60] S. Kathiresan *et al.*, "Polymorphisms Associated with Cholesterol and Risk of Cardiovascular Events," *New England Journal of Medicine,* vol. 358, no. 12, pp. 1240-1249, 2008/03/20 2008, doi: 10.1056/NEJMoa0706728.

[61] N. P. Paynter, D. I. Chasman, J. E. Buring, D. Shiffman, N. R. Cook, and P. M. Ridker, "Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3," *Ann Intern Med,* vol. 150, no. 2, pp. 65-72, Jan 20 2009, doi: 10.7326/0003-4819-150-2-200901200-00003.

[62] C. B. Do, D. A. Hinds, U. Francke, and N. Eriksson, "Comparison of Family History and SNPs for Predicting Risk of Complex Disease," *PLOS Genetics,* vol. 8, no. 10, p. e1002973, 2012, doi: 10.1371/journal.pgen.1002973.

[63] D. Shigemizu *et al.*, "The Construction of Risk Prediction Models Using GWAS Data and Its Application to a Type 2 Diabetes Prospective Cohort," *PLOS ONE,* vol. 9, no. 3, p. e92549, 2014, doi: 10.1371/journal.pone.0092549.

[64] Z. Wei *et al.*, "Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease," *Am J Hum Genet,* vol. 92, no. 6, pp. 1008-12, Jun 6 2013, doi: 10.1016/j.ajhg.2013.05.002.

[65] S. Okser, T. Pahikkala, and T. Aittokallio, "Genetic variants and their interactions in disease risk prediction - machine learning and network perspectives," *BioData Min,* vol. 6, no. 1, p. 5, Mar 1 2013, doi: 10.1186/1756-0381-6-5.

[66] H. Eleftherohorinou *et al.*, "Pathway Analysis of GWAS Provides New Insights into Genetic Susceptibility to 3 Inflammatory Diseases," *PLOS ONE,* vol. 4, no. 11, p. e8068, 2009, doi: 10.1371/journal.pone.0008068.

[67] C. Bernau *et al.*, "Cross-study validation for the assessment of prediction algorithms," *Bioinformatics (Oxford, England),* vol. 30, no. 12, pp. i105-i112, 2014, doi: 10.1093/bioinformatics/btu279.

[68] S. J. Schrodi *et al.*, "Genetic-based prediction of disease traits: prediction is very difficult, especially about the future," *Front Genet,* vol. 5, p. 162, 2014, doi: 10.3389/fgene.2014.00162.

[69] T. A. Manolio, "Bringing genome-wide association findings into clinical use," *Nat Rev Genet,* vol. 14, no. 8, pp. 549-58, Aug 2013, doi: 10.1038/nrg3523.

[70] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, "Five years of GWAS discovery," *Am J Hum Genet,* vol. 90, no. 1, pp. 7-24, Jan 13 2012, doi: 10.1016/j.ajhg.2011.11.029.

[71] L. Wang *et al.*, "SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia," *New England Journal of Medicine,* vol. 365, no. 26, pp. 2497-2506, 2011/12/29 2011, doi: 10.1056/NEJMoa1109016.

[72] D. A. Landau *et al.*, "Evolution and impact of subclonal mutations in chronic lymphocytic leukemia," *Cell,* vol. 152, no. 4, pp. 714-26, Feb 14 2013, doi: 10.1016/j.cell.2013.01.019.

[73] M. D. Mailman *et al.*, "The NCBI dbGaP database of genotypes and phenotypes," *Nature genetics,* vol. 39, no. 10, pp. 1181-1186, 2007, doi: 10.1038/ng1007-1181.

[74] M. Shanshal and R. Y. Haddad, "Chronic lymphocytic leukemia," *Dis Mon,* vol. 58, no. 4, pp. 153-67, Apr 2012, doi: 10.1016/j.disamonth.2012.01.009.

[75] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics,* vol. 25, no. 14, pp. 1754-60, Jul 15 2009, doi: 10.1093/bioinformatics/btp324.

[76] M. A. DePristo *et al.*, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nature Genetics,* vol. 43, no. 5, pp. 491-498, 2011/05/01 2011, doi: 10.1038/ng.806.

[77] I. Guyon, Andr, #233, and Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.,* vol. 3, pp. 1157-1182, 2003.

[78] N. A. Fonseca, J. Rung, A. Brazma, and J. C. Marioni, "Tools for mapping high-throughput sequencing data," *Bioinformatics,* vol. 28, no. 24, pp. 3169-77, Dec 15 2012, doi: 10.1093/bioinformatics/bts605.

[79] A. Hatem, D. Bozdag, A. E. Toland, and U. V. Catalyurek, "Benchmarking short sequence mapping tools," *BMC Bioinformatics,* vol. 14, p. 184, Jun 7 2013, doi: 10.1186/1471-2105-14-184.

[80] E. Alpaydin, *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. Cambridge, MA: The MIT Press, 2004.

[81] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995/09/01 1995, doi: 10.1023/A:1022627411411.

[82] T. Joachims, "Making large-scale svm learning practical," in *Advances in Kernel Methods - Support Vector Learning*, C. J. C. B. Bernhard Sch□ölkopf, Alexander J. Smola Ed., Cambridge, MA: MIT Press, 1999.

[83] I. Guyon, S. Gunn, A. B. Hur, and G. Dror, "Result analysis of the NIPS 2003 feature selection challenge," presented at the Proceedings of the 17th International Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, 2004.

[84] P. Smialowski, D. Frishman, and S. Kramer, "Pitfalls of supervised feature selection," *Bioinformatics (Oxford, England),* vol. 26, no. 3, pp. 440-443, 2010, doi: 10.1093/bioinformatics/btp621.

[85] L. Pasqualucci *et al.*, "Genetics of follicular lymphoma transformation," *Cell Rep,* vol. 6, no. 1, pp. 130-40, Jan 16 2014, doi: 10.1016/j.celrep.2013.12.027.

[86] N. Stransky *et al.*, "The mutational landscape of head and neck squamous cell carcinoma," *Science,* vol. 333, no. 6046, pp. 1157-60, Aug 26 2011, doi: 10.1126/science.1208130.

[87] S. Banerji *et al.*, "Sequence analysis of mutations and translocations across breast cancer subtypes," *Nature,* vol. 486, no. 7403, pp. 405-9, Jun 20 2012, doi: 10.1038/nature11154.

[88] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Res,* vol. 38, no. 16, p. e164, Sep 2010, doi: 10.1093/nar/gkq603.

[89] M. C. Di Bernardo *et al.*, "A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia," *Nat Genet,* vol. 40, no. 10, pp. 1204-10, Oct 2008, doi: 10.1038/ng.219.

[90] S. I. Berndt *et al.*, "Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia," *Nat Genet,* vol. 45, no. 8, pp. 868-76, Aug 2013, doi: 10.1038/ng.2652.

[91] H. E. Speedy *et al.*, "A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia," *Nat Genet,* vol. 46, no. 1, pp. 56-60, Jan 2014, doi: 10.1038/ng.2843.

[92] S. L. Slager *et al.*, "Genome-wide association study identifies a novel susceptibility locus at 6p21.3 among familial CLL," *Blood,* vol. 117, no. 6, pp. 1911-6, Feb 10 2011, doi: 10.1182/blood-2010-09-308205.

[93] S. G. Park, P. Schimmel, and S. Kim, "Aminoacyl tRNA synthetases and their connections to disease," *Proc Natl Acad Sci U S A,* vol. 105, no. 32, pp. 11043-9, Aug 12 2008, doi: 10.1073/pnas.0802862105.

[94] D. Kim, N. H. Kwon, and S. Kim, "Association of aminoacyl-tRNA synthetases with cancer," *Top Curr Chem,* vol. 344, pp. 207-45, 2014, doi: 10.1007/128_2013_455.

[95] K. De Keersmaecker *et al.*, "Fusion of EML1 to ABL1 in T-cell acute lymphoblastic leukemia with cryptic t(9;14)(q34;q32)," *Blood,* vol. 105, no. 12, pp. 4849-52, Jun 15 2005, doi: 10.1182/blood-2004-12-4897.

[96] N. Cheung *et al.*, "Targeting Aberrant Epigenetic Networks Mediated by PRMT1 and KDM4C in Acute Myeloid Leukemia," *Cancer Cell,* vol. 29, no. 1, pp. 32-48, Jan 11 2016, doi: 10.1016/j.ccell.2015.12.007.

[97] M. Emerenciano *et al.*, "Functional analysis of the two reciprocal fusion genes MLL-NEBL and NEBL-MLL reveal their oncogenic potential," *Cancer Lett,* vol. 332, no. 1, pp. 30-4, May 10 2013, doi: 10.1016/j.canlet.2012.12.023.

[98] Y. Wu, X. Zhang, Y. Liu, F. Lu, and X. Chen, "Decreased Expression of BNC1 and BNC2 Is Associated with Genetic or Epigenetic Regulation in Hepatocellular Carcinoma," *Int J Mol Sci,* vol. 17, no. 2, Jan 25 2016, doi: 10.3390/ijms17020153.

[99] L. Lou and B. Xu, "Induction of apoptosis of human leukemia cells by α-anordrin," *Chinese Journal of Cancer Research,* vol. 9, no. 1, pp. 1-5, 1997/03/01 1997, doi: 10.1007/BF02974711.

[100] C. S. Carlson *et al.*, "Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study," *PLoS Biol,* vol. 11, no. 9, p. e1001661, Sep 2013, doi: 10.1371/journal.pbio.1001661.

[101] M. Pino-Yanes *et al.*, "Genetic ancestry influences asthma susceptibility and lung function among Latinos," *J Allergy Clin Immunol,* vol. 135, no. 1, pp. 228-35, Jan 2015, doi: 10.1016/j.jaci.2014.07.053.

[102] B. I. Freedman, J. Divers, and N. D. Palmer, "Population ancestry and genetic risk for diabetes and kidney, cardiovascular, and bone disease: modifiable

environmental factors may produce the cures," *Am J Kidney Dis,* vol. 62, no. 6, pp. 1165-75, Dec 2013, doi: 10.1053/j.ajkd.2013.05.024.

[103] G. Abraham and M. Inouye, "Genomic risk prediction of complex human disease and its clinical application," *Curr Opin Genet Dev,* vol. 33, pp. 10-6, Aug 2015, doi: 10.1016/j.gde.2015.06.005.

[104] J. Usher-Smith, J. Emery, W. Hamilton, S. J. Griffin, and F. M. Walter, "Risk prediction tools for cancer in primary care," *Br J Cancer,* vol. 113, no. 12, pp. 1645-50, Dec 22 2015, doi: 10.1038/bjc.2015.409.

[105] C. Tomasetti, L. Li, and B. Vogelstein, "Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention," *Science,* vol. 355, no. 6331, p. 1330, 2017, doi: 10.1126/science.aaf9011.

[106] I. Martincorena and P. J. Campbell, "Somatic mutation in cancer and normal cells," *Science,* vol. 349, no. 6255, pp. 1483-9, Sep 25 2015, doi: 10.1126/science.aab4082.

[107] I. R. Watson, K. Takahashi, P. A. Futreal, and L. Chin, "Emerging patterns of somatic mutations in cancer," *Nat Rev Genet,* vol. 14, no. 10, pp. 703-18, Oct 2013, doi: 10.1038/nrg3539.

[108] R. P. Young *et al.*, "SNP-Based Risk Score Out Performs a Clinical Model for Dying of Lung Cancer in the NLST-ACRIN Sub-Study (N=10,054)," in *C30. LUNG CANCER SCREENING: WHO, WHY, WHERE, AND HOW MUCH*, (American Thoracic Society International Conference Abstracts: American Thoracic Society, 2017, pp. A5172-A5172.

[109] A. Aljouie, N. Patel, B. Jadhav, and U. Roshan, "Cross-validation and cross-study validation of chronic lymphocytic leukaemia with exome sequences and machine learning," *Int. J. Data Min. Bioinformatics,* vol. 16, no. 1, pp. 47-63, 2016, doi: 10.1504/ijdmb.2016.079801.

[110] N. Patel, B. Jhadav, A. Aljouie, and U. Roshan, "Cross-validation and cross-study validation of chronic lymphocytic leukemia with exome sequences and machine learning," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 9-12 Nov. 2015 2015, pp. 1367-1374, doi: 10.1109/BIBM.2015.7359878.

[111] D. Speed and D. J. Balding, "MultiBLUP: improved SNP-based prediction for complex traits," *Genome Res,* vol. 24, no. 9, pp. 1550-7, Sep 2014, doi: 10.1101/gr.169375.113.

[112] H. C. Erichsen and S. J. Chanock, "SNPs in cancer research and treatment," *Br J Cancer,* vol. 90, no. 4, pp. 747-51, Feb 23 2004, doi: 10.1038/sj.bjc.6601574.

[113] N. R. Wray, J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard, and P. M. Visscher, "Pitfalls of predicting complex traits from SNPs," *Nat Rev Genet,* vol. 14, no. 7, pp. 507-15, Jul 2013, doi: 10.1038/nrg3457.

[114] R. L. Grossman *et al.*, "Toward a Shared Vision for Cancer Genomic Data," *New England Journal of Medicine,* vol. 375, no. 12, pp. 1109-1112, 2016/09/22 2016, doi: 10.1056/NEJMp1607591.

[115] P. Danecek *et al.*, "The variant call format and VCFtools," *Bioinformatics (Oxford, England),* vol. 27, no. 15, pp. 2156-2158, 2011, doi: 10.1093/bioinformatics/btr330.

[116] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.,* vol. 12, pp. 2825-2830, 2011.

[117] P. Wanitchakool *et al.*, "Role of anoctamins in cancer and apoptosis," *Philos Trans R Soc Lond B Biol Sci,* vol. 369, no. 1638, p. 20130096, Mar 19 2014, doi: 10.1098/rstb.2013.0096.

[118] C. Duran and H. C. Hartzell, "Physiological roles and diseases of Tmem16/Anoctamin proteins: are they all chloride channels?," *Acta Pharmacol Sin,* vol. 32, no. 6, pp. 685-92, Jun 2011, doi: 10.1038/aps.2011.48.

[119] Y. Seo *et al.*, "Inhibition of ANO1 by luteolin and its cytotoxicity in human prostate cancer PC-3 cells," *PLOS ONE,* vol. 12, no. 3, p. e0174935, 2017, doi: 10.1371/journal.pone.0174935.

[120] W. M. Linehan, R. Srinivasan, and L. S. Schmidt, "The genetic basis of kidney cancer: a metabolic disease," *Nat Rev Urol,* vol. 7, no. 5, pp. 277-85, May 2010, doi: 10.1038/nrurol.2010.47.

[121] A. P. Fay, S. Signoretti, and T. K. Choueiri, "MET as a target in papillary renal cell carcinoma," *Clin Cancer Res,* vol. 20, no. 13, pp. 3361-3, Jul 1 2014, doi: 10.1158/1078-0432.Ccr-14-0690.

[122] S. Sadetzki *et al.*, "Description of selected characteristics of familial glioma patients - results from the Gliogene Consortium," *Eur J Cancer,* vol. 49, no. 6, pp. 1335-45, Apr 2013, doi: 10.1016/j.ejca.2012.11.009.

[123] M. L. Goodenberger and R. B. Jenkins, "Genetics of adult glioma," *Cancer Genet,* vol. 205, no. 12, pp. 613-21, Dec 2012, doi: 10.1016/j.cancergen.2012.10.009.

[124] B. Malmer *et al.*, "GLIOGENE an International Consortium to Understand Familial Glioma," *Cancer Epidemiol Biomarkers Prev,* vol. 16, no. 9, pp. 1730-4, Sep 2007, doi: 10.1158/1055-9965.Epi-07-0081.

[125] N. Paunu *et al.*, "A novel low-penetrance locus for familial glioma at 15q23-q26.3," *Cancer Res,* vol. 62, no. 13, pp. 3798-802, Jul 1 2002.

[126] K. Labreche *et al.*, "Diffuse gliomas classified by 1p/19q co-deletion, TERT promoter and IDH mutation status are associated with specific genetic risk loci," *Acta Neuropathol,* vol. 135, no. 5, pp. 743-755, 2018, doi: 10.1007/s00401-018-1825-z.

[127] Y. Liu, S. Shete, F. J. Hosking, L. B. Robertson, M. L. Bondy, and R. S. Houlston, "New insights into susceptibility to glioma," *Arch Neurol,* vol. 67, no. 3, pp. 275-8, Mar 2010, doi: 10.1001/archneurol.2010.4.

[128] S. Purcell *et al.*, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *American journal of human genetics,* vol. 81, no. 3, pp. 559-575, 2007, doi: 10.1086/519795.

[129] J. Marchini and B. Howie, "Genotype imputation for genome-wide association studies," *Nat Rev Genet,* vol. 11, no. 7, pp. 499-511, Jul 2010, doi: 10.1038/nrg2796.

[130] C. C. Spencer, Z. Su, P. Donnelly, and J. Marchini, "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip," *PLoS Genet,* vol. 5, no. 5, p. e1000477, May 2009, doi: 10.1371/journal.pgen.1000477.

[131] B. N. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS Genet,* vol. 5, no. 6, p. e1000529, Jun 2009, doi: 10.1371/journal.pgen.1000529.

[132] H. Li, "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data," *Bioinformatics,* vol. 27, no. 21, pp. 2987-93, Nov 1 2011, doi: 10.1093/bioinformatics/btr509.

[133] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001/10/01 2001, doi: 10.1023/A:1010933404324.

[134] V. Gotea, J. J. Gartner, N. Qutob, L. Elnitski, and Y. Samuels, "The functional relevance of somatic synonymous mutations in melanoma and other cancers," *Pigment Cell Melanoma Res,* vol. 28, no. 6, pp. 673-84, Nov 2015, doi: 10.1111/pcmr.12413.

[135] N. Deng, H. Zhou, H. Fan, and Y. Yuan, "Single nucleotide polymorphisms and cancer susceptibility," *Oncotarget,* vol. 8, no. 66, pp. 110635-110649, 2017, doi: 10.18632/oncotarget.22372.

[136] F. Supek, B. Minana, J. Valcarcel, T. Gabaldon, and B. Lehner, "Synonymous mutations frequently act as driver mutations in human cancers," *Cell,* vol. 156, no. 6, pp. 1324-1335, Mar 13 2014, doi: 10.1016/j.cell.2014.01.051.

[137] J. P. Thakkar *et al.*, "Epidemiologic and molecular prognostic review of glioblastoma," *Cancer Epidemiol Biomarkers Prev,* vol. 23, no. 10, pp. 1985-96, Oct 2014, doi: 10.1158/1055-9965.Epi-14-0275.

[138] A. F. Tamimi and M. Juweid, "Epidemiology and Outcome of Glioblastoma," in *Glioblastoma*, D. V. S Ed. Internet: Codon, 2017.

[139] K. Clark *et al.*, "The cancer imaging archive (TCIA)," *Maintaining and operating a public information repository,* Article vol. 26, no. 6, pp. 1045-1057, 12/1 2013, doi: 10.1007/s10278-013-9622-7.

[140] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Cham, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., 2015// 2015: Springer International Publishing, pp. 234-241.

[141] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[142] W. G. Cochran, "The X2 test of goodness of fit," *Annals of Mathematical Statistics,* vol. 23, pp. 315-345, 1952, doi: 10.1214/aoms/1177729380.

[143] "Broad Institute TCGA Genome Data Analysis Center Firehose std-data 2016 01 28 run." Broad Institute of MIT and Harvard. Cambridge, MA.

[144] M. W. Woolrich *et al.*, "Bayesian analysis of neuroimaging data in FSL," *Neuroimage,* vol. 45, no. 1 Suppl, pp. S173-86, Mar 2009, doi: 10.1016/j.neuroimage.2008.10.055.

[145] Y. Wang *et al.*, "Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates," *PLoS One,* vol. 9, no. 1, p. e77810, 2014, doi: 10.1371/journal.pone.0077810.

[146] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *NIPS 2017 Autodiff Workshop*, Long Beach, California, 2017.

[147] L. Bottou, "Large-Scale Machine Learning with Stochastic Gradient Descent," in *Proceedings of COMPSTAT'2010*, Heidelberg, Y. Lechevallier and G. Saporta, Eds., 2010// 2010: Physica-Verlag HD, pp. 177-186.

[148] I. J. Goodfellow *et al.*, "Generative Adversarial Nets," in *NIPS*, 2014.

[149] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.,* vol. 15, no. 1, pp. 1929-1958, 2014.

[150] W.-Z. Gao, L.-M. Guo, T.-Q. Xu, Y.-H. Yin, and F. Jia, "Identification of a multidimensional transcriptome signature for survival prediction of postoperative glioblastoma multiforme patients," *Journal of Translational Medicine,* vol. 16, no. 1, p. 368, 2018/12/20 2018, doi: 10.1186/s12967-018-1744-8.

[151] B. Liu *et al.*, "A prognostic signature of five pseudogenes for predicting lower-grade gliomas," *Biomed Pharmacother,* vol. 117, p. 109116, Sep 2019, doi: 10.1016/j.biopha.2019.109116.

[152] L. Macyszyn *et al.*, "Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques," *Neuro Oncol,* vol. 18, no. 3, pp. 417-25, Mar 2016, doi: 10.1093/neuonc/nov127.

[153] Y. Tan, W. Mu, X.-c. Wang, G.-q. Yang, R. J. Gillies, and H. Zhang, "Improving survival prediction of high-grade glioma via machine learning techniques based on MRI radiomic, genetic and clinical risk factors," *European Journal of Radiology,* 2019/07/13/ 2019, doi: https://doi.org/10.1016/j.ejrad.2019.07.010.

[154] A. Toft *et al.*, "Abstract B3: Prognostic and predictive biomarkers in recurrent WHO grade 3 malignant glioma patients treated with bevacizumab and irinotecan," *Molecular Cancer Therapeutics,* vol. 14, no. 12 Supplement 2, p. B3, 2015, doi: 10.1158/1535-7163.TARG-15-B3.

[155] G. Steponaitis *et al.*, "High CHI3L1 expression is associated with glioma patient survival," *Diagn Pathol,* vol. 11, p. 42, Apr 27 2016, doi: 10.1186/s13000-016-0492-4.

[156] R.-C. Chai *et al.*, "Systematically profiling the expression of eIF3 subunits in glioma reveals the expression of eIF3i has prognostic value in IDH-mutant lower grade glioma," *Cancer Cell International,* vol. 19, no. 1, p. 155, 2019/06/04 2019, doi: 10.1186/s12935-019-0867-1.

[157] M. Tian *et al.*, "Impact of gender on the survival of patients with glioblastoma," *Biosci Rep,* vol. 38, no. 6, p. BSR20180752, 2018, doi: 10.1042/BSR20180752.

[158] A. Fogli *et al.*, "The tumoral A genotype of the MGMT rs34180180 single-nucleotide polymorphism in aggressive gliomas is associated with shorter patients' survival," *Carcinogenesis,* vol. 37, no. 2, pp. 169-176, Feb 2016, doi: 10.1093/carcin/bgv251.

[159] A. Bunevicius *et al.*, "Common genetic variations of deiodinase genes and prognosis of brain tumor patients," *Endocrine,* Aug 26 2019, doi: 10.1007/s12020-019-02016-6.

[160] D. Cui *et al.*, "R132H mutation in IDH1 gene reduces proliferation, cell survival and invasion of human glioma by downregulating Wnt/beta-catenin signaling," *Int J Biochem Cell Biol,* vol. 73, pp. 72-81, Apr 2016, doi: 10.1016/j.biocel.2016.02.007.

[161] K. Wang *et al.*, "Radiological features combined with IDH1 status for predicting the survival outcome of glioblastoma patients," *Neuro-oncology,* vol. 18, no. 4, pp. 589-597, 2016, doi: 10.1093/neuonc/nov239.

[162] H. Yan *et al.*, "IDH1 and IDH2 Mutations in Gliomas," *New England Journal of Medicine,* vol. 360, no. 8, pp. 765-773, 2009/02/19 2009, doi: 10.1056/NEJMoa0808710.

[163] S. E. Combs *et al.*, "Prognostic significance of IDH-1 and MGMT in patients with glioblastoma: one step forward, and one step back?," *Radiat Oncol,* vol. 6, p. 115, Sep 13 2011, doi: 10.1186/1748-717x-6-115.

[164] U. Smrdel, M. S. Vidmar, and A. Smrdel, "Glioblastoma in Patients over 70 Years of Age," *Radiol Oncol,* vol. 52, no. 2, pp. 167-172, 2018, doi: 10.2478/raon-2018-0010.

[165] F. Chollet *et al.*, "Keras," https://keras.io, 2015. (accessed: Oct/25/2019)