Spring 5-2020

Integrated Machine Learning and Bioinformatics Approaches for Prediction of Cancer-Driving Gene Mutations

Oluyemi Odeyemi
Chapman University

Integrated Machine Learning and Bioinformatics Approaches for Prediction of Cancer-Driving Gene Mutations

A Dissertation by

Oluyemi Odeyemi

Chapman University

Orange, California

Schmid College of Science and Technology

Submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Computational and Data Sciences

May 2020

Committee in charge:

Gennady M. Verkhivker, Ph.D., Committee Chair

Cyril Rakovski, Ph.D.

Divya Sain, Ph.D.

Moom Roosan, Ph.D

The dissertation of Oluyemi Odeyemi is approved.

Gennady Verkhivker, Ph.D., Committee Chair

Cyril Rakovski, Ph.D., Committee Member

Divya Sain, Ph.D., Committee Member

Moom Roosan, Ph.D., Committee Member

May 2020

Integrated Machine Learning and Bioinformatics Approaches for Prediction of Cancer-Driving Gene Mutations

by Oluyemi Odeyemi

ACKNOWLEDGEMENT

I would like to thank Dr. Gennady Verkhivker, Dr. Cyril Rakovski, Dr. Hesham El-Askary, Dr. Divya Sain, Dr. John Peach, Ryan Peeler, Sidy Danioko, Steve Agajanian, Kristalee Lio, Ayo Bello and everyone who offered thoughts and advice that were useful in completing this research.

DEDICATION

To my family, friends and loved ones.

ABSTRACT

Integrated Machine Learning and Bioinformatics Approaches for Prediction of Cancer-Driving Gene Mutations

by Oluyemi Odeyemi

Cancer arises from the accumulation of somatic mutations and genetic alterations in cell division checkpoints and apoptosis, this often leads to abnormal tumor proliferation. Proper classification of cancer-linked driver mutations will considerably help our understanding of the molecular dynamics of cancer. In this study, we compared several cancer-specific predictive models for prediction of driver mutations in cancer-linked genes that were validated on canonical data sets of functionally validated mutations and applied to a raw cancer genomics data.

By analyzing pathogenicity prediction and conservation scores, we have shown that evolutionary conservation scores play a pivotal role in the classification of cancer drivers and were the most informative features in the driver mutation classification. Through extensive comparative analysis with structure-functional experiments and multicenter mutational calling data from Pan-Cancer Atlas studies, we have demonstrated the robustness of our models and addressed the validity of computational predictions. We evaluated the performance of our models using the standard diagnostic metrics such as sensitivity, specificity, area under the curve and F-measure. To address the interpretability of cancer-specific classification models and obtain novel insights about molecular signatures of driver mutations, we have complemented machine learning predictions with structure-functional analysis of cancer driver mutations in several key tumor suppressor genes and oncogenes. Through the experiments carried out in this study, we found that evolutionary-based features have the strongest signal in the machine learning classification

of driver mutations and provide orthogonal information to the ensembled-based scores that are prominent in the ranking of feature importance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 General Introduction

1.1 Genetic Variation

Cancer is driven by changes at the cellular and molecular levels. Cancer development and proliferation are associated with the accumulation of mutations. Notably, quite a few of identified mutations are responsible for cellular variations leading to cancer. Most variations are neutral and benign (passenger) in nature while a small fraction of the mutations drive the cancer development process. Genetic variation describes the mutation in the genome's DNA sequence. Genetic variation is responsible for the distinct traits' humans exhibit. It is the result of subtle differences in the DNA. Variation occurs in germ and somatic cells. The only variation that arises in germ cells can be inherited from one individual to another and so affect the dynamics of the population, consequently leading to evolutionary changes. Mutation is the primary source of genetic variation, however sexual reproduction and recombination contribute significantly to genetic variation. A mutation is an alteration in the genome sequence. New mutations occur when there are errors during DNA replication that are not corrected by DNA repair enzymes. Mutation can be neutral, beneficial or damaging to an organism. Most somatic mutations are salient but can occasionally interfere with major cellular functions. Accumulation of genetic alteration can result in tumor development in oncogenes, tumor-suppressor genes and stability genes (Vogelstein and Kinzler 2004; Vogelstein et al., 2013; Feinberg et al. 2006). Early somatic mutation can lead to developmental disorders whereas incessant accumulation of mutation can cause cancer. Human cells innately have several safety protocols to protect themselves against the lethal effects of mutation inducing cancers. Therefore, it is the defective genes that result in cancer proliferation (Yeang et al., 2008; Li et al., 2016).

Genetic variations are generally divided into three main classes namely, single base-pair substitution, insertion or deletion(indel) and structural variation. Single base-pair substitution is a type of mutation in which a single nucleotide base is deleted or inserted from a DNA sequence such as in transition (interchange of purine or pyrimidine nucleic acids) and transversion (interchange of a purine and pyrimidine nucleic acids). Single nucleotide polymorphisms (SNPs) result from the substitution of a single base-pair. (SNPs) are the most occurring type of genetic variation in people. A SNP represents a difference in a single DNA base, and they are found approximately in 1 out of every 300 bases. An example of SNPs would be a substitution of a thymine nucleic acid with guanine.

Indel variation refers to the insertion and/or deletion of nucleotides in genomic DNA. Indels play a significant role in the identification and detection of human diseases such as cancer. Indels are a common kinase activation system in cancer (Sehn, 2015; Porter et al., 2015; Sanders and Mason 2016). Indels are among the most common types of structural variants. Cumulatively, there are between 1.5 and 2.7 million indel polymorphisms in the human population, with ~ 0.4 million short indels in each individual (McMahon et al., 2017). When an indel occurs within the coding region of the DNA, it is referred to as 'in-frame" and if the number of DNA lost or gained is divisible by 3, it is referred to as frameshift because the triplet reading code is altered for all subsequent nucleotides. Frameshift indels often lead to premature stop codons and most times have more functional impact than in-frame indels (Sanders and Mason, 2016; Copley, 2010). Yang et al., 2010 found a strong correlation between indels and base substitutions in cancer-related genes and showed that they tend to concentrate at the same locus in the coding sequences with the same samples. Also, they observed that a high proportion of indels are found in somatic variation when compared with meiotic mutations. Yang et al. 2010 also concluded that indels can

often be the driver-mutation in cancer proliferation that they attributed to the major influence of indels on gene function.

Structural variation is the large-scale differences in the genomic DNA. They are the region of DNA with base size larger than 1 kb. This genetic variation includes copy number variants (CNVs) and chromosomal rearrangement events such as duplications, inversions, translocation, deletions, and insertions. Identification of structural variation is crucial to genome interpretation but has been historically challenging due to limitations inherent to available genome technologies. Structural variations are mainly responsible for the evolutionary diversity of human genomes at individual and population levels (Dennis and Eichler, 2016; Kosugi et al., 2019). Pang et al., 2010 and Alkan et al., 2011 observed that the genomic difference between individuals caused by structural variations has been estimated to be more than 3 times higher than those by SNVs. Therefore, structural variations may have higher impacts on gene functions than SNVs and short indels. Consequently, structural variations are affiliated

with human diseases such as cancer (Stankiewicz and Lupski, 2010).

1.2 Cancer-Linked Driver Mutation

Genomic instability is central to cancer development (McFarland et al., 2017; Burrell et al., 2013). Genomic instability is responsible for driver and passenger mutations. Driver mutations are responsible for driving carcinogenic processes. Whereas, passenger mutations have no proliferative impact on cellular systems. Studies have shown that from sequenced cancers passenger mutation accounts for over 90% of all genomic variation. The role of passenger mutations is not clearly understood, with some studies arguing that passengers are misclassified ‘mini-drivers’ or effectively neutral and potentially harmful to cancer (Castro-Giner et al.,2015

and McFarland et al., 2013). McFarland et al., 2017 observed that accumulated passenger mutations can be moderately harmful to cancer cells. They further observed that although passenger mutations exhibit individually weak effects on cancer progression, their collective impact is in line with a skewed high number of driver mutations, leading to strife between passengers and drivers. Consequently, passenger's damaging effects are most evident in higher mutation rates (McFarland et al., 2017).

The identification of driver mutations in cancer remains a challenging task. There are several computational strategies aimed at detecting driver genes and ranking mutations for their carcinogenicity prospect (Torkamani et al., 2009). Consequently, several driver genes are not tagged as disease-related (Brown et al., 2019).

Presently, the reoccurrence of a mutation in patients remains one of the top reliable markers of mutation driver status. Nonetheless, some mutations are more likely to occur than others due to differences in background mutations rates arising from different types of DNA replication and repair systems (Brown et al., 2019). From the study of Brown et al., 2019, they showed that mutations not yet observed in a tumor had relatively low mutability, thus indicating that background mutability might limit the occurrence of mutation. Also, they concluded that mutability of driver mutations is often lower than that of passenger mutation and consequently adjusting mutation recurrence frequency by mutability significantly improved prediction of driver mutation.

1.3 Functional Prediction and Conservation Scores

Many algorithms have been designed to predict the molecular impact of amino acid substitutions on protein function and measure the conservation of nucleotide positions. These

algorithms map functional predictions and annotations for human splice site variants and non-synonymous single-nucleotide variants thereby providing key supporting evidence to clinicians when interpreting variants per the American College Medical Genetics (ACMG) guidelines. The most common functional prediction algorithms are MetaSVM, MetaLR, CADD, PolyPhen2, VEST3, PROVEAN, MutationTaster, MutationAssessor, FATHMM, and SIFT while the most popular conservation score metrics are PhyloP, LRT and GERP2.

1.4 Machine Learning

Machine learning (ML) is a branch of artificial intelligence based on the concept that systems can learn from data through pattern recognition. In recent years, machine learning has been used to develop predictive models for a more robust understanding of biological systems. A typical machine learning task consists of the following: 1) Exploratory data analysis, 2) Data preprocessing, 3) Model training, 4) Feature optimization and model parameterization and 5) Final predictions (Kandoi, 2019). Exploratory data analysis (EDA) and data preprocessing are pivotal to any ML task. These steps involve data cleaning steps- handling missing values, removal of redundant features, feature extraction, identification of predictors and target variables. The accuracy of a machine learning model is dependent on its predictors' capability of distinguishing one class from another.

ML is broadly classified into supervised (classification and regression) and unsupervised learning. Classification tasks have a categorical target variable, for example, gram-positive vs gram-negative microorganisms while regression has continuous target variables such as a colony-forming unit of *Staphylococcus aureus*. Unsupervised learning involves tasks without pre-defined target variables such as dimension reduction and clustering.

Some popular algorithms used for classification, regression, and unsupervised learning tasks include ensemble tree-based algorithms, support vector machines, generalized linear regression, general linear regression, k-means and self-organizing maps (Kandoi, 2019). Machine learning plays a cardinal role in computational biology and bioinformatics. Drug target identification, protein structure prediction, microbial morphology, gene function predictions are some of the common applications of ML (Pierre and Soren, 2003; Yang 2010; Mitra 2019).

In this study we used machine learning models to: (i) identify and characterize cancer driving mutations (SNVs) using functional, evolutionary-conservation and ensemble-score predictors, (ii) reclassify genomic variants of unknown significance associated with cancer development and (iii) classify survival status of unresectable cancers based on clinical and genomic features

1.5 Dissertation Organization

This dissertation is divided into six chapters. A summary of each chapter is given below.

Chapter 1 provides a background introduction to this dissertation.

Chapter 2 includes a published article motivating the need to study structure-functional analysis of cancer-driving mutation in oncogenes and tumor suppressor genes. By examining sequence, structure and ensembled-based features we were able to show that evolutionary conservation scores play a critical role in the classification of cancer drivers thereby providing the strongest signal in machine learning prediction. The article has been published under the title, "*Machine Learning Classification and Structure-Functional Analysis of Cancer Mutations Reveal Unique Dynamic and Network Signatures of Driver Sites in Oncogenes and Tumor Suppressor Genes* " by Agajanian S, Odeyemi O, Bischoff N, Ratra S, Verkhivker GM in the Journal of Chemical

Information and modeling 2018 Oct 22;58(10):2131-2150. doi: 10.1021/acs.jcim.8b00414. Epub 2018 Oct 3.

Chapter 3 includes parts of an article "*Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for Classification and Biomolecular Modeling of Cancer Driver Mutations"* by Steve Agajanian, Odeyemi Oluyemi, and Gennady Verkhivker published in Frontiers in Molecular Biosciences 2019; 6: 44. By analyzing raw nucleotide sequences for cancer driver mutation classification, we integrated various ensemble-based approaches with deep convolutional neural networks to predict cancer driver mutation in genomic datasets.

Chapter 4 includes a manuscript currently under preparation or submission to a peer-reviewed paper. In the paper titled "*Machine Learning-Based Overall Survival Status Classification and Cox PH Model Analysis of Selected Unresectable Cancers*", by examining genomic features such as iAMP21, TCF-HLF, ETV6-RUNX1, MLL Rearranged molecular subtypes, gene expression type, fraction of genome altered and clinical features such as age at diagnosis, cell of tumor origin (T-cell & B-cell), we build an ensemble-based predictive model to classify overall survival status of unresectable cancers namely acute lymphoid leukemia, colorectal adenocarcinoma, glioblastoma, and pediatric neuroblastoma. Also, we used the same features for the Cox PH survival analysis model.

Chapter 5 includes a manuscript currently under preparation for submission to a peer-reviewed journal. We built a machine learning model to reclassify variants of unknown significance (VUS) based on established clinical values. The manuscript is titled, "*Machine Learning Reclassification of Variants of Uncertain Significance*".

Chapter 6 includes the general conclusions of this dissertation and the contribution of this study to cancer therapy research

1.6 Reference

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, *12*(5), 363–376. doi: 10.1038/nrg2958

Brown, A. L., Li, M., Goncearenco, A., & Panchenko , A. R. (n.d.). Finding driver mutations in cancer: Elucidating the role of background mutational processes. *PLOS Computational Biology*, *15*(4). doi: https://doi.org/10.1371/journal.pcbi.1006981

Burrell, R. A., Mcgranahan, N., Bartek, J., & Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, *501*(7467), 338–345. doi: 10.1038/nature12625

Castro-Giner, F., Ratcliffe, P., & Tomlinson, I. (2015). The mini-driver model of polygenic cancer evolution. *Nature Reviews Cancer*, *15*(11), 680–685. doi: 10.1038/nrc3999

Copley, S. D. (2010). Evolution and the Enzyme. *Comprehensive Natural Products II Chemistry and Biology*, *8*, 9–46. doi: https://doi.org/10.1016/B978-008045382-8.00670-5

Dennis, M. Y., & Eichler, E. E. (2016). Human adaptation and evolution by segmental duplication. *Current Opinion in Genetics & Development*, *41*, 44–52. doi: 10.1016/j.gde.2016.08.001

Kandoi, G., Machine learning tools for mRNA isoform function prediction (2019). Graduate Theses and Dissertations. 17479. https://lib.dr.iastate.edu/etd/17479

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, *20*(1). doi: 10.1186/s13059-019-1720-5

Li, H. T., Zhang, J., Xia, J., & Zheng, C. H. (2016). Identification of driver pathways in cancer based on combinatorial patterns of somatic gene mutations. *Neoplasma*, *63*(01), 57–63. doi: 10.4149/neo_2016_007

Mcfarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R., & Mirny, L. A. (2013). Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences*, *110*(8), 2910–2915. doi: 10.1073/pnas.1213968110

Mcfarland, C. D., Yaglom, J. A., Wojtkowiak, J. W., Scott, J. G., Morse, D. L., Sherman, M. Y., & Mirny, L. A. (2017). The Damaging Effect of Passenger Mutations on Cancer Progression. *Cancer Research*, *77*(18), 4763–4772. doi: 10.1158/0008-5472.can-15-3283-t

Mcmahon, K., Paciorkowski, A. R., Walters-Sen, L. C., Milunsky, J. M., Bassuk, A., Darbro, B., … Gropman, A. (2017). Neurogenetics in the Genome Era. *Swaimans Pediatric Neurology*, 257–267. doi: 10.1016/b978-0-323-37101-8.00034-5

Mitra, S. (2019). *Introduction to machine learning and bioinformatics*. Boca Raton, FL: CRC Press/Taylor & Francis Group.

Pang, A. W., Macdonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., … Scherer, S. W. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biology*, *11*(5). doi: 10.1186/gb-2010-11-5-r52

Pierre, B., & Soren, B. (2003). *Bioinformatics The Machine Learning Approach*. New Delhi: Affiliated East-West Press P. Ltd.

Porter, J., Berkhahn, J., & Zhang, L. (2015). A Comparative Analysis of Read Mapping and Indel Calling Pipelines for Next-Generation Sequencing Data. *Emerging Trends in*

Computational Biology, Bioinformatics, and Systems Biology, 521–535. doi: 10.1016/b978-0-12-802508-6.00029-6

Sanders, S. J., & Mason, C. E. (2016). The Newly Emerging View of the Genome. *Genomics, Circuits, and Pathways in Clinical Neuropsychiatry*, 3–26. doi: 10.1016/b978-0-12-800105-9.00001-9

Sanders, S. J., & Mason, C. E. (2016). The Newly Emerging View of the Genome. *Genomics, Circuits, and Pathways in Clinical Neuropsychiatry*, 3–26. doi: 10.1016/b978-0-12-800105-9.00001-9

Sehn, J. K. (2015). Insertions and Deletions (Indels). *Clinical Genomics*, 129–150. doi: 10.1016/b978-0-12-404748-8.00009-5

Stankiewicz, P., & Lupski, J. R. (2010). Structural Variation in the Human Genome and its Role in Disease. *Annual Review of Medicine*, *61*(1), 437–455. doi: 10.1146/annurev-med-100708-204735

Torkamani, A., Verkhivker, G., & Schork, N. J. (2009). Cancer driver mutations in protein kinase genes. *Cancer Letters*, *281*(2), 117–127. doi: 10.1016/j.canlet.2008.11.008

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science*, *339*(6127), 1546–1558. doi: 10.1126/science.1235122

Vogelstein, B., & Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature Medicine*, *10*(8), 789–799. doi: 10.1038/nm1087

Yang, H., Zhong, Y., Peng, C., Chen, J.-Q., & Tian, D. (2010). Important role of indels in somatic mutations of human cancer genes. *BMC Medical Genetics*, *11*(1). doi: 10.1186/1471-2350-11-128

Yang, Z. R. (2010). *Machine learning approaches to bioinformatics*. Singapore: World Scientific.

Yeang, C.-H., Mccormick, F., & Levine, A. (2008). Combinatorial patterns of somatic gene mutations in cancer. *The FASEB Journal*, *22*(8), 2605–2622. doi: 10.1096/fj.08-10898

2 Machine Learning Classification and Structure-Functional Analysis of Cancer Mutations Reveal Unique Dynamic and Network Signatures of Driver Sites in Oncogenes and Tumor Suppressor Genes

Steve Agajanian, Yemi Odeyemi, Nathaniel Bischoff, Simrath Ratra and Gennady Verkhivker

Author's Contribution

YO contributed to the design of the study and interpretation of the results. YO wrote part of codes for the data preprocessing and data analysis. All the authors read and approved the final manuscript**.**

2.1 Abstract

Background: Cancer arises from the accumulation of somatic mutations and genetic alterations in cell division checkpoints and apoptosis, this often leads to abnormal tumor proliferation. Proper classification of cancer-linked driver mutations will considerably help our understanding of the molecular dynamics of cancer.

Methods: In this study, we compared two cancer-specific predictive models- logistic regression and random forest for prediction of driver mutations in cancer-linked genes that were validated on canonical data sets of functionally validated mutations and applied to a raw cancer genomics data.

Results: By analyzing pathogenicity prediction and conservation scores, we have shown that evolutionary conservation scores play a pivotal role in the classification of cancer drivers and were the most informative features in the driver mutation classification. Through extensive

comparative analysis with structure-functional experiments and multicenter mutational calling data from Pan-Cancer Atlas studies, we have demonstrated the robustness of our models and addressed the validity of computational predictions. We evaluated the performance of our models using the standard diagnostic metrics such as sensitivity, specificity, area under the curve and F-measure. To address the interpretability of cancer-specific classification models and obtain novel insights about molecular signatures of driver mutations, we have complemented machine learning predictions with structure-functional analysis of cancer driver mutations in several key tumor suppressor genes and oncogenes.

Conclusion: Through the experiments carried out in this study, we found that evolutionary-based features have the strongest signal in the machine learning classification of driver mutations and provide orthogonal information to the ensembled-based scores that are prominent in the ranking of feature importance.

2.2 Introduction

Cancer arises from a gradual buildup of somatic mutations and genetic alterations that undermine cell division checkpoints and apoptosis, this often leads to abnormal tumor proliferation (Iranzo et al.,2018; Reddy et al., 2017; Li and Thirumala, 2016). The primary objective of cancer research is the discovery and characterization of functional effects of variants that contribute to tumor development. Several DNA sequencing programs have prompted interest in cancer studies, consequently yielding vital information needed for understanding genomic basis for tumor development. Somatic variations have been extensively characterized through deep-sequencing analyses of genome-wide association studies (GWAS) and the coding exomes of various cancer types, showing that there are ~140 genes whose intragenic mutations contribute

to cancer, with a relatively small fraction of recurrent somatic variants providing a proliferative edge to cancer cells and often being detected based on mutational frequency in omics studies.

A large percentage of the somatic mutations are passengers that occur randomly as a result of mutagenesis, without a known functional impact and biological effect. By analyzing the variations driving cancer development in more than 7000 tumors, molecular evolution-informed studies have shown that an insignificant fraction of mutated genes is required to transform a single normal cell into a cancer cell across diverse cancer types. Novel methods have also focused on the discovery of presumed drivers in the genome's non-coding regions. Albeit major driver mutations can substantially promote tumor proliferation, some passengers can be singularly weak and yet concertedly lethal, indicating that disease progression is often difficult to justify based on a classical binary passenger-driver model. Cancer-linked variations that are inconsequential to tumor proliferation and are present at low frequency in cancer cohorts can form a set of "mini-drivers", showing that mutational motifs in cancer genomes are highly heterogeneous and can span a continuum of phenotypic impacts. Recent NGS breakthroughs have led to the inauguration of interdisciplinary cancer genomic projects and key data portals, such as The Catalogue of Somatic Mutations in Cancer (COSMIC) database, The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) and the International Cancer Genome Consortium (ICGC) cancer genome projects(Hudson et al., 2010; Zhang et al., 2011). The COSMIC database includes data from full exome sequencing of 1020 cancer cell lines. TCGA data include information about 40 cancer projects from >20 000 genes, ~3.1 million mutations (Jensen et al., 2017) and 33 cancer types. The ICGC data portal includes 86 cancer projects of 22 cancer primary sites with 81 782 588 annotated simple somatic mutations. (Klonowska et al., 2016; Hinkson et al.,2017). The cBio Cancer Genomics Portal provides access to cancer

genomics data sets from >5000 tumor samples, 215 cancer studies, and 981 genes (Cerami et al., 2012; Gao et al., 2013). Oncogenomics advancement has provided large complex data needed for data-centric analyses of genetic variations in various cancer types (Nakagawa and Fujita 2018).

In recent times, many bioinformatics tools that compute the structural impact of a given SNV have been developed for predictions of presumed cancer-linked drivers and functional annotation of somatic mutations (Cheng et al., 2016; Ding et al., 2014; Raphael et al., 2014). Computational features for analyzing mutations in the protein-coding regions(Sim et al.,2012; Adzhubei et al., 2010; Chun et al., 2009; Reva et al., 2011; Schwarz et al., 2010; Gonzalez-Perez and Lopez-Bigas, 2011; Choi et al., 2012; Shihab et al., 2013) and several sequence-based scores(Davydov et al., 2010; Garber et al., 2009)were efficiently used for prediction of cancer driver mutations in the noncoding regions. Combined annotation-dependent depletion (CADD) and genome-wide annotation of variants (GWAVA) (Kircher et al., 2014; Ritchie et al., 2014) methods were developed for characterization and classification of the noncoding variants by combining various genomic annotations into integrated score measures with the aid of support vector machine classifiers. Cancer driver annotation (CanDrA) (Mao et al.,2013) and cancer-specific high-throughput annotation of somatic mutations (CHASM) (Carter et al.,2009) are cancer-specific ML approaches that utilized evolutionary, functional and genetic features of somatic mutations computed by various prediction algorithms. Cancer-related analysis of variants toolkit (CRAVAT), which is a web-based CHASM application, is now often used for prioritization of genes and variants important for specific cancer tissue types (Douville et al., 2013).

In the effort to consolidate a comprehensive functional annotation for SNVs discovered in exome sequencing studies, a database of human nonsynonymous SNVs (dbNSFP) was

developed as a one-stop resource for analysis of disease-causing mutations (Liu et al., 2011; 2013; 2016). Based on detailed comparisons and machine learning-based integration of 18 scoring methods for prediction of nonsynonymous SNVs, the two ensemble scores RadialSVM and LR were designed, showing superior performance over their 10 component scores (Dong et al., 2015). The latest database dbWGFP of functional predictions for SNVs collected approximately 8.6 billion possible human whole-genome SNVs, with capabilities to impute a total of 48 functional prediction scores for each SNV, including 32 functional prediction scores by 13 approaches, 15 conservation features from 4 different tools including ensemble-based predictors MSRV, RadialSVM, and LR scores(Wu et al., 2016).

ML studies indicated that the incorporation of both sequence and structure-based scores can provide important information for classification of cancer genes and driver mutations, but the underlying molecular reasons for a weak consensus between functional features remains poorly understood and hidden in the feature selection process. Modern deep learning methods can leverage large data sets for finding hidden patterns and making robust predictions in cancer genomics, and drug discovery applications (Angermueller et al., 2016; Zhang et al., 2017; Min et al., 2016; Jing et al., 2018; Zhou and Troyanskava,2015; Yuan et al., 2016). However, it is often overlooked that the performance and interpretability of machine learning models are equally important for predictions and understanding of cancer mutation signatures.

In this study, we compared two ML models, logistic regression (logit) and random forest (RF) classifiers by considering various cancer-specific golden sets of functionally known cancer mutations(Carter et al., 2009; Martelotto et al., 2014) and using a set of diverse functional prediction and conservation scores that included computations of 48 functional scores using dbWGFP server (Wu et al., 2016). By evaluating a diverse spectrum of functional features, we

show that conservation-derived and ensemble-based scores were the most informative predictors in the ML classification of cancer-linked driver mutation. We used logit and RF models to predict and analyze driver mutations in Cbioportal cancer genomics data set (Cerami et al., 2012; Gao et al., 2013) by considering 145 601 mutations covering various cancer subtypes and over 300 genes.

2.3 Methods

2.3.1 Data Source and Dataset

We primarily sourced our datasets from Cbioportal, TCGA, and CanDra (from the COSMIC database). We used various well studied benchmarking data sets of functionally validated and manually curated variations for our training and validation sets. The initial set included a total of 3591 SNVs from oncogenes such as HRAS, BRAF, KIT, PIK3CA, KRAS, EGFR, ERRB2, DICER1, ESR1, IDH1, IDH2, MYOD, SF3B1, and tumor suppressor genes such as RUNX1, BRCA1, BRCA2 and TP53 (Martelotto et al., 2014). Due to our goal of classifying mutation as either driver or passenger, we classified neutral and uncertain SNVs as passenger whereas the driver mutations were left intact. The ML model was trained and validated only on missense mutations. The first benchmarking had 3706 mutations, with only 3591 SNVs that included 140 neutral, 849 non-neutral, and 2602 of uncertain function which was assigned as passengers. The second employed data set was taken from the original CanDrA study (Mao et al.,2013) with 1550 SNVs. The CanDrA data set was initially derived by combining ovarian carcinoma (OVC) and glioblastoma multiforme (GBM) mutational data extracted from TCGA (Weinstein et al., 2013) and COSMIC databases (Forbes et al.,2015). In the GBM sets, 134 SNVs were drivers and 585 SNVs were passenger mutations, while in the OVC datasets, 122 SNVs were driver mutations and 709 were passengers (Figure 2-2). The

CanDra data set (Mao et al., 2013) was integrated with the benchmarking data set (Martelotto et al., 2014) to generate one master training set used to build the ML models. The tree-based random forest and logistic regression classification models were used to predict cancer-linked driver variants from the cBioPortal cancer genomics data set (Cerami et al., 2012; Gao et al., 2013) by considering data from over 200 cancer studies covering 20 primary cancer sites (ovary, thyroid, CNS/brain, prostate, kidney, bowel, lymphoid, stomach, bladder, head/neck, breast, myeloid, uterus, lung, PNS, skin, liver, pancreas, soft tissue, cervix).

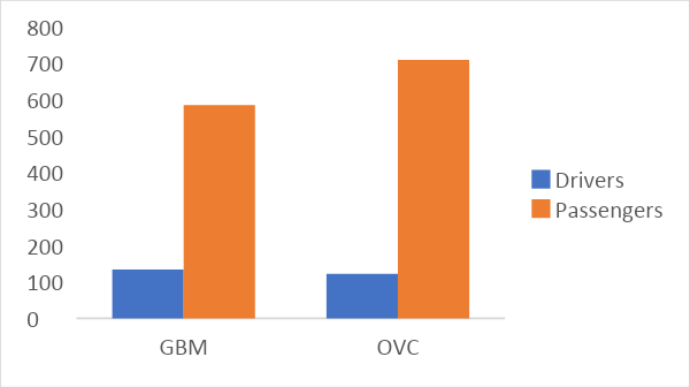


Figure 2-1 Bar plot showing the distribution of cancer-linked driver mutation in glioblastoma multiforme and ovarian carcinoma

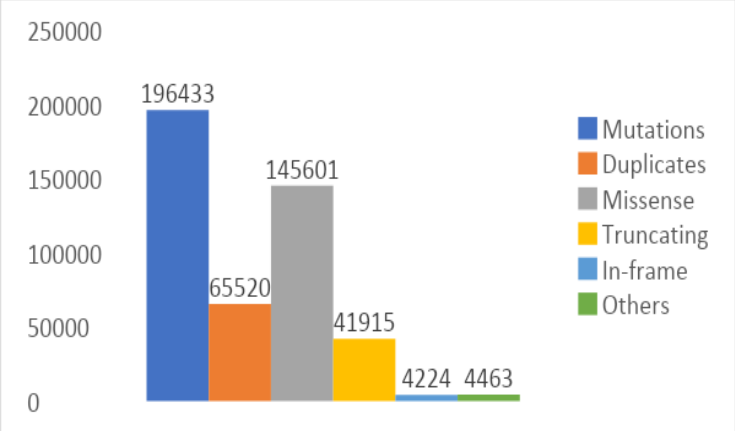


Figure 2-2. Bar plot showing the distribution of studied cancer genes and mutations from Cbioportal database

Table 2-1 Some of the studied genes from Cbioportal database categorized by cancer subtypes

Protein function	**Genes**
Cell death regulation signaling	NFKB1, NFKB2, CHUK, DIRAS3, FAS, HLA-G, BAD, BCL2, BCL2L1, APAF1, CASP9, CASP8, CASP10, CASP3, CASP7, CASP6, GSK3B, ARL11, WWOX
Telomere Maintenance	TERC, TERT
RTK Signaling	EGFR, ERBB2, ERBB3, ERBB4, PDGFA, PDGFRB, KIT, FGF1, IGF1, IGF1R, VEGFA, VEGFB, KDR
P13K-AKT-mTOR signaling	PIK3CA, PIK3R1, PIK3R2, PTEN, PDPK1, AKT1, AKT2, FOXO1, FOXO3, MTOR, RICTOR, TSC1, TSC2, RHEB, AKT1s1, RPTOR
Ras-Raf-MEK-Erk/JNK signaling	KRAS, HRAS, BRAF, RA1, MAP3K1, MAP3K2, MAP3K3, MAP3K4, MAP3K5, MAP2K1, MAP2K2, MAP2K3, MAP2K4, MAPK1, MAPK3, DAB2, RAB25
Regulation of ribosomal protein synthesis	RPS6KA1, RPS6KA2, RPS6KB1, RPS6KB2, EIF5A2, EIF4E, EIF4EBP1, RPS6, HIFIA
Angiogenesis	VEGFA, VEGFB, KDR, CXCL8, CXCR1, CXCR2
Folate transport	SLC19A1, FOLR1, FOLR2, FOLR3, IZUMO1R
Invasion and metastasis genes	MMP1, MMP2, MMP3, MMP7, MMP9, MMP10, MMP11, MMP12, MMP13, MMP14, MMP15, MMP16, MMP19, MMP28, ITGB3, ITGB3, PTK2, CDH1, SPARC, WFDC2

A total of 80,081 unique missense mutations from 378 cancer genes was considered (after excluding 65 520 duplicate mutations in patients with multiple samples from a total of 145 601 mutations Figure 2-2). glioblastoma/TP53 pathway genes; glioblastoma/RTK/Ras/PI3K/AKT signaling genes; glioblastoma/RB pathway genes; Genes implicated in multiple cancer types and associated with the following major functions: cell cycle control genes; DNA damage response genes; growth/proliferation signaling genes; RTK signaling genes; p53 signaling genes; notch signaling genes survival/cell death regulation signaling genes; telomere maintenance genes; Ras-Raf-MEK-Erk/JNK signaling genes; PI3K-AKT-mTOR signaling genes; regulation of ribosomal protein synthesis and cell growth genes; angiogenesis genes; folate transport genes; invasion and metastasis genes. A total of 56,634 unique missense mutations were finally analyzed and classified by our RF and logit models (Table 2-1).

2.3.2 Machine Learning: Random Forest and Logistic Regression Classifiers

Several factors were considered in the choice of our classifiers for our models such as easy interpretation, speed (run time) and statistical importance. Afterward, we used and compared the performance of logistic regression (logit) and random forest (RF). Logistic regression (logit) is an ML classifier that is used to predict the probability of a categorical binary outcome. In logit, the dependent variable is a dichotomous(binary) in nature that contains data coded as 0 or 1. It answer questions such as how does the probability of obesity (yes/ no) change for every additional can of soda drank per day? Logit models the log odds of an event. Logistic regression models are based on specific statistical assumptions such as the target variable takes the form of a binomial distribution, the absence of outliers in the data and no multicollinearity among the predictors.

Mathematically, the logit classifier predicts P(Y=1) as a function of X.

$$Log\left(\frac{Y}{1-Y}\right) = b_{\mathrm{o}} + b_1X_1 + b_2X_2 + \cdots + b_{\mathrm{n}}X_{\mathrm{n}} \qquad (2.1)$$

Where $\left(\frac{Y}{1-Y}\right)$ is the log likelihood of the target variable, b_{o} is the y-intercept, $b_{\mathrm{n}}X_{\mathrm{n}}$ are the slope coefficient and the predictors.

RF is a tree-based ensemble classifier where each tree is exposed to a random subset of features and a random subset of samples drawn with replacement. In this model, the number of trees, the number of features seen by each tree, and the maximum depth of each tree can be tuned to fit the data. A data point to be classified is passed to every tree in the forest, and the predicted class is the winner of a collective vote by all trees. Because of the bootstrapped nature of all the trees in the forest, random forest classifiers are typically robust to noise and can accurately fit nonlinear relationships present in data. In the framework of this approach, we used Bayesian optimization to optimize the hyper-parameters of the model with 100 iterations in our optimization run. By changing values for the hyper-parameters, we monitored how this affects the F-score which is the harmonic mean of the precision and recall, to obtain the final set of hyper-parameters. We provided accuracy to the optimizer as the average of 10-fold cross-validation run across three different random seeds. The final set of hyper-parameters chosen by Bayesian optimization were maximum depth = 200, maximum features = 32, number of estimators = 70. The number of estimators determines the number of trees in the forest, maximum depth determines the maximum depth that each tree can grow to, and maximum features determine the number of features randomly selected at each node when choosing the best tree split. The depth of the random forest model and measurements of information Gini impurity implemented in scikit-learn (Pedregosa et al., 2011) were employed and refined to achieve the best results and avoid overfitting problem (Biau, 2012)

$$Gini\ impurity = \sum_{i=1}^{c} f_i(1 - f_i) \qquad (2.2)$$

Where f_i is the frequency label of i at a node and c is the number of unique labels. The model training and tuning was done using scikit-learn free software machine learning library for the Python programming language (pedregosa et al., 2011). In addition, we also explored logistic regression classification model that allows for fast training and prediction phases of the model without tuning requirement and need for feature rescaling (Palazon-Bru et al., 2017).

For the model diagnostic, accuracy, recall, precision, and F_1 metrics were used to evaluate the performance of each classifier. These parameters are defined as follows:

$$Accuracy = \frac{TP + TN}{all};\ Precision = \frac{TP}{TP + FP} \qquad (2.3)$$

$$Recall = \frac{TP}{TP + FN};\ F1 = 2\,\frac{Precision \times Recall}{Precision + Recall} \qquad (2.4)$$

The true positive (TP) and true negative (TN) are defined as the number of mutations that are classified correctly as driver and passenger mutations, respectively. Likewise, false positive (FP) and false negative (FN) are defined as the number of mutations that are misclassified into the other mutational classes. An *F* score is a measure of precision and recall and is often used in binary classification problems. Precision is defined as the number of positive samples the model predicts correctly (true positives) divided by the true positives plus the false positives. Recall is defined as true positives divided by true positives plus false negatives. The positive predictive value (PPV) and negative predictive value (NPV) were assessed for various components of the model. PPV is defined as the proportion of mutations predicted to be driver mutations that were experimentally validated as drivers:

$$Positive\ Predictive\ Value\ (PPV) = \frac{TP}{TP + FP} \quad (2.5)$$

NPV is the proportion of mutations predicted to be passengers that were experimentally validated as neutral:

$$Negative\ Predictive\ value = \frac{TN}{TN + FN} \quad (2.6)$$

The model performance was evaluated using receiver operating characteristic area under the curve. The receiver operating curve (ROC) is a graph where sensitivity is plotted as a function of 1 – specificity. The sensitivity or true positive rate (TPR) is defined as the percentage of non-neutral mutations that are correctly identified as driver mutations:

$$Sensitivity(TPR) = \frac{TP}{TP + FN} \quad (2.7)$$

The specificity or true negative rate (TNR) is defined as the percentage of mutations that are correctly identified as passengers:

$$Specificity(TNR) = \frac{TN}{TN + FP} \quad (2.8)$$

The area under the ROC is denoted AUC. A reliable and valid AUC estimate can be interpreted as the probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example. The AUC effectively measures discrimination of true positive rate versus false positive rate, providing means to evaluate the ability of the model to correctly classify drivers and passenger mutations.

2.3.3 Mutational Predictor Scores: Feature Selection and Feature Importance Analysis

The first set of selected features were acquired from dbWGFP web server of functional predictions for human whole-genome single nucleotide variants that provided 32 functional prediction scores and 15 evolutionary features (Wu et al., 2016).

Function prediction scores are scores that predict the probability of a SNV to cause a damaging change in protein function whereas evolutionary scores are scores providing different conservation metrics of a given nucleotide site across multiple species. Some of the score features such as RadialSVM, SIFT, PolyPhen, LRT, MSRV, Mutation Assessor , MutationTaster, FATHMM, LR (Dong et al., 2015 ; Sim et al., 2012 ; Adzhubei et al., 2010 ; Chun and Fay, 2009 ;Wu et al., 2016 ; Reva et al., 2011 ; Schwarz et al., 2010 ; Shihab et al., 2013) can be applied only to SNVs in the protein coding regions, whereas other scores such as Gerp++, SiPhy, PhyloP, Grantham, CADD and GWAVA (Davydov et al., 2010 ; Garber et al., 2009 ; Grantham, 1974 ; Kircher et al., 2014 ; Ritchie et al. , 2014) can evaluate SNVs spreading over the whole genome. The ensemble-based scores RadialSVM and LR are integrated features that used support vector machine (SVM) and LR machine learning approaches to combine information from 10 individual component scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, Gerp++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP), as well as the maximum frequency observed in the 1000 genomes populations (Dong et al., 2015).

In the final dataset, we imputed 48 feature scores for each SNV (Wu et al., 2016) and the resulting data set was divided into training and test sets: 80-20% ratio respectively. Our resulting dataset was preprocessed to handle missing values, skewness, distribution disparity etc.

The training data was subjected to stepAIC feature selection, where features are dropped in a step by step manner based on each feature statistical significance hierarchy, and the model trained on the resulting features. After the feature selection resulting in a final data set of 32 features.

2.4 Results and Discussion

2.4.1 ML Classification of Cancer Driver Mutations on Canonical Data Sets: Ensemble-Based and Conservation Features Consistently Outperform Structural Scores

Logit and RF models were trained by examining a combination of two canonical cancer-specific sets of functionally validated mutations (Carter et al., 2009; Martelotto et al., 2014) and using a group of various predictors that included functional scores obtained from dbWGFP server (Table 2-2). In this analysis, we compared the performance of both classifiers and focused on identifying statistical significant predictors that drive cancer predictions in both models (Figure 2-3). In accordance with prior studies, the ensemble-based scores LR and RadialSVM dominated the feature importance distribution in both models, significantly outweighing the importance of other predictors (Figure 2-3A, B). In the RF model, the LR and RadialSVM scores were the top ranked predictors with the information value scores of 0.27 and 0.23 respectively, followed by a group of sequence-based evolutionary conservation predictors (Mutation Assessor, Gerp++, GerpRS, SiPhy, and PhyloP) that also showed significant feature importance (Figure 2-3A)

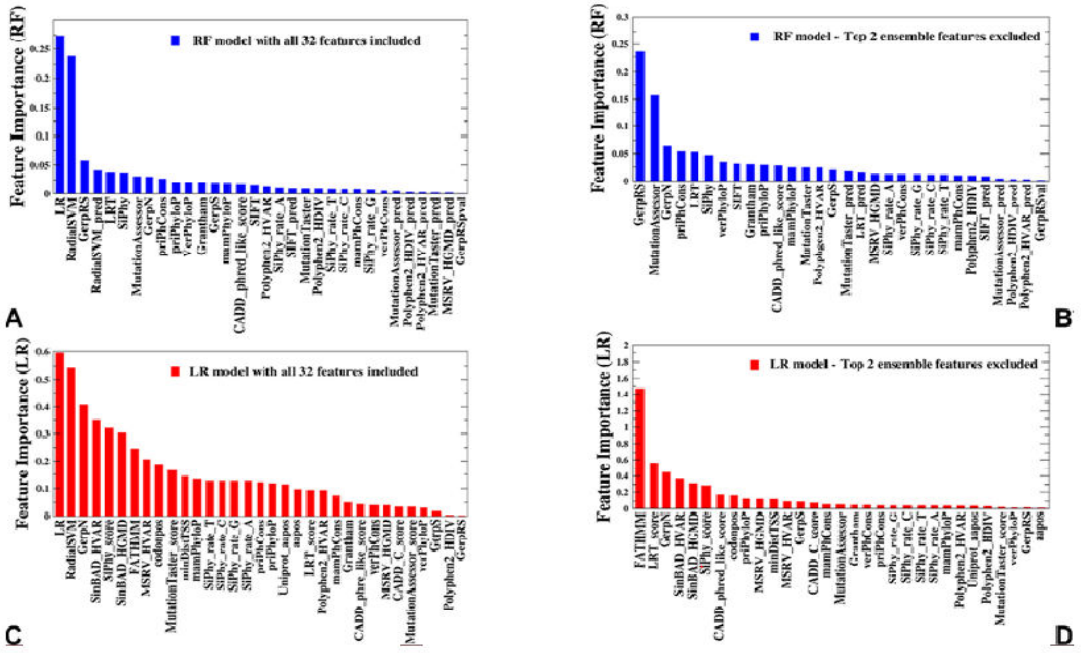


Figure 2-3. Feature importance analysis of the RF and logit machine learning models on the canonical cancer-specific data set of functionally validated mutations. Feature importance of a final data set of 32 features in machine learning predictions using RF model (A, B) and logit model (C,D). Feature importance is measured using the information value and weight of evidence criteria. **(A)** Feature importance analysis of the RF model including all 32 features. **(B)** Feature importance analysis of the RF model after excluding two top ensemble-based features LR and RadialSVM. The feature importance values are shown in blue filled bars and annotated. **(C)** Feature importance analysis of the LR model including all 32 features. **(D)** Feature importance analysis of the LR model after excluding two top ensemble-based features LR and RadialSVM. The feature importance values are shown in red filled bars and annotated. The information value (IV) metric typically indicates that attributes with less than 0.02 score have no predictive power, while those that fall between 0.02 and 0.1 are considered to have weak predictive power, the features with importance values in the range 0.1–0.2 indicate medium predictive power, and the features with importance exceeding 0.2 often indicate strong predictive power.

Interestingly, the ensemble features RadialSVM and LR integrate information from components that individually also indicate a substantial information value (GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, and PhyloP). LR and RadialSVM features were initially developed to differentiate non-pathogenic polymorphisms from pathogenic (mutant) missense variants (Dong et al., 2015). The major rationale behind consideration of ensemble-based scores alongside with individual components they include was therefore to evaluate transferability and performance of various predictors for passenger/driver classification. In this analysis, we compared the two ensemble-based scores, both with each other and with all other features. The variable importance analysis confirmed that the ensemble based LR and RadialSVM dwarfs' other features in both models. Despite these scores reversed rank in the logit model, they remained the top two features with the information scores of 0.332 and 0.441 for LR and RadialSVM scores respectively (Figure 2-3B). Another significant observation of feature importance analysis was that the ensemble predictors not only eclipse their own component scores but also demonstrated greater importance than all the other predictors.

To evaluate the models ability to recap the predictions in the absence of two ensemble-based metrics, we repeated our experiments by excluding the ensemble-based metrics and ranked the importance of the remaining mutational scores based on their statistical significance (Figure 2-3,C,D). The top ranked features were commensurable to functional scores based on evolutionary conservation patterns, such as Mutation Assessor, which was derived to evaluate sequence homology of protein families and subfamilies within and between species using combinatorial entropy formalism (Reva et al., 2011). Other highly ranked individual features included GerpN neutral evolution score and Gerp element scores (Davydov et al., 2010) that

estimate amino acid substitution deficits to identify constrained elements in multiple alignments. Also, a similar feature importance ranking was observed in the RF-based application of a different ensemble learner REVEL (rare exome variant ensemble learner) developed for predicting pathogenicity of missense variants, (Ioannidis et al., 2016) where the ensemble-derived predictors along with Gerp++ and Mutation Assessor individual scores also showed the highest feature importance. Another significant observation of our analysis was that a group of evolutionary-based features (GerpRS, GerpN, Gerp++, Mutation Assessor) can be adequately potent not only for prediction of pathogenic missense SNVs but also in cancer-specific passenger/driver classification of somatic mutations.

In the logit model, MSRV, SinBAD_HVAR, and SinBAD_HGMD scores were the top-ranking scores. MSRV, an integrated feature obtained from a set of 24 physicochemical properties and several conservation scores to compute disease-causing nonsynonymous SNV mutations (Jiang et al., 2007). SinBaD scores is also an integrated feature that were initially developed by a logistic regression model with ninety binary features derived from multiple sequence alignment designed for evaluation of mutational effects in protein coding and promoter regions (Lehmann and Chen et al., 2013). Another highly ranked feature was likelihood ratio test (LRT) score that adopted the log-likelihood ratio of the conserved relative to the neutral model to predict functional significance of mutations (Chun and Fay 2009). Noteworthily, some of the top- ranking features such as SiPhy, PhyloP, and Grantham were initially developed to provide prediction scores for variants spreading over the whole genome. Concurrently, functional scores that estimate the impact of mutational variants on the protein coding regions (MutationTaster, LRT, SIFT, PolyPhen2_HVAR) contributed less significantly to the feature performance (Figure2-3 C,D). Some of the population-based scores that differentiates pathogenic missense

mutants from the typical polymorphisms (SIFT, PolyPhen2) were not as important in our classification models than cancer-specific features, such as FATHMM designed to distinguish somatic drivers (Figure 2-3 C, D).

Conversely, we assessed for multicollinearity among each feature of the integrated scores so as to identify dominant individual components and leverage the complementary of different prediction algorithms for model performance augmentation. The importance measure for an individual feature can reflect correlations with other features as well as its intrinsic predictive ability because importance may be shared among correlated features. Collinearity indicates a linear relationship in a model. When features are highly correlated, they cannot independently predict the value of the response variable. In statistical terms, they reduce the significance of the features. To assess for possible redundancies in our model, we computed and analyzed Spearman's rank pairwise correlations between different prediction scores (Figure 2-4). From our analysis, RadialSVM and LR are highly correlated with one of its original component scores, Mutation Assessor, and moderately correlated with other top performing sequence-based features (PhyloP, PhCons, GerpN, GerpRS, and GerpS). These independent features recorded high correlation with cancer pathogenicity and can provide classification of cancer driver variations.

In our logit model, we observed a significant correlation between RadialSVM and other ensemble-based SinBAD_HVAR,SinBAD_HGMD and MSRV scores (Figure 2-4 B). Synchronously, the ensemble scores were only moderately correlated with other top performing sequence-based features - LRT, PhyloP, PhCons, SiPhy, PhyloP, GerpN, GerpRS, and GerpS. Furthermore, we observed that GerpRS, GerpN, Mutation Assessor and LRT were weakly correlated with each other and have a fairly low correlation with other scores in the RF model Figure 2-4 A). Hence, no pairwise high collinearity was observed in the RF model. Our results

are consistent with the evaluation of functional and conservation scores in prediction of disease-causing SNVs and application of radial SVM machine learning models to predict patterns of cancer driver mutations (Dong et al., 2016)

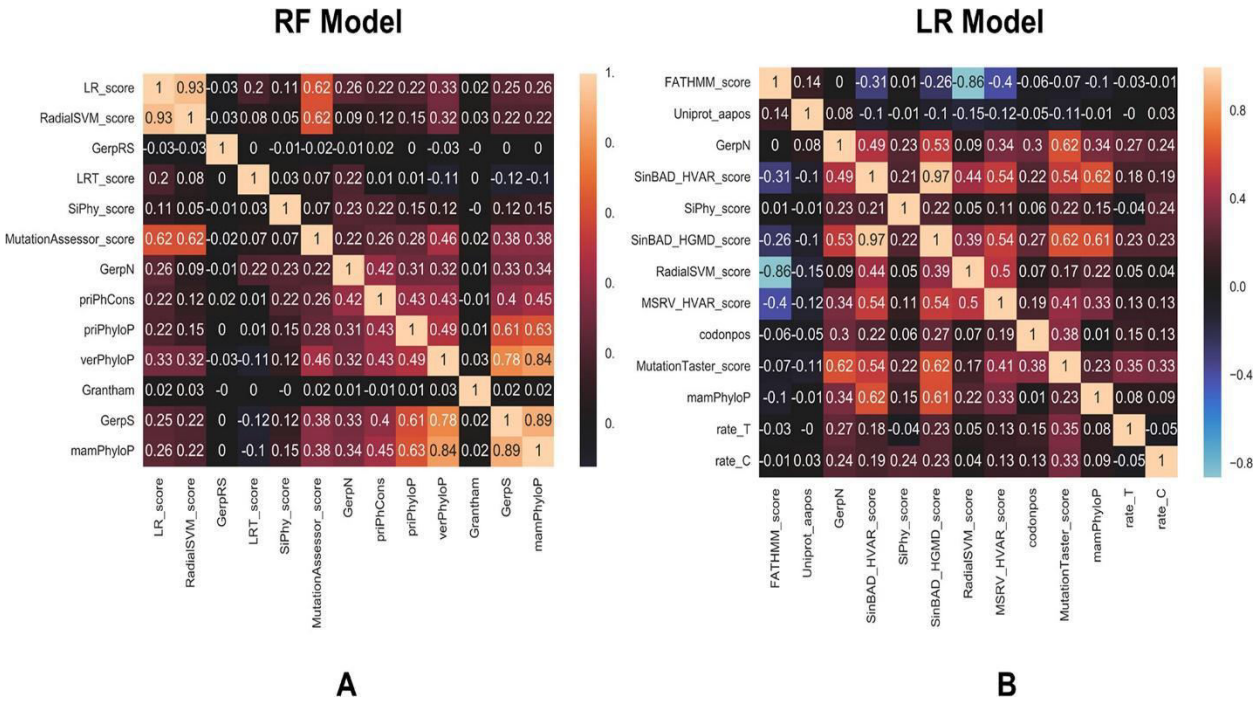


Figure 2-4. Pairwise Spearman's rank correlation coefficients between different prediction scores. The heat map of pairwise Spearman's rank correlation coefficients is shown for top ranking features in the RF model (A) and logit model respectively (B). The high-ranking features include ensemble based LR and RadialSVM scores.

Hence, these evolutionary-based features may provide orthogonal information to the ensemble scores, leading to the superior feature importance of LR and RadialSVM features.

2.4.2 Model Evaluation: Random Forest (RF) and Logistic Regression (Logit)

For the model diagnostics, we compared the predictive performance of logit and RF models as well the contribution of individual features using area under the curve (AUC) from the receiver operating characteristic (ROC) plots, in which true positive rate TPR (sensitivity) is plotted as a

function of 1 – true negative rate TNR (specificity) (Figure 2-5). The plots showed the improved performance of the RF model after the feature selection process was complete (AUC = 0.97) as compared to the original AUC = 0.85 for the initial set of features (Figure 2-5A). By analyzing the contribution of each features to the performance of RF model, we observed that GerpRS achieved as high performance in the testing data set (AUC = 0.91) as the ensemble-based LR score (AUC = 0.88) and RadialSVM score (AUC = 0.87) (Figure 2-5 B). These observations further supported our conclusions that evolutionary conservation scores can often outperform other features, including ensemble based LR and RadialSVM scores. A comparative AUC analysis in the presence and absence of top ensemble scores showed only an insignificant effect on performance of both models (Figure 2-5 C).Interestingly, we observed strong and similar performance in logit and RF models (AUC > 0.9) on the testing set when the top two ensemble features were excluded. These findings showed that ensemble-based scores and functional features based on evolutionary conservation metrics and statistical descriptors of substitution patterns can dominate feature importance ranking. Despite differences in the prediction scores, the obtained highly important features reflected a common fundamental signature, namely that mutations of evolutionarily conserved residues in functional regions are likely to be harmful and can often signify the emergence of cancer driver mutations. In simpler terms, the probabilistic evaluation of deleterious mutations that underlies most of the dominant features appeared to be also sufficient for robust classification of driver mutations in the canonical cancer data.

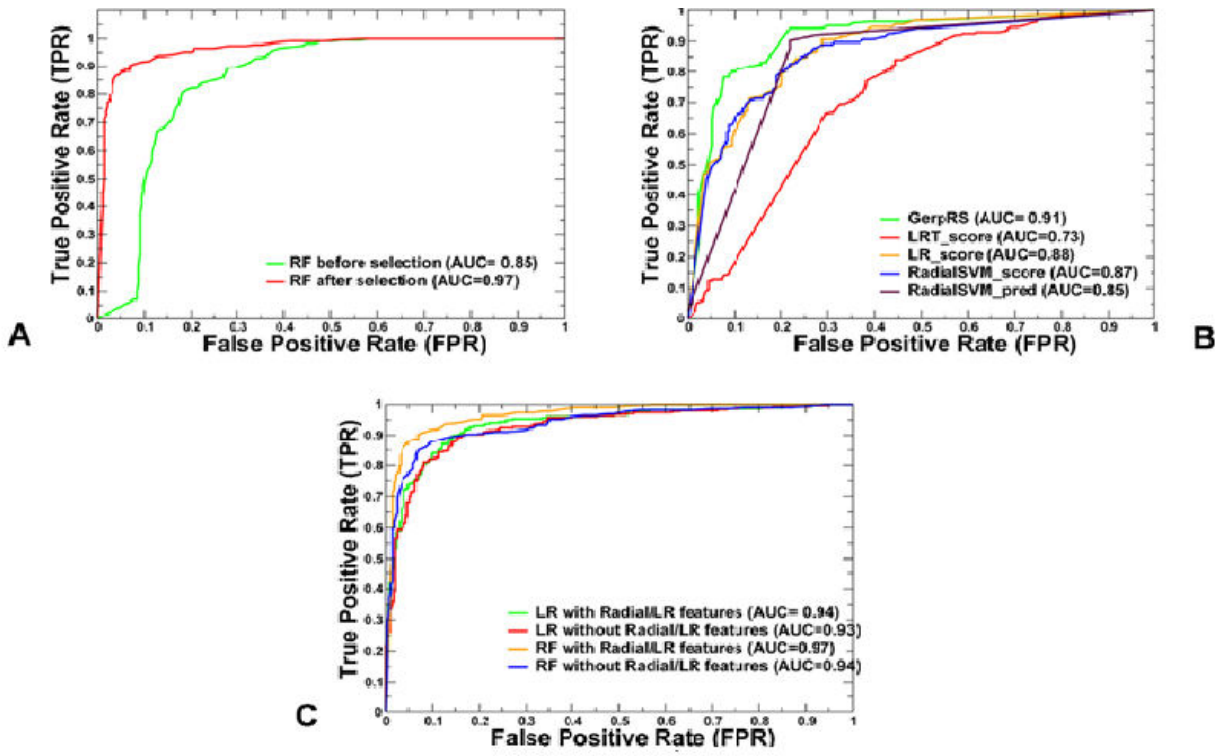


Figure 2-5. ROC plots of sensitivity (TPR) as a function of 1 – specificity, where specificity is TNR. **(A)** ROC curves for overall performance of the RF model before and after feature selection. **(B)** ROC curves for top 5 prediction scores in the RF model. The ensemble-based LR (AUC = 0.88) and RadialSVM scores (AUC = 0.85) are top performing features in the RF model. A higher AUC score indicates better performance. **(C)** Comparison of ROC curves for RF and LR models with all selected features and without the top two ensemble scores LR and RadialSVM. These plots illustrated a comparative performance of machine learning models for top prediction scores.

These findings indicated that ensemble-based scores and functional features based on evolutionary conservation measures and statistical descriptors of substitution patterns can dominate feature importance ranking. Despite differences in the prediction scores, the obtained highly important features reflected a common fundamental signature, namely that mutations of evolutionarily conserved residues in functional regions are likely to be deleterious and can often

signify the emergence of cancer driver mutations. In other words, the probabilistic evaluation of deleterious mutations that underlies most of the dominant features appeared to be also sufficient for robust classification of driver mutations in the canonical cancer data sets.

2.5 References

Adzhubei, I. A.; Schmidt, S.; Peshkin, L.; Ramensky, V. E.; Gerasimova, A.; Bork, P.; Kondrashov, A. S.; Sunyaev, S. R. A Method and Server for Predicting Damaging Missense Mutations. Nat. Methods 2010, 7, 248– 249, DOI: 10.1038/nmeth0410-248

Angermueller, C.; Parnamaa, T.; Parts, L.; Stegle, O. Deep Learning for Computational Biology. Mol. Syst. Biol. 2016, 12, 878, DOI: 10.15252/msb.20156651

Bailey, M. H.; Tokheim, C.; Porta-Pardo, E.; Sengupta, S.; Bertrand, D.; Weerasinghe, A.; Colaprico, A.; Wendl, M. C.; Kim, J.; Reardon, B.; Ng, P. K.; Jeong, K. J.; Cao, S.; Wang, Z.; Gao, J.; Gao, Q.; Wang, F.; Liu, E. M.; Mularoni, L.; Rubio-Perez, C.; Nagarajan, N.; Cortes-Ciriano, I.; Zhou, D. C.; Liang, W. W.; Hess, J. M.; Yellapantula, V. D.; Tamborero, D.; Gonzalez-Perez, A.; Suphavilai, C.; Ko, J. Y.; Khurana, E.; Park, P. J.; Van Allen, E. M.; Liang, H.; Lawrence, M. S.; Godzik, A.; Lopez-Bigas, N.; Stuart, J.; Wheeler, D.; Getz, G.; Chen, K.; Lazar, A. J.; Mills, G. B.; Karchin, R.; Ding, L.Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 2018, 173, 371, DOI: 10.1016/j.cell.2018.02.060

Biau, G. Analysis of a Random Forest Model. J. Mach. Learn. Res. 2012, 13, 1063– 1095

Carter, H.; Chen, S.; Isik, L.; Tyekucheva, S.; Velculescu, V. E.; Kinzler, K. W.; Vogelstein, B.; Karchin, R. Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational Prediction of Driver Missense Mutations. Cancer Res. 2009, 69, 6660– 6667, DOI: 10.1158/0008-5472.CAN-09-1133

Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B. E.; Sumer, S. O.; Aksoy, B. A.; Jacobsen, A.; Byrne, C. J.; Heuer, M. L.; Larsson, E.; Antipin, Y.; Reva, B.; Goldberg, A. P.; Sander, C.; Schultz, N. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery 2012, 2, 401– 404, DOI: 10.1158/2159-8290.CD-12-0095

Cheng, F.; Zhao, J.; Zhao, Z. Advances in Computational Approaches for Prioritizing Driver Mutations and Significantly Mutated Genes in Cancer Genomes. Briefings Bioinf. 2016, 17, 642– 656, DOI: 10.1093/bib/bbv068

Choi, Y.; Sims, G. E.; Murphy, S.; Miller, J. R.; Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS One 2012, 7, e46688, DOI: 10.1371/journal.pone.0046688

Chun, S.; Fay, J. C. Identification of Deleterious Mutations within Three Human Genomes. Genome Res. 2009, 19, 1553– 1561, DOI: 10.1101/gr.092619.109

Davydov, E. V.; Goode, D. L.; Sirota, M.; Cooper, G. M.; Sidow, A.; Batzoglou, S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Comput. Biol. 2010, 6, e1001025, DOI: 10.1371/journal.pcbi.1001025

Ding, L.; Wendl, M. C.; McMichael, J. F.; Raphael, B. J. Expanding the Computational Toolbox for Mining Cancer Genomes. Nat. Rev. Genet. 2014, 15, 556– 570, DOI: 10.1038/nrg3767

Ding, L.; Bailey, M. H.; Porta-Pardo, E.; Thorsson, V.; Colaprico, A.; Bertrand, D.; Gibbs, D. L.; Weerasinghe, A.; Huang, K. L.; Tokheim, C.; Cortes-Ciriano, I.; Jayasinghe, R.; Chen, F.; Yu, L.; Sun, S.; Olsen, C.; Kim, J.; Taylor, A. M.; Cherniack, A. D.; Akbani, R.; Suphavilai, C.; Nagarajan, N.; Stuart, J. M.; Mills, G. B.; Wyczalkowski, M. A.; Vincent, B. G.; Hutter, C. M.; Zenklusen, J. C.; Hoadley, K. A.; Wendl, M. C.; Shmulevich, L.; Lazar, A. J.; Wheeler, D. A.;

Getz, G.Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. Cell 2018, 173, 305, DOI: 10.1016/j.cell.2018.03.033

Dong, C.; Guo, Y.; Yang, H.; He, Z.; Liu, X.; Wang, K. iCAGES: integrated CAncer GEnome Score for comprehensively prioritizing driver genes in personal cancer genomes. Genome Med. 2016, 8, 135, DOI: 10.1186/s13073-016-0390-0

Dong, C.; Wei, P.; Jian, X.; Gibbs, R.; Boerwinkle, E.; Wang, K.; Liu, X. Comparison and Integration of Deleteriousness Prediction Methods for Nonsynonymous SNVs in Whole Exome Sequencing Studies. Hum. Mol. Genet. 2015, 24, 2125– 2137, DOI: 10.1093/hmg/ddu733

Douville, C.; Carter, H.; Kim, R.; Niknafs, N.; Diekhans, M.; Stenson, P. D.; Cooper, D. N.; Ryan, M.; Karchin, R. CRAVAT: Cancer-Related Analysis of Variants Toolkit. Bioinformatics 2013, 29, 647– 648, DOI: 10.1093/bioinformatics/btt017

Forbes, S. A.; Beare, D.; Gunasekaran, P.; Leung, K.; Bindal, N.; Boutselakis, H.; Ding, M.; Bamford, S.; Cole, C.; Ward, S.; Kok, C. Y.; Jia, M.; De, T.; Teague, J. W.; Stratton, M. R.; McDermott, U.; Campbell, P. J. COSMIC: Exploring the World's Knowledge of Somatic Mutations in Human Cancer. Nucleic Acids Res. 2015, 43, D805– D811, DOI: 10.1093/nar/gku1075

Gao, J.; Aksoy, B. A.; Dogrusoz, U.; Dresdner, G.; Gross, B.; Sumer, S. O.; Sun, Y.; Jacobsen, A.; Sinha, R.; Larsson, E.; Cerami, E.; Sander, C.; Schultz, N. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Sci. Signaling 2013, 6, pl1, DOI: 10.1126/scisignal.2004088

Gao, J.; Chang, M. T.; Johnsen, H. C.; Gao, S. P.; Sylvester, B. E.; Sumer, S. O.; Zhang, H.; Solit, D. B.; Taylor, B. S.; Schultz, N.; Sander, C. 3D Clusters of Somatic Mutations in Cancer

Reveal Numerous Rare Mutations as Functional Targets. Genome Med. 2017, 9, 4, DOI: 10.1186/s13073-016-0393-x

Garber, M.; Guttman, M.; Clamp, M.; Zody, M. C.; Friedman, N.; Xie, X. Identifying Novel Constrained Elements by Exploiting Biased Substitution Patterns. Bioinformatics 2009, 25, i54–i62, DOI: 10.1093/bioinformatics/btp190

Gonzalez-Perez, A.; Lopez-Bigas, N. Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score. Am. J. Hum. Genet. 2011, 88, 440– 449, DOI: 10.1016/j.ajhg.2011.03.004

Grantham, R. Amino acid difference formula to help explain protein evolution. Science 1974, 185, 862– 864, DOI: 10.1126/science.185.4154.862

Hinkson, I. V.; Davidsen, T. M.; Klemm, J. D.; Kerlavage, A. R.; Kibbe, W. A.; Chandramouliswaran, I. A Comprehensive Infrastructure for Big Data in Cancer Research: Accelerating Cancer Research and Precision Medicine. Front. Cell Dev. Biol. 2017, 5, 83, DOI: 10.3389/fcell.2017.00083

Hudson, T. J.; Anderson, W.; Artez, A.; Barker, A. D.; Bell, C.; Bernabe, R. R.; Bhan, M. K.; Calvo, F.; Eerola, I.; Gerhard, D. S.; Guttmacher, A.; Guyer, M.; Hemsley, F. M.; Jennings, J. L.; Kerr, D.; Klatt, P.; Kolar, P.; Kusada, J.; Lane, D. P.; Laplace, F.; Youyong, L.; Nettekoven, G.; Ozenberger, B.; Peterson, J.; Rao, T. S.; Remacle, J.; Schafer, A. J.; Shibata, T.; Stratton, M. R.; Vockley, J. G.; Watanabe, K.; Yang, H.; Yuen, M. M.; Knoppers, B. M.; Bobrow, M.; Cambon-Thomsen, A.; Dressler, L. G.; Dyke, S. O.; Joly, Y.; Kato, K.; Kennedy, K. L.; Nicolas, P.; Parker, M. J.; Rial-Sebbag, E.; Romeo-Casabona, C. M.; Shaw, K. M.; Wallace, S.; Wiesner, G. L.; Zeps, N.; Lichter, P.; Biankin, A. V.; Chabannon, C.; Chin, L.; Clement, B.; de Alava, E.; Degos, F.; Ferguson, M. L.; Geary, P.; Hayes, D. N.; Johns, A. L.; Kasprzyk, A.; Nakagawa, H.;

Penny, R.; Piris, M. A.; Sarin, R.; Scarpa, A.; van de Vijver, M.; Futreal, P. A.; Aburatani, H.; Bayes, M.; Botwell, D. D.; Campbell, P. J.; Estivill, X.; Grimmond, S. M.; Gut, I.; Hirst, M.; Lopez-Otin, C.; Majumder, P.; Marra, M.; McPherson, J. D.; Ning, Z.; Puente, X. S.; Ruan, Y.; Stunnenberg, H. G.; Swerdlow, H.; Velculescu, V. E.; Wilson, R. K.; Xue, H. H.; Yang, L.; Spellman, P. T.; Bader, G. D.; Boutros, P. C.; Flicek, P.; Getz, G.; Guigo, R.; Guo, G.; Haussler, D.; Heath, S.; Hubbard, T. J.; Jiang, T.; Jones, S. M.; Li, Q.; Lopez-Bigas, N.; Luo, R.; Muthuswamy, L.; Ouellette, B. F.; Pearson, J. V.; Quesada, V.; Raphael, B. J.; Sander, C.; Speed, T. P.; Stein, L. D.; Stuart, J. M.; Teague, J. W.; Totoki, Y.; Tsunoda, T.; Valencia, A.; Wheeler, D. A.; Wu, H.; Zhao, S.; Zhou, G.; Lathrop, M.; Thomas, G.; Yoshida, T.; Axton, M.; Gunter, C.; Miller, L. J.; Zhang, J.; Haider, S. A.; Wang, J.; Yung, C. K.; Cros, A.; Liang, Y.; Gnaneshan, S.; Guberman, J.; Hsu, J.; Chalmers, D. R.; Hasel, K. W.; Kaan, T. S.; Lowrance, W. W.; Masui, T.; Rodriguez, L. L.; Vergely, C.; Bowtell, D. D.; Cloonan, N.; deFazio, A.; Eshleman, J. R.; Etemadmoghadam, D.; Gardiner, B. B.; Kench, J. G.; Sutherland, R. L.; Tempero, M. A.; Waddell, N. J.; Wilson, P. J.; Gallinger, S.; Tsao, M. S.; Shaw, P. A.; Petersen, G. M.; Mukhopadhyay, D.; DePinho, R. A.; Thayer, S.; Shazand, K.; Beck, T.; Sam, M.; Timms, L.; Ballin, V.; Lu, Y.; Ji, J.; Zhang, X.; Chen, F.; Hu, X.; Yang, Q.; Tian, G.; Zhang, L.; Xing, X.; Li, X.; Zhu, Z.; Yu, Y.; Yu, J.; Tost, J.; Brennan, P.; Holcatova, I.; Zaridze, D.; Brazma, A.; Egevard, L.; Prokhortchouk, E.; Banks, R. E.; Uhlen, M.; Viksna, J.; Ponten, F.; Skryabin, K.; Birney, E.; Borg, A.; Borresen-Dale, A. L.; Caldas, C.; Foekens, J. A.; Martin, S.; Reis-Filho, J. S.; Richardson, A. L.; Sotiriou, C.; Thoms, G.; van't Veer, L.; Birnbaum, D.; Blanche, H.; Boucher, P.; Boyault, S.; Masson-Jacquemier, J. D.; Pauporte, I.; Pivot, X.; Vincent-Salomon, A.; Tabone, E.; Theillet, C.; Treilleux, I.; Bioulac-Sage, P.; Decaens, T.; Franco, D.; Gut, M.; Samuel, D.; Zucman-Rossi, J.; Eils, R.; Brors, B.; Korbel, J. O.; Korshunov, A.; Landgraf, P.;

Lehrach, H.; Pfister, S.; Radlwimmer, B.; Reifenberger, G.; Taylor, M. D.; von Kalle, C.; Majumder, P. P.; Pederzoli, P.; Lawlor, R. A.; Delledonne, M.; Bardelli, A.; Gress, T.; Klimstra, D.; Zamboni, G.; Nakamura, Y.; Miyano, S.; Fujimoto, A.; Campo, E.; de Sanjose, S.; Montserrat, E.; Gonzalez-Diaz, M.; Jares, P.; Himmelbauer, H.; Bea, S.; Aparicio, S.; Easton, D. F.; Collins, F. S.; Compton, C. C.; Lander, E. S.; Burke, W.; Green, A. R.; Hamilton, S. R.; Kallioniemi, O. P.; Ley, T. J.; Liu, E. T.; Wainwright, B. J. International Network of Cancer Genome Projects. Nature 2010, 464, 993– 998, DOI: 10.1038/nature08987

Ioannidis, N. M.; Rothstein, J. H.; Pejaver, V.; Middha, S.; McDonnell, S. K.; Baheti, S.; Musolf, A.; Li, Q.; Holzinger, E.; Karyadi, D.; Cannon-Albright, L. A.; Teerlink, C. C.; Stanford, J. L.; Isaacs, W. B.; Xu, J.; Cooney, K. A.; Lange, E. M.; Schleutker, J.; Carpten, J. D.; Powell, I. J.; Cussenot, O.; Cancel-Tassin, G.; Giles, G. G.; MacInnis, R. J.; Maier, C.; Hsieh, C. L.; Wiklund, F.; Catalona, W. J.; Foulkes, W. D.; Mandal, D.; Eeles, R. A.; Kote-Jarai, Z.; Bustamante, C. D.; Schaid, D. J.; Hastie, T.; Ostrander, E. A.; Bailey-Wilson, J. E.; Radivojac, P.; Thibodeau, S. N.; Whittemore, A. S.; Sieh, W. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am. J. Hum. Genet. 2016, 99, 877– 885, DOI: 10.1016/j.ajhg.2016.08.016

Iranzo, J., Martincorena, I., & Koonin, E. V. (2017). The cancer-mutation network and the number and specificity of driver mutations. doi: 10.1101/237016

Jensen, M. A.; Ferretti, V.; Grossman, R. L.; Staudt, L. M. The NCI Genomic Data Commons as an Engine for Precision Medicine. Blood 2017, 130, 453– 459, DOI: 10.1182/blood-2017-03-735654

Jiang, R.; Yang, H.; Zhou, L.; Kuo, C. C.; Sun, F.; Chen, T. Sequence-Based Prioritization of Nonsynonymous Single-Nucleotide Polymorphisms for the Study of Disease Mutations. Am. J. Hum. Genet. 2007, 81, 346– 360, DOI: 10.1086/519747

Jing, Y.; Bian, Y.; Hu, Z.; Wang, L.; Xie, X. S. Deep Learning for Drug Design: an Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era. AAPS J. 2018, 20, 58, DOI: 10.1208/s12248-018-0210-0

Kircher, M.; Witten, D. M.; Jain, P.; O'Roak, B. J.; Cooper, G. M.; Shendure, J. A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants. Nat. Genet. 2014, 46, 310– 315, DOI: 10.1038/ng.2892

Klonowska, K.; Czubak, K.; Wojciechowska, M.; Handschuh, L.; Zmienko, A.; Figlerowicz, M.; Dams-Kozlowska, H.; Kozlowski, P. Oncogenomic Portals for the Visualization and Analysis of Genome-Wide Cancer Data. Oncotarget 2016, 7, 176– 192, DOI: 10.18632/oncotarget.6128

Lehmann, K. V.; Chen, T. Exploring Functional Variant Discovery in Non-Coding Regions with SInBaD. Nucleic Acids Res. 2013, 41, e7, DOI: 10.1093/nar/gks800

Li, X., & Thirumalai, D. (2016). Interplay of Driver, Mini-Driver, and Deleterious Passenger Mutations on Cancer Progression. doi: 10.1101/084392

Liu, X.; Jian, X.; Boerwinkle, E. dbNSFP: A Lightweight Database of Human Nonsynonymous SNPs and their Functional Predictions. Hum. Mutat. 2011, 32, 894– 899, DOI: 10.1002/humu.21517

Liu, X.; Jian, X.; Boerwinkle, E. dbNSFP v2.0: Database of Human Nonsynonymous SNPs and their Functional Predictions and Annotations. Hum. Mutat. 2013, 34, E2393– E2402, DOI: 10.1002/humu.22376

Liu, X.; Wu, C.; Li, C.; Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum. Mutat. 2016, 37, 235– 241, DOI: 10.1002/humu.22932

Martelotto, L. G.; Ng, C. K.; De Filippo, M. R.; Zhang, Y.; Piscuoglio, S.; Lim, R. S.; Shen, R.; Norton, L.; Reis-Filho, J. S.; Weigelt, B. Benchmarking Mutation Effect Prediction Algorithms Using Functionally Validated Cancer-Related Missense Mutations. Genome Biol. 2014, 15, 484, DOI: 10.1186/s13059-014-0484-1

Mao, Y.; Chen, H.; Liang, H.; Meric-Bernstam, F.; Mills, G. B.; Chen, K. CanDrA: Cancer-Specific Driver Missense Mutation Annotation with Optimized Features. PLoS One 2013, 8, e77945, DOI: 10.1371/journal.pone.0077945

Min, S.; Lee, B.; Yoon, S. Deep Learning in Bioinformatics. Briefings Bioinf. 2016, 18, 851–869, DOI: 10.1093/bib/bbw068

Nakagawa, H.; Fujita, M. Whole Genome Sequencing Analysis for Cancer Genomics and Precision Medicine. Cancer Sci. 2018, 109, 513– 522, DOI: 10.1111/cas.13505

Palazon-Bru, A.; Folgado-de la Rosa, D. M.; Cortes-Castell, E.; Lopez-Cascales, M. T.; Gil-Guillen, V. F. Sample Size Calculation to Externally Validate Scoring Systems Based on Logistic Regression Models. PLoS One 2017, 12, e0176726, DOI: 10.1371/journal.pone.0176726

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 2011, 12, 2825– 2830

Raphael, B. J.; Dobson, J. R.; Oesper, L.; Vandin, F. Identifying Driver Mutations in Sequenced Cancer Genomes: Computational Approaches to Enable Precision Medicine. Genome Med. 2014, 6, 5, DOI: 10.1186/gm524

Reddy, B. Y., Miller, D. M., & Tsao, H. (2017). Somatic driver mutations in melanoma. Cancer, 123(S11), 2104–2117. doi: 10.1002/cncr.30593

Reva, B.; Antipin, Y.; Sander, C. Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics. Nucleic Acids Res. 2011, 39, e118, DOI: 10.1093/nar/gkr407

Ritchie, G. R.; Dunham, I.; Zeggini, E.; Flicek, P. Functional Annotation of Noncoding Sequence Variants. Nat. Methods 2014, 11, 294– 296, DOI: 10.1038/nmeth.2832

Schwarz, J. M.; Rodelsperger, C.; Schuelke, M.; Seelow, D. MutationTaster Evaluates Disease-Causing Potential of Sequence Alterations. Nat. Methods 2010, 7, 575– 576, DOI: 10.1038/nmeth0810-575

Shihab, H. A.; Gough, J.; Cooper, D. N.; Stenson, P. D.; Barker, G. L.; Edwards, K. J.; Day, I. N.; Gaunt, T. R. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions Using Hidden Markov Models. Hum. Mutat. 2013, 34, 57– 65, DOI: 10.1002/humu.22225

Sim, N. L.; Kumar, P.; Hu, J.; Henikoff, S.; Schneider, G.; Ng, P. C. SIFT Web Server: Predicting Effects of Amino Acid Substitutions on Proteins. Nucleic Acids Res. 2012, 40, W452– W457, DOI: 10.1093/nar/gks539

Weinstein, J. N.; Collisson, E. A.; Mills, G. B.; Shaw, K. R.; Ozenberger, B. A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J. M. The Cancer Genome Atlas Pan-Cancer Analysis Project. Nat. Genet. 2013, 45, 1113– 1120, DOI: 10.1038/ng.2764

Wu, J.; Wu, M.; Li, L.; Liu, Z.; Zeng, W.; Jiang, R. dbWGFP: A Database and Web Server of Human Whole-Genome Single Nucleotide Variants and their Functional Predictions. Database 2016, 2016, baw024, DOI: 10.1093/database/baw024

Yuan, Y.; Shi, Y.; Li, C.; Kim, J.; Cai, W.; Han, Z.; Feng, D. D. DeepGene: An Advanced Cancer Type Classifier Based on Deep Learning and Somatic Point Mutations. BMC Bioinf. 2016, 17, 476, DOI: 10.1186/s12859-016-1334-9

Zhang, J.; Baran, J.; Cros, A.; Guberman, J. M.; Haider, S.; Hsu, J.; Liang, Y.; Rivkin, E.; Wang, J.; Whitty, B.; Wong-Erasmus, M.; Yao, L.; Kasprzyk, A. International Cancer Genome Consortium Data Portal - A One-Stop Shop for Cancer Genomics Data. Database 2011, 2011, bar026, DOI: 10.1093/database/bar026

Zhang, L.; Tan, J.; Han, D.; Zhu, H. From Machine Learning to Deep Learning: Progress in Machine Intelligence for Rational Drug Discovery. Drug Discovery Today 2017, 22, 1680–1685, DOI: 10.1016/j.drudis.2017.08.010

Zhou, J.; Troyanskaya, O. G. Predicting Effects of Noncoding Variants with Deep Learning-Based Sequence Model. Nat. Methods 2015, 12, 931– 934, DOI: 10.1038/nmeth.3547

3 Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for Classification and Biomolecular Modeling of Cancer Driver Mutations

Steve Agajanian, Oluyemi Odeyemi, Gennady Verkhivker

Author's Contribution

YO and SA performed the research. YO, SA and GV analyzed the results and wrote the manuscript.

3.1 Abstract

Introduction: Incorporation of machine learning with genome-wide association studies in the prediction of the structural and clinical significance of cancer driver variants are at the forefront of our understanding of the molecular dynamics of cancer.

Methods: In this study, we integrated two ensemble tree-based classifiers (random forest and gradient boosting machines) with deep convolutional neural networks (CNN) for prediction of cancer driver mutations in the genomic datasets. The feasibility of CNN in using raw nucleotide sequences for classification of cancer driver mutations was initially explored by employing label encoding, one-hot encoding, and embedding to preprocess the DNA information. These classifiers were benchmarked against their tree-based alternatives to evaluate the performance on a relative scale. We then integrated DNA-based scores generated by CNN with various categories of conservational, evolutionary and functional features into a generalized random forest classifier.

Results: The results of this study have demonstrated that CNN can learn high-level features from genomic information that are complementary to the ensemble-based predictors often employed for the classification of cancer mutations.

Conclusion: By combining deep learning-generated score with only two main ensemble-based functional features, we can achieve a superior performance of various machine learning classifiers.

3.2 Introduction

Recent advances in next-generation sequencing (NGS) platforms have yielded a large amount of cancer genomic data which can be leveraged for classifying driver genes. With the increasing number of validated driver mutations, researchers began utilizing machine learning models to predict new cancer-linked drivers. The explosion of data generated in genome-wide association study (GWAS) and next-generation sequencing (NGS) has been the motivation for the development of large bioinformatics data resources such as The Cancer Gene Census of the Catalog of Somatic Mutations in Cancer (COSMIC) database (http://cancer.sanger.ac.uk), Cancer Genome Atlas (TCGA), Genomics Data Commons Portal (https://portal.gdc.cancer.gov/), the International Cancer Genome Consortium (ICGC) and formation of international cancer genomic projects. (Forbes et al., 2015; Weinstein et al., 2013; Jensen et al., 2017; Hudson et al., 2010; Zhang et al., 2011; Klonowska et al., 2016; Hinkson et al., 2017). COSMIC has evolved from ~290 well-characterized cancer genes (Futreal et al., 2004) to more than 500 entries (Forbes et al., 2015) where some cancer genes can be commonly mutated across cancer types, while other genes are mostly cancer-specific. The Cancer Genomics Portal (cBioPortal) (https://www.cbioportal.org/) is an open-access resource for big data cancer genomics analyses (Cerami et al., 2012; Gao et al., 2013). These datasets have allowed for robust genome-wide analyses of genetic variation in various cancer types (Poulos and Wong, 2018). A comparatively small percentage of somatic variants known as driver mutations have substantial biological effects and can be amassed over a period of time as a

result of a range of mutational processes, rather than inherited (Haber and Settleman, 2007; Lawrence et al., 2013; Vogelstein et al., 2013). A robust analysis of cancer-linked driver genes and mutations has provided a classification of 751,876 distinct missense mutations, yielding a dataset of 3,442 functionally validated driver variations (Bailey et al., 2018). Another substantial dataset of ~1,050 experimentally tested and functionally validated drivers (Ng et al., 2018) has expanded our knowledge of cancer-linked variants in oncogenes and tumor suppressor genes. TCGA organized the Multi-Center Mutation Calling in Multiple Cancers (MC3) network project which produced a complete and consistent collection of somatic mutation calls for the ~10,400 tumor samples data (Ellrott et al., 2018). Computational strategies that evaluate the impact of somatic mutations are often identified by types of input features, models, prediction targets such as driver gene, statistical and computational assumptions (Gonzalez-Perez et al., 2013; Cheng et al., 2016).

Several statistical and ML-based somatic variant callers are now available for somatic mutation identification and detection, such as VarScan2 (Koboldt et al., 2012), Strelka2 (Kim et al., 2018), and SomaticSniper (Larson et al., 2012). DeepVariant (deep CNN strategy) can characterize variants in NGS data by identifying statistical relationships around presumptive variant sites (Poplin et al., 2018). To enable standardized somatic variant processing from cancer sequencing data, deep learning (DL) approach and tree-based random forest (RF)were utilized, showing that these ML methodologies could achieve high and similar classification performance across all processed variant classes (Ainscough et al., 2018)

Several computational techniques have been proposed for cancer driver genes prediction. Some of these methods use cohort-based analysis to identify driver genes such as MuSiC (Dees et al., 2012), OncodriveFM (Gonzalez-Perez and Lopez-Bigas, 2012), OncodriveFML (Mularoni

et al., 2016) and OncodriveCLUST (Tamborero et al., 2013). The success of hybrid methods for scoring coding variants has indicated that the integration of different tools may enhance predictive accuracy for both coding and non-coding variants (Li et al., 2015). DeepDriver- A deep learning-based method predicts driver mutations by CNN trained with a mutation-based feature matrix constructed using similarity metrics (Luo et al., 2019). Since various approaches are often found to predict unique or partially overlapping subsets of cancer driver mutations, a consensus-based strategy was recently proposed, showing considerable promise and outperforming the individual approaches (Bertrand et al., 2018). An integrated ML-based evaluation framework for the analysis of driver gene predictions compared to the performance of these approaches, indicating that the driver mutations predicted by each tool can differ significantly(Tokheim C. et al., 2016; Tokheim C. J. et al., 2016).

Computational strategies formulated to characterize driver genes have become increasingly crucial to facilitate an automated evaluation of biological and clinical effects (Raphael et al., 2014; Gnad et al., 2013; Martelotto et al., 2014; Cheng et al., 2016; Ding et al., 2014). Functional computational prediction methods include Functional Analysis Through Hidden Markov Models (FATHMM) (Shihab et al., 2013), Sorted Intolerant From Tolerant (SIFT) (Sim et al., 2012), PolyPhen-2 (Adzhubei et al., 2010), Mutation Assessor (Reva et al., 2011), MutationTaster (Schwarz et al., 2010) and Protein Variation Effect Analyzer (PROVEAN) (Choi et al., 2012). Cancer-specific High-throughput Annotation of Somatic Mutations (CHASM) (Carter et al., 2009; Douville et al., 2013; Masica et al., 2017), Cancer Driver Annotation (CanDrA) (Mao et al., 2013), and FATHMM (Shihab et al., 2013). Many new approaches have recently addressed a problem of locating driver mutations within the non-coding genome regions (Piraino and Furney, 2016).To consolidate functional annotation for

SNVs discovered in exome sequencing studies, a database of human non-synonymous SNVs (dbNSFP) was developed (Liu et al., 2011, 2013, 2016; Dong et al., 2015; Wu et al., 2016). This resource allows for the computation of a total of 48 functional prediction scores for each SNV, including 32 functional prediction scores by 13 approaches and 15 conservation features (Wu et al., 2016). In our recent investigation, two cancer-specific machine learning classifiers were proposed that utilized 48 functional scores from the dbWGFP server in the classification of cancer driver mutations (Agajanian et al., 2018).

In this study, we explore and integrate RF and deep convolutional neural network (D-CNN) ML methods for the classification and prediction of cancer driver genes. Initially, we explore the feasibility of CNN models to classify cancer driver mutations directly from raw nucleotide sequence information without the predetermined functional scores. The performance of these classifiers was compared with tree-based classifiers (RF and gradient boosting machines -GBM) algorithms to provide a comparative evaluation of different classification models. These raw sequence-derived scores are advantageous because they can be obtained for any mutation with a known chromosome and position, whereas the functional scoring features can be limited to subsets of genomic mutations. By designing a practical classification algorithm that could leverage information from raw DNA nucleotides, classifiable mutations domain can be immensely broadened resulting in more general and robust ML tools. The results of this study reveal that CNN models can learn high importance features from genomic information that are complementary to the ensemble-based predictor scores traditionally used in ML classification of cancer mutations. We show that integration of the DL-derived predictor score with only several ensemble-based features can improve performance in capturing driver mutations across a

spectrum of ML classifiers and recapitulate the results obtained with a large number of functional features.

3.3 Materials and Methods

3.3.1 Data Source

Our primary data sources are cBioPortal (https://www.cbioportal.org/), University of California, Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/) and dbWGFP web server CBioPortal is a publicly curated database hosted by Center for Molecular Oncology at Memorial Sloan Kettering Cancer Center (MSK) (Cerami et al., 2012 & Gao et al., 2013). UCSC Genome Browser is an open access repository offering genome sequence data from a variety of vertebrate and invertebrate species and major model organisms, integrated with a large collection of aligned annotations. It is hosted by the University of California, Santa Cruz (UCSC) (Tyner et al., 2017). The dbWGFP is a web server of human whole-genome single nucleotide variants and their functional predictions (Wu et al., 2016).

3.3.2 Dataset and Feature Selection

In our previous work, (Agajanian et al., 2018) we used RF classifier to predict cancer-linked drivers using two consolidated golden datasets (Mao et al., 2013; Martelotto et al., 2014). In this study, we expanded this dataset by including the passengers from the missense mutations analysis and predicted cancer driver mutations in Cbioportal database (Agajanian et al., 2018). Based on our prior analysis, we generated a dataset consisting of functionally validated 12,941 passenger mutations and 6,389 cancer driver mutations. The passenger/driver classifications for 2,570 of these mutations were present in the two previous golden datasets, and our RF classifier made predictions on the remaining 16,760 missense mutations from the

Cbioportal database. From the performance of our initial model (Agajanian et al., 2018), we hypothesize that the incorporation of the missense mutations in the Cbioportal database with the two golden datasets would produce an informative dataset for this study. The features used for the initial random forest predictions were retrieved from dbWGFP web server (Wu et al., 2016) of functional predictions for human whole-genome single nucleotide variants. A total of 32 sequence-based, functional and evolutionary features characterized in our prior study (Agajanian et al., 2018) were initially used for ML studies with the new dataset of cancer mutations. In cancer driver mutation predictions, traditional input data comprise of unique features that cannot be directly used in CNN models due to their lack of spatiality in the data. Using chromosomal position that corresponded to the mutated nucleotide, we could obtain the surrounding nucleotides of the mutation of interest to perform classification with only this raw string of nucleotides. For easier representation of the original nucleotide and its mutated version, we placed two nucleotide sequences on top of each other, one containing the original nucleotide, and the other contained the mutated copy. This would only result in a one nucleotide difference between the two, allowing to effectively utilize CNN's sliding window mechanism.

The schematic workflow diagram of the CNN approach employed in this study is presented in Figure 3-1

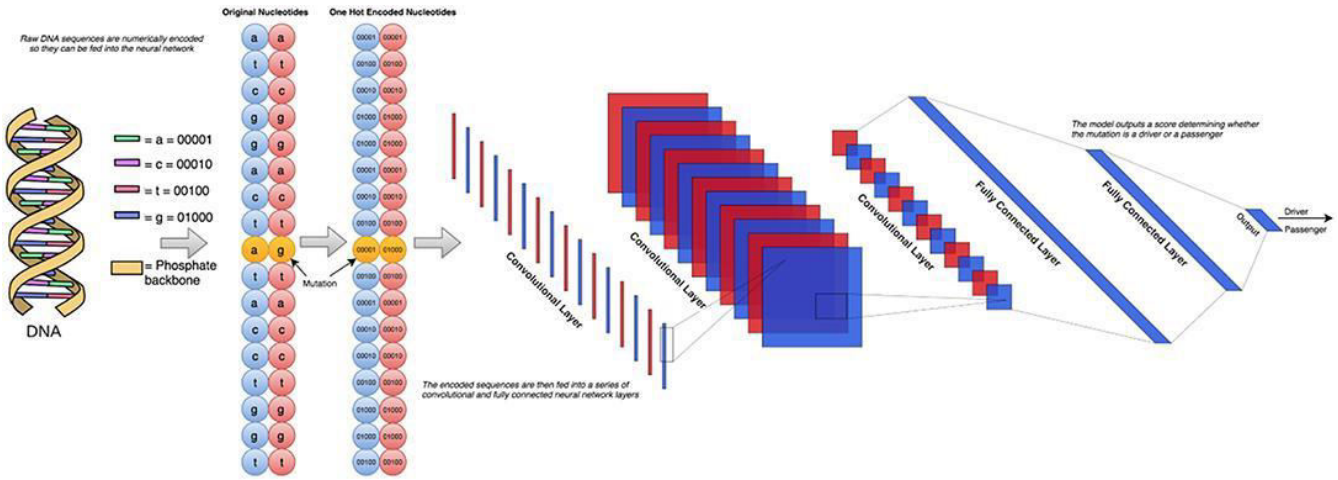


Figure 3-1. The schematic workflow diagram of the CNN approach employed in this study. To determine the optimal architecture, we performed a grid search over a total of 72 different neural network architectures. These 72 architectures consisted of between 1 and 3 convolutional layers and 1–3 fully connected layers following. The number of nodes in each of these layers was also varied between 2 and 256 in powers of 2. The simplest architecture covered in this search contains 1 convolutional layer with 2 filters feeding into 1 fully connected layer with 2 nodes, and the most complex would have 3 convolutional layers feeding into 3 fully connected layers, all containing 256 nodes.

To generate this dataset, we parsed information from University of California, Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/) (Tyner et al., 2017) which takes a chromosome (CHR) and a position (POS) on that chromosome as arguments and returns all nucleotides within the sequence. To generate this dataset, we parsed information from the University of California, Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/) (Tyner et al., 2017) which takes a chromosome (CHR) and a position (POS) on that chromosome as arguments and returns all nucleotides within the sequence. Using the dataset containing 6,389 driver mutations and 12,941 passengers, we were able to obtain five different datasets of various window sizes around each given CHR/POS pair. The explored window sizes (10, 50, 100, 500,

and 5,000) produced nucleotide strings of length 21, 101, 201, 1,001, and 10,001, respectively. To represent the type of mutation (A->C, A->G, etc.) we stacked two of the same nucleotide sequences on top of each other, having one contains the original nucleotide at the position passed in initially, and the other containing the mutated version (Figure 2A). This computation yielded a total input matrix size of (2, 21), (2, 101), (2, 201), (2, 1001), and (2, 10001), respectively. Three various preprocessing techniques were then applied to the dataset to allow it to be passed into the CNN model in the numerical form: label encoding (Figure 2B), one-hot encoding (Figure 2C; Goh et al., 2017), and embedding (Figure 2D). Label encoding works by designating each nucleotide its own unique ID (A = 0, C = 1, G = 2, T = 3) This introduces a ranking system on the nucleotide sequences that may have implications for the neural network learning (Figure 2B). This technique was implemented using the Scikit-learn LabelEncoder package for the Python programming language. We also tried introducing dummy variables (one-hot encoding) to the dataset by designating each nucleotide its own bit encoded string (A = [0,0,0,0,1], C= [0,0,0,1,0], G =[0,0,1,0,0], T = [0,1,0,0,0) (Figure 2C). This tends to be a favorable preprocessing function for weight-based classifiers because no ranking system is imposed on the samples. This approach tends to be the representation choice for discrete features due to its interpretation by default. Because each nucleotide gets its own index in a 5-bit string, a 1 in any particular index means that nucleotide is present in that location. For example, since A = [0,0,0,0,1], this can essentially be read as “There are 0 ‘n,' 0 ‘g,' 0 ‘t,' 0 ‘c,' and 1 ‘a' nucleotides present at this location.” Since the one-hot encoding preprocessing technique complicates the string, the resulting dimensionalities were (2, 105), (2, 505), (2, 1005), (2, 5005), and (2, 50005), respectively.

The final preprocessing method employed for the DNA sequences involved learned embeddings created with the word2vec algorithm (Mikolov et al., 2013). This approach analyzes the

sequential context of the nucleotides assigning them a numeric representation in vector space. Using this representation, the nucleotide segments with similar meaning in the word2vec model would yield similar vectors in an N-dimensional representation. This technique was implemented using the Word2Vec model from the genism library for the Python programming language. Since the vocabulary in this application is fairly small, consisting of only 5-bit components, we chose to convert the nucleotide to 2-dimensional vectors which are sufficient to effectively encode this set. This resulted in the input sizes (2, 42), (2, 202), (2, 402), (2, 2002), and (2, 20002), respectively (Figures 3-1,3-2). The implementation and execution of these three preprocessing techniques provides adequate and efficient nucleotide representations for the CNN classifier.

3.3.3 Machine Learning Models

We used and compared the performance of CNN models and ensemble(tree-based) classifiers. For the ensemble methods, we adopted our earlier established protocol for obtaining hyper-parameters (Agajanian et al., 2018). The model training and tuning were done using the Sklearn open-source ML library for the Python programming language (Pedregosa et al., 2011; Biau, 2012). The Keras framework was used for training, validation, and testing of CNN models (Erickson et al., 2017). The dataset was partitioned into training and test set in a 70%-30% ratio. Due to the skewness of the label in favor of passenger mutation, oversampling and under-sampling techniques were used to address potential bias in the prediction models (Blagus and Lusa, 2013). To handle for overfitting in our models, we performed five-fold cross-validation, splitting the training set up into 5 equal-sized subsets. The model trains on four subsets and predicts the fifth. This is repeated five times so that each of the five subsets has been predicted on.

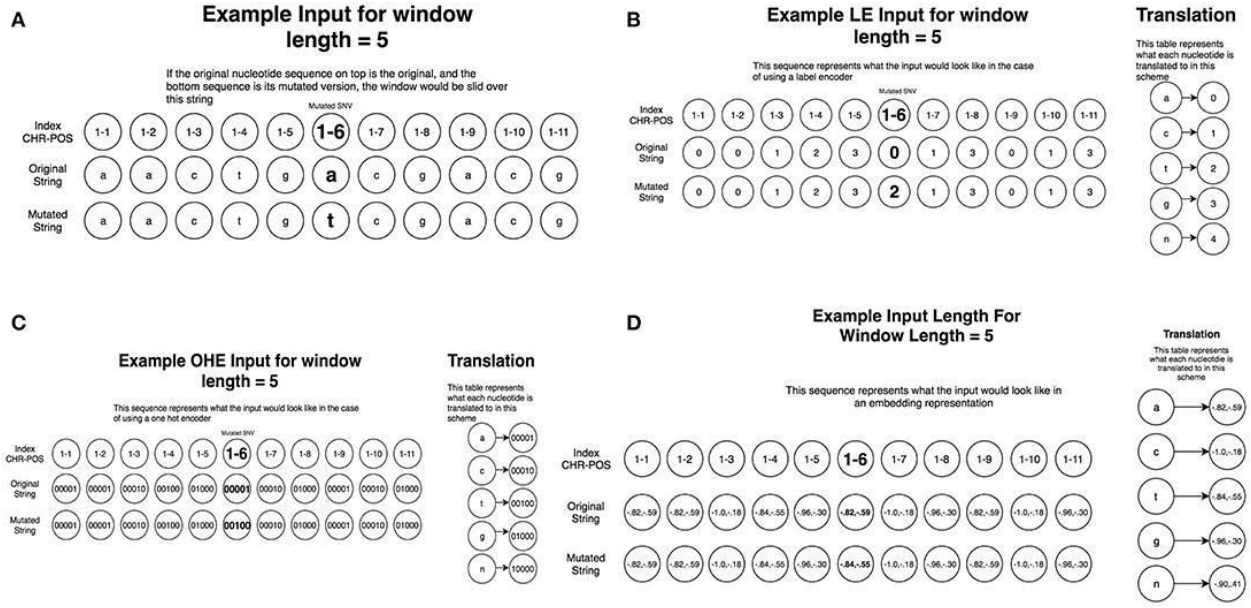


Figure 3-2. Preprocessing of the nucleotide information for CNN machine learning of cancer driver mutations. Two different preprocessing techniques were then applied to the dataset to allow it to be passed into the CNN model in the numerical form: label encoding and one-hot encoding. **(A)** A schematic diagram of window sliding protocol. To represent the original nucleotide and its mutated version, two nucleotide sequences are placed on top of each other, one containing the original string, and the other containing the mutated version. This representation allows to utilize the sliding window format of the CNN models. **(B)** A schematic diagram of label encoding preprocessing protocol. Label encoding assigns each nucleotide its own unique ID (A->0, C->1 etc.) This imposes an ordering on the nucleotide sequences. **(C)** A schematic diagram of one hot encoding preprocessing protocol. One-hot encoding assigns each nucleotide its own bit encoded string (A -> [0,0,0,0,1], C-> [0,0,0,1,0]). This tends to be a favorable preprocessing function for weight-based classifiers because no artificial ordering is imposed on the samples. **(D)** A schematic diagram of the embedding preprocessing scheme created with the word2vec algorithm.

A workflow diagram of the CNN approach (Figure 3-1) was engineered to determine the optimal architecture. For this, we performed a grid search over a total of 72 different neural network architectures. These 72 architectures consisted of between 1 and 3 convolutional layers and 1–3 fully connected layers following. The number of nodes in each of these layers was also varied between 2 and 256 in powers of 2. The simplest architecture covered in this search contains 1 convolutional layer with 2 filters feeding into 1 fully connected layer with 2 nodes, and the most complex would have 3 convolutional layers feeding into 3 fully connected layers, all containing 256 nodes. The ReLU activation function was used, which returns max (0, X). All 72 different architectures (Table 3-1) were tested using this cross-validation algorithm and the architecture that had the highest F1 score across all 3-folds was chosen. Our neural networks were trained for 100 epochs, which means that they will pass through the entire dataset 100 times to complete their training. In between each epoch, the model recorded its predictions on the validation fold, and the epoch with the best performance on the validation set was recorded. The best architecture was used for predictions on the test set. Dropout layers was added in between layers so that inputs into a layer are randomly set to 0 with a certain probability. This prevents overfitting in the model, forcing it to learn without random features present.

To evaluate the performance of each model, several metrics such as accuracy, precision, recall and F1 score were calculated to assess the performance of classification models. These parameters are defined as follows:

$$Accuracy = \frac{TP + TN}{all}; \quad Precision = \frac{TP}{TP + FP} \tag{3-1}$$

$$Recall = \frac{TP}{TP + FN}; \quad F_1 = 2\,\frac{Precision \times Recall}{Precision + Recall} \tag{3-2}$$

True Positive (TP) and True Negative (TN) are defined as the number of mutations that are classified correctly as driver and passenger mutations, respectively. False Positive (FP) and False Negative (FN) are defined as the number of mutations that are misclassified into the other mutational classes.

Table 3-1. The parameters of displayed CNN architectures in classification of cancer driver mutations

Architecture	No of Layers	No of Nodes Per Layer
0	2	32,2
1	3	16,8,2
2	3	16,16,2
3	3	32,16,2
4	3	32,8,2
5	3	64,32,2
6	3	64,16,2
7	4	64,64,16,2
8	4	128,64,16,2
9	4	128,64,32,2
10	5	128,64,32,16,2

Precision is defined as the proportion of positive samples the model predicts correctly (true positives) divided by the true positives plus the false positives. Recall is defined as true positives divided by true positives plus false negatives. The model performance was evaluated using receiver operating characteristic area under the curve. The receiver operating curve (ROC) is a graph where sensitivity is plotted as a function of 1-specificity.

The area under the ROC is denoted AUC. The sensitivity or true positive rate (TPR) is defined as the percentage of non-neutral mutations that are correctly identified as driver mutations:

$$Sensitivity(TPR) = \frac{TP}{TP + FN} \qquad (3\text{-}3)$$

The specificity or true negative rate (TNR) is defined as the percentage of mutations that are correctly identified as passengers:

$$Specificity(TNR) = \frac{TN}{TN + FP} \qquad (3\text{-}4)$$

These metrics allow us to differentiate models by providing assessment options to properly evaluate the performance of a model. F1 score, recall and precision were the primary discriminatory metrics we used to evaluate classification performance. Under this data distribution, a model that only predicted passengers would yield an accuracy of 66.95%, but an F1 score of 0. In the case that two models exhibited the same F1 score, we used the AUC score as an alternative evaluation measure. The AUC score summarizes the trade-off between the TPR and FPR for a predictive model using a number of probability thresholds. AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1). A powerful classifier learns a likelihood function that consistently maps instances of the negative class to likelihoods lower than the positive class. A model that is reliable able to do this would receive an AUC of 1, whereas a model that only predicted the negative class would also receive an AUC of 0. They are appropriate when the observations are balanced between each class, whereas precision-recall curves are appropriate for imbalanced datasets.

3.4 Results

3.4.1 Deep Learning Classification of Cancer Driver Mutations from Nucleotide Information

We started by trying to recapitulate our predictions by using a number of CNN architectures informed by raw nucleotide sequence data that assessed the ability to make predictions based mainly on raw genomic data. The combination of the 3 various preprocessing methods enabled us to select the most informative representation of the nucleotides. The one-hot encoded sequences produced the model with the best performance, and for simple representation we reported only the dimensions and performance of the one-hot encoded model. This preprocessing model resulted in input matrices of size (2, 105), (2, 505), (2, 1005), (2, 5005), and (2, 50005) corresponding to the different window sizes (10, 50, 100, 500, 1,000) surrounding the original nucleotide. Noteworthily, that the word2vec algorithm also learned meaningful representations of the nucleotides. The missing place indicator, "n," was predictably separated from the original nucleotides, which were arranged in 2 neat clusters (Figure 3D). Cluster 1 consisted of the tyrosine (T) and adenine (A) nucleotides, and cluster 2 comprised of cytosine (C) and guanine(G) nucleotides. These two clusters are easily identified due to the fact that their constituent components are very close to each other while simultaneously being far away from the other cluster.

We used the 72 different deep learning architectures (Table 3-1) and the results for the window size of 10 are presented since they revealed more variance (Figure 3-3). The figures below display the 10 best performing models out of the 72 models. The training accuracy continued to increase for the duration of training (Figure 3-3A), while on the validation testing set of cancer

mutations, the best DL architecture achieved an average validation accuracy of 86.68% with an F1 score of 0.61 (Figure 3-3B).

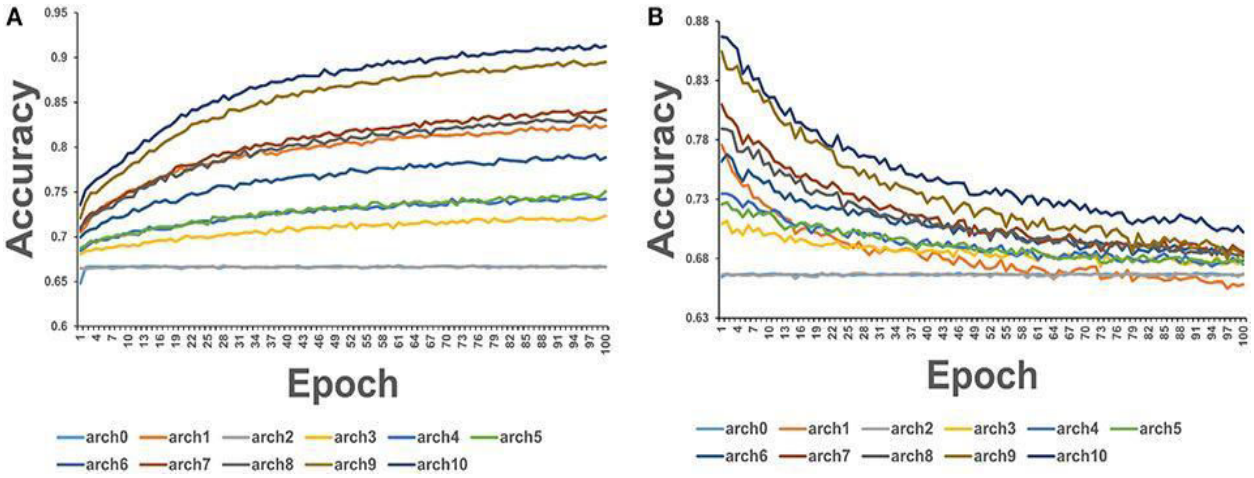


Figure 3-3. The average accuracy of CNN models using exclusively nucleotide information. **(A)** Average accuracy across all 3-folds on an epoch by epoch basis on the training set with the sliding window size = 10. **(B)** Average accuracy across all 3-folds on an epoch by epoch basis on the validation set with the sliding window size = 10.

Noteworthily, we discovered that the CNN model seemed to learn early on, overfitting with each successive epoch (Figure 3B). In fact, the model had its highest validation accuracy on the first epoch, and subsequently declined after. Furthermore, the AUC and F1 scores of the model plateaued all through the computation process. This is further contextualized by the ensembled method's performance on the same dataset. The RF classifier exhibited an average validation accuracy of 69.86% and F1 measure of 0.58, and the GBM classifier recorded an average validation accuracy of 66.59% and an F1 measure of 0.57. Although we investigated a number of different architectures using various nucleotide-encoding protocols, a direct brute-force application of DL/CNN models to predict driver mutations only as a function of surrounding nucleotides appeared to be challenging. As a result, we employed a number of

informative features to recapitulate the level of robust performance achieved in our earlier work with functional features and sequence-based conservation (Agajanian et al., 2018).

Initially, we used the random forest classifier on the cancer mutation dataset with conservation and functional features retrieved from dbWGFP server and used in our earlier study (Agajanian et al., 2018). A repository of human non-synonymous SNVs (dbNSFP) was developed with the goal of providing resources for disease-causing mutations analysis (Liu et al., 2011, 2013, 2016; Dong et al., 2015; Wu et al., 2016) storing ~8.5 billion possible human whole-genome SNVs, with capabilities to compute a total of 48 functional prediction scores for each SNV, including 15 conservation features from four various tools including ensemble-based features MSRV, RadialSVM and LR scores; and 32 functional prediction scores by 13 approaches. Evolutionary scores refer to scores providing different conservation measures of a given nucleotide site across multiple species. Whereas, functional prediction scores refer to scores that predict the likelihood of a given SNV to cause a deleterious functional change in the protein. Some of the features such as GWAVA, Grantham, Gerp++, SiPhy, PhyloP and CADD can evaluate SNVs spreading over the whole genome. Whereas, other scores such as FATHMM, SIFT, PolyPhen, SinBaD, LRT, RadialSVM, LR, MSRV, Mutation Assessor and MutationTaster can be applied only to SNVs in the protein coding regions. The ensemble-based scores LR and RadialSVM are integrated predictors that used ML methods to combine information from ten individual component scores (LRT, FATHMM, SiPhy, PhyloP, SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, Gerp++, MutationTaster, Mutation Assessor) (Agajanian et al., 2018).

In this baseline experiment we analyzed performance of 32 input predictors on the expanded dataset (Figure 3-4A). Similar to our earlier investigation (Agajanian et al., 2018), we discovered that the ensemble-based scores RadialSVM and LR considerably overshadowed the contributions

of other features (Figure 3-4). By incorporating the DL score in the original features set, we applied the RF model for predicting cancer driver mutations with this expanded set of predictors. The first question was to evaluate the feature importance of the RF model with the DL score included and determine whether the nucleotide-based scoring predictors can contribute to the prediction performance in a significant way (Figure 3-4). In the second step of RF classification experiments, we added DL score to the original list of 32 features (Figure 3-4B). Interestingly, the DL score ranked third following the ensemble-based LR and RadialSVM scores (Figure 3-4B). Moreover, it was evident that these three feature scores completely dominated feature importance distribution, with the DL score contributing almost as much as the ensemble-based RadialSVM feature (Figure 3-4B). Quite remarkably, the DL-based score derived by CNN exclusively from primary nucleotide information can deliver significant information content.

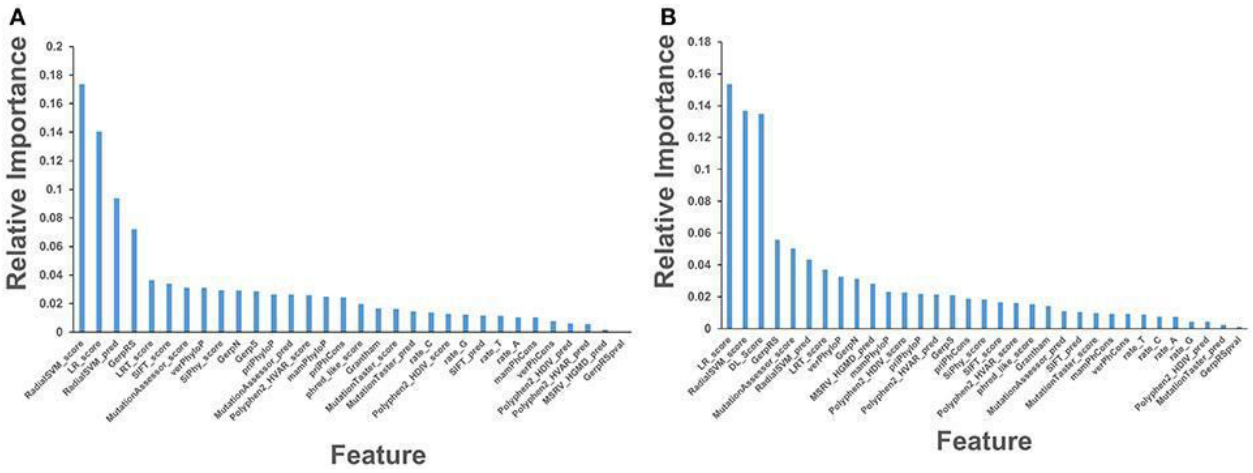


Figure 3-4. **(A)** Feature importance of 32 functional and sequence conservation features with DL score feature produced by CNN model excluded. **(B)** Feature importance of 33 features with the DL score included in the RF classification. The feature importance values are shown in blue filled bars and annotated. Feature importance is measured using the information value and weight of evidence criteria.

3.4.2 Incorporation of CNN Predictions with Ensemble-Based Predictors in Cancer Driver Mutations Models

From the preliminary assessment, we analyzed feature selection again with the goal of recapitulate similar accuracy with only eight predictors: LR, LRT, RadialSVM, DL, GerpRS, GerpN , verPhyloP, and SiPhy scores (Figure 3-5A). The eight-predictor RF model produced a similar ranking in which the ensemble-based scores and DL score contributed the most (Figure 3-5A). Evolutionary conservation scores derived from multiple sequence alignments and reflecting functional specificity, such as SiPhy (Garber et al., 2009) and GerpRS (Davydov et al., 2010) also showed significant information score values (Figure 3-5A). Subsequently, we tested the performance of the RF model and feature importance by performing ML of cancer driver mutations using only top three predictors (Figure 3-5B)

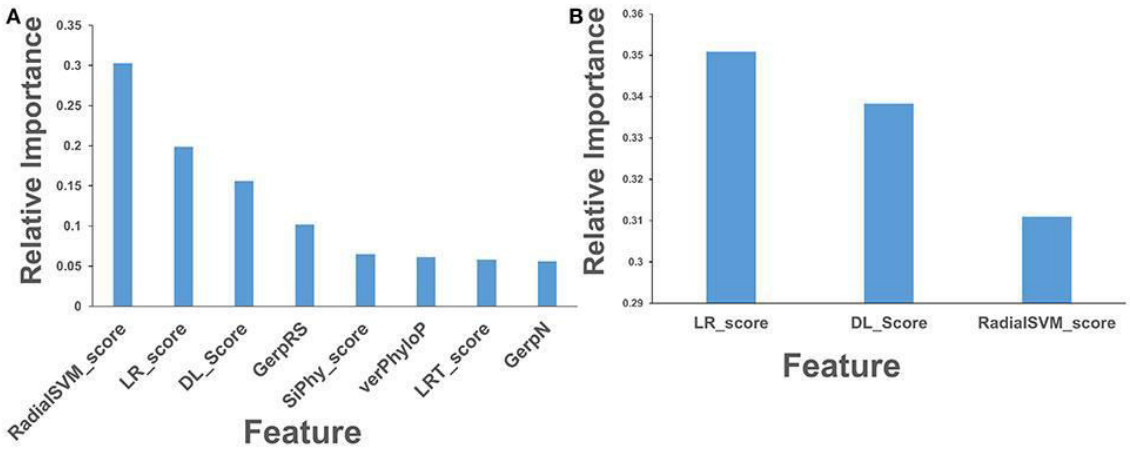


Figure 3-5. **(A)** Feature importance ranking based on RF classification with only 8 most informative features. **(B)** Feature importance ranking based on RF classification with only top three predictors that included ensemble-based RadialSVM, LR scores, and DL score produced by CNN model.

The predictability of the RF models with various set of predictors was evaluated using area under the curve (AUC) plots (Figure 3-6).

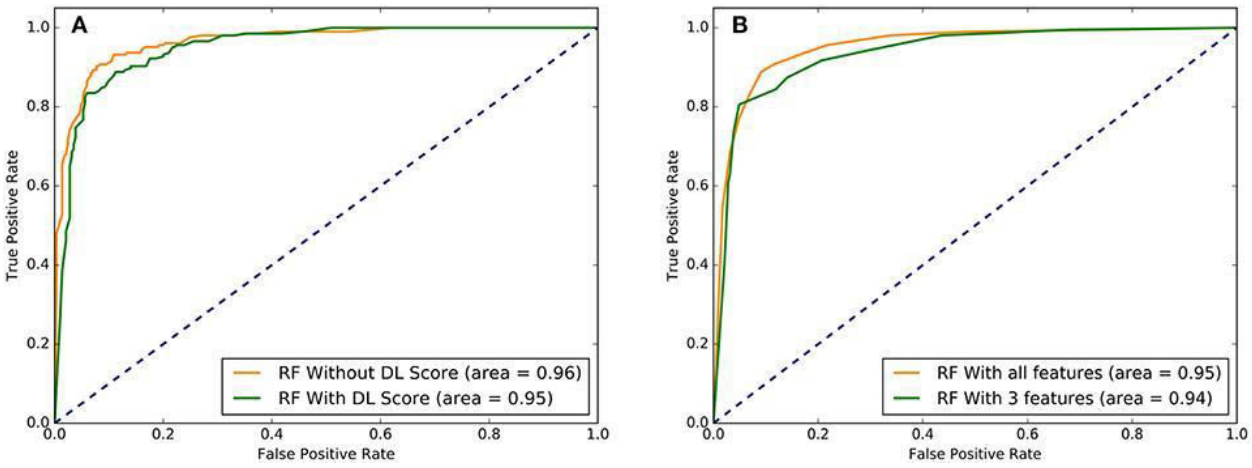


Figure 3-6. The AUC/ROC plots of sensitivity (TPR) versus specificity (TNR). (A) The ROC curves for overall performance of the RF model with 32 functional features excluding DL score (in green) and 33 features that included DL score (in red). (B) The ROC curves for the RF model with all 33 features (in green) and with the top 3 performing features that included LR score, Radial_SVM score, and DL score (in red). Higher AUC score indicates better performance. These plots illustrated a comparative performance of machine learning models for top prediction scores.

First, we evaluated the difference in the AUC curves for RF-based classification with 32 functional features and with additional DL score (Figure 3-6A). The results showed a very similar high-level prediction performance with AUC = 0.95–0.96. It is worth noting that due to high AUC value for RF classification with 32 informative functional features, the addition of DL could not significantly enhance it. However, we showed that this nucleotide-derived predictor score provides an additional information content and is complementary to the ensemble-based RadialSVM score and LR score. In this context, it was instructive to observe that addition of DL score may marginally improve separation between sensitivity and 1 – specificity (FPR) at higher values of these parameters (Figure 3-6A).

Interestingly, RF model that relied on RadialSVM score, LR score, and DL score(top three predictors) yielded 94% AUC , thereby showing that these predictors may be adequate for robust classification of cancer driver mutations on a fairly large dataset of somatic mutations employed in this study. We closely examined the discriminatory ability of the classification models (Table 3-2).

Table 3-2. Statistics and comparative performance metrics of various ML classification of cancer driver mutations models for the top eight predictors.

	Random forest	Boosted trees	Support vector machines
False negative rate	0.91	0.91	0.75
True negative rate	0.11	0.12	0.02
False positive rate	0.12	0.11	0.80
True positive rate	0.86	0.85	0.95
Recall	0.90	0.90	0.89
Precision	0.90	0.90	0.89
F1- measure	0.90	0.90	0.89
Accuracy	0.89	0.89	0.89

All algorithms achieved a high classification accuracy of ~90%. The TPR values were higher for the RF and SVM classifiers, but all algorithms resulted in similar high performance classification on the dataset with only limited number of major predictors that included DL score (Table 3-2). Summarily, our findings supported the notion that ML-derived ensemble functional features may play a cardinal role in classification of driver mutations. The major finding of these ML experiments was that combination of ensemble-based features and DL score are

complementary and when combined can yield comparable classification accuracy. The critical lesson from this analysis is that integrated high-level features obtained by ML strategies from primary nucleotide and protein sequence information may be adequate for predicting key functional phenotypes.

3.5 References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248

Agajanian, S., Odeyemi, O., Bischoff, N., Ratra, S., and Verkhivker, G. M. (2018). Machine learning classification and structure-functional analysis of cancer mutations reveal unique dynamic and network signatures of driver sites in oncogenes and tumor suppressor genes. *J. Chem. Inf. Model*. 58, 2131–2150. doi: 10.1021/acs.jcim.8b00414

Ainscough, B. J., Barnell, E. K., Ronning, P., Campbell, K. M., Wagner, A. H., Fehniger, T. A., et al. (2018). A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nat. Genet*. 50, 1735–1743. doi: 10.1038/s41588-018-0257-y

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell* 173, 371–385.e318. doi: 10.1016/j.cell.2018.02.060

Bertrand, D., Drissler, S., Chia, B. K., Koh, J. Y., Li, C., Suphavilai, C., et al. (2018). Consensus driver improves upon individual algorithms for predicting driver alterations in different cancer types and individual patients. *Cancer Res*. 78, 290–301. doi: 10.1158/0008-5472.CAN-17-1345

Biau, G. (2012). Analysis of a random forest model. *J. Mach. Learn. Res.* 13, 1063–1095. Available online at: http://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., et al. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res*. 69, 6660–6667. doi: 10.1158/0008-5472.CAN-09-1133

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2, 401–404. doi: 10.1158/2159-8290.CD-12-0095

Cheng, F., Zhao, J., and Zhao, Z. (2016). Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Brief. Bioinformatics* 17, 642–656. doi: 10.1093/bib/bbv068

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* 7:e46688. doi: 10.1371/journal.pone.0046688

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol*. 6:e1001025. doi: 10.1371/journal.pcbi.1001025

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. *Genome Res*. 22, 1589–1598. doi: 10.1101/gr.134635.111

Ding, L., Wendl, M. C., McMichael, J. F., and Raphael, B. J. (2014). Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet*. 15, 556–570. doi: 10.1038/nrg3767

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., et al. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet*. 24, 2125–2137. doi: 10.1093/hmg/ddu733

Douville, C., Carter, H., Kim, R., Niknafs, N., Diekhans, M., Stenson, P. D., et al. (2013). CRAVAT: cancer-related analysis of variants toolkit. *Bioinformatics* 29, 647–648. doi: 10.1093/bioinformatics/btt017

Ellrott, K., Bailey, M. H., Saksena, G., Covington, K. R., Kandoth, C., Stewart, C., et al. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst*. 6, 271–281.e277. doi: 10.1016/j.cels.2018.03.002

Erickson, B. J., Korfiatis, P., Akkus, Z., Kline, T., and Philbrick, K. (2017). Toolkits and libraries for deep learning. *J. Digit Imag.* 30, 400–405. doi: 10.1007/s10278-017-9965-6

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*. 43, D805–811. doi: 10.1093/nar/gku1075

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. (2004). A census of human cancer genes. *Nat. Rev. Cancer* 4, 177–183. doi: 10.1038/nrc1299

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal*. 6:pl1. doi: 10.1126/scisignal.2004088

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* 25, i54–62. doi: 10.1093/bioinformatics/btp190

Gnad, F., Baucom, A., Mukhyala, K., Manning, G., and Zhang, Z. (2013). Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* 14 (Suppl 3):S7. doi: 10.1186/1471-2164-14-S8-S7

Goh, G. B., Hodas, N. O., and Vishnu, A. (2017). Deep learning for computational chemistry. *J. Comput. Chem.* 38, 1291–1307. doi: 10.1002/jcc.24764

Gonzalez-Perez, A., and Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Res*. 40:e169. doi: 10.1093/nar/gks743

Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R., Creixell, P., Karchin, R., et al. (2013). Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods* 10, 723–729. doi: 10.1038/nmeth.2562

Haber, D. A., and Settleman, J. (2007). Cancer: drivers and passengers *Nature* 446, 145–146. doi: 10.1038/446145a

Hinkson, I. V., Davidsen, T. M., Klemm, J. D., Kerlavage, A. R., and Kibbe, W. A. (2017). A comprehensive infrastructure for big data in cancer research: accelerating cancer research and precision medicine. *Front. Cell. Dev. Biol*. 5:83. doi: 10.3389/fcell.2017.00083

Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabe, R. R., et al. (2010). International network of cancer genome projects. *Nature* 464, 993–998. doi: 10.1038/nature08987

Jensen, M. A., Ferretti, V., Grossman, R. L., and Staudt, L. M. (2017). The NCI Genomic Data Commons as an engine for precision medicine. *Blood* 130, 453–459. doi: 10.1182/blood-2017-03-735654

Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Kallberg, M., et al. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* 15, 591–594. doi: 10.1038/s41592-018-0051-x

Klonowska, K., Czubak, K., Wojciechowska, M., Handschuh, L., Zmienko, A., Figlerowicz, M., et al. (2016). Oncogenomic portals for the visualization and analysis of genome-wide cancer data. *Oncotarget* 7, 176–192. doi: 10.18632/oncotarget.6128

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 22, 568–576. doi: 10.1101/gr.129684.111

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., et al. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* 28, 311–317. doi: 10.1093/bioinformatics/btr665

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218. doi: 10.1038/nature12213

Li, J., Drubay, D., Michiels, S., and Gautheret, D. (2015). Mining the coding and non-coding genome for cancer drivers. *Cancer Lett*. 369, 307–315. doi: 10.1016/j.canlet.2015.09.015

Liu, X., Jian, X., and Boerwinkle, E. (2011). dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat*. 32, 894–899. doi: 10.1002/humu.21517

Liu, X., Jian, X., and Boerwinkle, E. (2013). dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat*. 34, E2393–2402. doi: 10.1002/humu.22376

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat*. 37, 235–241. doi: 10.1002/humu.22932

Luo, P., Ding, Y., Lei, X., and Wu, F. X. (2019). deepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Front. Genet*. 10:13. doi: 10.3389/fgene.2019.00013

Mao, Y., Chen, H., Liang, H., Meric-Bernstam, F., Mills, G. B., and Chen, K. (2013). CanDrA: cancer-specific driver missense mutation annotation with optimized features. *PLoS ONE* 8:e77945. doi: 10.1371/journal.pone.0077945

Martelotto, L. G., Ng, C. K., De Filippo, M. R., Zhang, Y., Piscuoglio, S., Lim, R. S., et al. (2014). Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biol*. 15:484. doi: 10.1186/s13059-014-0484-1

Masica, D. L., Douville, C., Tokheim, C., Bhattacharya, R., Kim, R., Moad, K., et al. (2017). CRAVAT 4: cancer-related analysis of variants toolkit. *Cancer Res*. 77, e35–e38. doi: 10.1158/0008-5472.CAN-17-0338

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). *Efficient Estimations of Word Representations in Vector Space. arXiv:1301.3781 [cs.CL]*. Available online at: https://arxiv.org/abs/1301.3781

Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A., and Lopez-Bigas, N. (2016). OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol*. 17:128. doi: 10.1186/s13059-016-0994-0

Parthiban, V., Gromiha, M. M., Abhinandan, M., and Schomburg, D. (2007). Computational modeling of protein mutant stability: analysis and optimization of statistical potentials and structural features reveal insights into prediction model development. *BMC Struct. Biol*. 7:54. doi: 10.1186/1472-6807-7-54

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Piraino, S. W., and Furney, S. J. (2016). Beyond the exome: the role of non-coding somatic mutations in cancer. *Ann. Oncol*. 27, 240–248. doi: 10.1093/annonc/mdv561

Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol*. 36, 983–987. doi: 10.1038/nbt.4235

Poulos, R. C., and Wong, J. W. H. (2018). Finding cancer driver mutations in the era of big data research. *Biophys. Rev*. 11, 21–29. doi: 10.1007/s12551-018-0415-6

Raphael, B. J., Dobson, J. R., Oesper, L., and Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med*. 6:5. doi: 10.1186/gm524

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res*. 39:e118. doi: 10.1093/nar/gkr407

Schwarz, J. M., Rodelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* 7, 575–576. doi: 10.1038/nmeth0810-575

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., et al. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat*. 34, 57–65. doi: 10.1002/humu.22225

Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*. 40, W452–457. doi: 10.1093/nar/gks539

Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013). OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* 29, 2238–2244. doi: 10.1093/bioinformatics/btt395

Tokheim, C., Bhattacharya, R., Niknafs, N., Gygax, D. M., Kim, R., Ryan, M., et al. (2016). Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res*. 76, 3719–3731. doi: 10.1158/0008-5472.CAN-15-3190

Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., and Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci. U.S.A*. 113, 14330–14335. doi: 10.1073/pnas.1616440113

Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., et al. (2017). The UCSC genome browser database: 2017 update. *Nucleic Acids Res.* 45, D626–D634. doi: 10.1093/nar/gkw1134

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet*. 45, 1113–1120. doi: 10.1038/ng.2764

Wu, J., Wu, M., Li, L., Liu, Z., Zeng, W., and Jiang, R. (2016). dbWGFP: a database and web server of human whole-genome single nucleotide variants and their functional predictions. *Database* 2016:baw024. doi: 10.1093/database/baw024

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., et al. (2011). International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. *Database* 2011: bar026. doi: 10.1093/database/bar026

4 Machine Learning Based Classification of Survival and Cox PH Model Analysis of Selected Unresectable Cancers

Oluyemi Oluyemi, Kristalee Lio, Cyril Rakovski

Author's Contribution

YO and CR conceived and designed the research. YO performed the research. YO, KL and CR analyzed the results and wrote the manuscript. YO and CR wrote the final version of the manuscript and supervised the project.

4.1 Abstract

Background: Acute lymphoblastic leukemia, colorectal adenocarcinoma, glioblastoma, and neuroblastoma are well studied unresectable cancer types. Ensemble machine learning classifiers can be used for the analysis of the overall survival status of selected cancer types. The primary objectives of this study were to build: (1) high accuracy machine learning model to classify cancer survival status and (2) Cox proportional-hazards survival analysis for pediatric cancers.

Methods: Data for the four selected cancer types were retrieved from the cBioPortal cancer genome database. The high-performance machine learning approaches, gradient boost machine (GBM), extreme gradient boosting (XGBoost) and a deep neural network architecture were used to and trained on the cancer survival datasets. Five survival status models were built for each of the analyzed cancer types and a fifth one was built for the Memorial Sloan Kettering cancer center-MSK-IMPACT dataset. Furthermore, Cox proportional- hazards models were implemented to conduct survival analysis on pediatric acute lymphoblastic leukemia and pediatric neuroblastoma that allow identification and effect size estimation of the statistically significant predictors of these cancers.

Results: The ensemble tree classifiers and the deep neural network model obtained area under the curves values between 0.75 and 0.95 across all five models. The Cox-proportional- hazards model showed that age at diagnosis, tumor origin of T-cells(effect size/HR: 0.35) , tumor origin of B-cells(HR:0.39), congenital abnormality-no(HR: 0.59),congenital abnormality-yes(HR: 8.30), molecular subtype-hyperdiploidy without trisomy of both chromosomes 4 and 10(HR: 0.70) and molecular subtype-iAMP21(HR: 1.67) , TCF3_PBX1_status-negative(HR: 0.76)and Trisomy 4 and 10 -negative(HR: 1.52) play a significant role in survival rate of pediatric acute lymphoblastic leukemia. Whereas, in the pediatric neuroblastoma INSS-stages 1(HR: 0.44), INSS-stage 2a(HR: 0.53),INSS-stage 4(HR: 0.51), INSS stage 4s(HR: 2.69), the number of complete sets of cellular chromosomes -ploidy_diploid/DI =1 (HR:0.69) and the number of complete sets of cellular chromosomes-DI>1(HR: 2.39) and gender- male (HR: 0.89) were the statistically significant features in the Cox PH model.

Conclusion: Machine learning can be used to predict the overall survival status of cancer using age at diagnosis, mutational profile among other risk factors. Age at diagnosis, cell of tumor origin, congenital abnormality, molecular subtype, INSS stage, and ploidy make a significant impact on the survival rate in pediatric cancers.

4.2 Introduction

The burden of cancer in the US is on the rise, there is an incessant need to develop efficient strategies for prevention and standard care (Allemani et al., 2015; Cavalli & Cavalli, 2013). In 2018, cancer accounted for more than 600,000 deaths in the US (Miller et al.,2018; NCBI cancer statistics,2018). However, the US has better survival statistics when compared to most other developed countries (Allemani et al.,2015; Centers for Disease Control and Prevention, 2019; Gorey, 2009; Coleman et al., 2008). The National Institute of Health (NIH)

introduced a couple of programs with a special focus on the early diagnosis as a key approach to improving cancer survival rates (Cronin et al., 2018). Survival of cancers is dependent on numerous factors namely: (1) stage of cancer (2) cancer type; (3) prior treatment strategy;(3) level of fitness. In a study conducted by Lim et al. (2006), they found evidence of cancer survival dependence on seasonality of diagnosis and sunlight exposure. Survival statistics are often used for prognosis estimation and evaluation of treatment options (Cancer.net, 2018)

Surgical resection remains the main curative treatment approach for most cancers. However, more than 60% of patients have unresectable cancer at the time of diagnoses such as extra-regional lymph node metastasis and local invasion. Despite recent innovations in chemotherapy strategy, the median survival for patients with skin cancer was approximately 113 months, for aerodigestive cancers was 70 months, and 107 months for patients with salivary gland cancers. The survival rates for gastric cancer are among the worst of any cancer with a 3.1% five-year patient survival (Carey et al., 2019; Yang et al., 2011 and Brenner et al., 2009).

Data from National Cancer Intelligence Network (NCIN) showed that 70% of acute lymphoblastic leukemia (ALL) patients will survive leukemia for five years or more after diagnosis (Hoezler et al., 2016; National Cancer Intelligence Network, 2014; Tobias and Hochhauser 2015; Wang et al.,2019).

Unresectable cancers are cancers that cannot be removed completely by surgery. Cancer can be classified as unresectable for several reasons : (1) tumor location-a tumor may be interconnected with vital blood vessel; (2) tumor size-too large to remove; (3) metastases-cancer cells that have spread to other parts; (4) other medical condition that could increase the risk of surgery-severe diabetes raises the risk of surgery to critical levels (Fabre et al., 2015; Imai et al., 2016; Kato et al., 2015; Nitsche et al.,2015 and Yang et al., 2015)

ALL is the second most common acute leukemia in the US and the American Cancer Society ACS estimated that there will be ~ 6000 new cases of ALL and ~ 1500 deaths from ALL in 2019 (Howlader et al., 2018; Terwilliger and Abdul-Hay, 2017). ALL is a malignancy of hematopoietic stem cells that originates from B and T -cells. ALL pathogenesis involves differentiation and abnormal growth of the lymphoid cells' clonal population. Recent pediatric studies have identified genetic syndromes that are predisposing to minor cases of ALL namely, Bloom, Down and Nijmegen breakdown syndromes (Terwilliger and Abdul-Hay, 2017; Shah et al., 2013; Bieloral et al.,2013; Chessells et al., 2001 and German 1997). Other factors include exposure to Human Immunodeficiency Virus (HIV) (Geriniere et al., 1994), Epstein-Barr Virus (EBV)(Seghal et al., 2010), pesticide and ionizing radiation(Spector et al.,2006). However, in most cases, ALL appears as a new malignancy in previously healthy patients. ALL is often driven by mutations, gene aneuploidy, and chromosomal translocation (Mohseni et al., 2018; Hunger et al., 2015). Several genetic subsets of B-cell and T-cell ALL such as BCR-ABL1, MLL rearrangements, Hypodiploid, iAMP21, ETV6-RUNX1, TCF3-HLF have to be found to play a significant role in tumor development. Certain changes in leukemia cells' genes or chromosomes can affect prognosis and children tend to have a less favorable outcome if the leukemia cells have a translocation between chromosomes 9 and 22 (Philadelphia chromosome); hypodiploidy (fewer than 44 chromosomes)and complex karyotype while on the other hand patients tend to have a better outlook if the leukemia cells have hyperdiploidy (more than 50 chromosomes) and a translocation between chromosome 12 and 21. Recent studies have shown that the risk for developing ALL is highest in children younger than five years of age and that the risk declines until the age of twenty-five and begins to rise again after fifty years of age. NCI SEER data indicated that the median age at diagnosis is 16 years and the median age at death stood at 56

years. The risk factors for ALL are radiation exposure, certain chemical exposure – benzene, certain viral infections -HTLV1, certain genetic syndromes-Down, and Klinefelter syndromes, age, race/ethnicity among other factors (Jain et al., 2016). Despite the high response rate to chemotherapy, only approximately 35% of adult patients with acute lymphoblastic leukemia attain long-term remission (Terwilliger and Abdul-Hay, 2017; Jabbour et al.,2015). The survival of young patients with ALL has considerably improved in the 21st century, however, there is insufficient data to determine if patients of all ethnic backgrounds have benefited equally (Pulte et al., 2013)

Similarly, colorectal cancer the third most commonly diagnosed cancer in the US is estimated to record over 50,000 deaths in 2019 (American Cancer Society,2020; Howlader et al., 2016). The treatment for colorectal cancer has evolved in the last decade. FOLFOX and other chemotherapy regimens remain the gold standard for therapy, however, the addition of known epidermal growth factor receptor (EGFR) inhibitors have improved colorectal cancer outcomes (Miyamoto et al.,2017). Fortunately, colorectal cancer has a 90% five-year survival rate if detected early (American Cancer Society,2020). Gliomas, another lethal cancer are the most common cranial tumors in adults representing 80% of malignant brain tumors. They originate from glial tissue. Glioblastoma, the most prevalent glioma histology (~45% of all gliomas) has 37.2%,5.1% and 2.6% survival rates for one-, five- and ten-year periods from diagnosis (Ostrom et al.,2015). The minority of these tumors are caused by tuberous sclerosis and neurofibromatosis and other Mendelian disorders. Recent studies indicate that isocitrate dehydrogenase mutation, cytosine–phosphate–guanine island methylator phenotype and O^6-methylguanine-DNA methyltransferase methylation and other biomarkers significantly improved survival in glioma (Ostrom et al., 2015). The Cancer Genome Atlas Research Network glioblastoma studies

revealed recurrent alterations in the following pathways: (i) tumor protein 53 (TP53) signaling (CDKN2A deletion, TP53 mutation, mouse double minute 1/4 [MDM1/MDM4] amplification); (ii) retinoblastoma (RB) signaling (cyclin-dependent kinase inhibitor 2A/2C [CDKN2A/CDKN2C] deletion, RB mutation, cluster of differentiation 4/6 [CD4/CD6] amplification); and (iii) receptor tyrosine kinase signaling (phosphatase and tensin homolog/neurofibromin 1/phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha [PTEN/NF1/PIK3CA] mutation, epidermal growth factor receptor/platelet-derived growth factor receptor [EGFR/PDGFRA] amplification)(Cancer Genome Atlas Research Network, 2008).Glioblastomas are often genetically diagnosed as IDH-wildtype, IDH-mutant or Glioblastoma with unknown IDH status. Noushmehr et al. (2010) and Mur et al. (2013) studies found a substantial link between a genome-wide glioma cytosine–phosphate–guanine island methylator phenotype (G-CIMP) and IDH mutations in all glioma subgroups.

Neuroblastoma is the most common extracranial solid tumor in infants and originates from the neural crest element of the sympathetic nervous system (Stewart et al., 2016; Cheung and Dyer, 2013). It is a rare type of cancerous tumor and is the most common cancer in children accounting for 6% of all cancers in infants. 90% of cases are diagnosed at age five (Stewart et al., 2016). It occurs slightly more in males than females. Some of the risk factors associated with neuroblastoma are age, heredity, and congenital anomalies. Biopsy, blood, and urine catecholamine tests, imaging tests (x-rays, MRI, MIBG, CT scans), neurological exam and bone marrow aspiration are among several tests and procedures done for the diagnosis of neuroblastoma. Neuroblastoma's spread description is known as staging. Staging helps clinicians to determine the appropriate treatment option for a patient. Upon the diagnosis of neuroblastoma, the tumor spread is assessed using the international neuroblastoma staging system committee

(INSS) system. The INSS consists of six stages and are summarized as follows: (1) stage 1- tumor can be excised completely by surgery; (2) stage 2A- the tumor is localized and cannot be completely removed during surgery and surrounding lymph are not cancerous; (3) stage 2B- the tumor is localized and may or may not be completely removed during surgery and the surrounding lymph nodes are cancerous; (4) stage 3- the tumor cannot be excised with surgery, often cancer has spread to surrounding lymph nodes; (5) stage 4- the tumor has metastasized to distant lymph nodes, bones, liver, skin and other organs; and (6) stage 4S- the tumor is localized (as in 1 and 2) and it has metastasized to the skin, liver and none in infants (Monclair et al., 2009). INSS classification systems in combination with age at diagnosis, MYCN gene status, chromosome 11q status, tumor cell ploidy, differentiated cell grade are used to determine the clinical risk of a tumor and its response to therapy. A patient's age and the INSS stage plays a key role in the treatment approach for the disease. The treatment regimen for neuroblastoma consists of three approaches namely surgery, radiation therapy, iodine 131-MIBG therapy (treatment with radioactive iodine), chemotherapy and a combination of high-dose chemotherapy and radiation therapy with stem cell transplants

In this study, we built a machine learning tree-based ensemble classification model for the survival status of patients with selected unresectable cancers: acute lymphoblastic leukemia, colorectal adenocarcinoma, glioblastoma, and neuroblastoma. Likewise, we implemented a separate Cox-proportional hazard survival analysis for two unresectable pediatric cancers - acute lymphoblastic leukemia and neuroblastoma.

4.3 Methods

4.3.1 Survival Status Dataset

The dataset used in the study was retrieved from cBioPortal (www.cbioportal.org). The dataset consists of clinical and genomic features such as age at diagnosis, ploidy, molecular subtypes. In this study, we are primarily interested in ensemble classification of cancer survival status of the following cancer sets namely pediatric acute lymphoblastic leukemia(ALL), glioblastoma (GLIO), colorectal adenocarcinoma (COAD), pediatric neuroblastoma (P-NB) and MSK-IMPACT clinical sequencing cohort - targeted tumor-sequencing of 10000 clinical cases. Secondarily, we are interested in the survival analysis of ALL and P-NB due to the low median age at diagnosis of pediatric acute lymphoblastic leukemia and pediatric neuroblastoma. Table 4-1 shows the total number of patients and samples used in this study.

Table 4-1. Total number of patients of selected cancer types and MSK_IMPACT combined cancer type

Profiling Sample (TARGET)	Total Number of Patients
Acute Lymphoid Leukemia (Pediatric)	1551
Glioblastoma	585
Colorectal Adenocarcinoma	594
Neuroblastoma (Pediatric)	1076
MSK_IMPACT cohort*	10336

*includes other cancer types

Survival Status Data Preparation and Feature Selection

We carried out data management and exploratory analysis steps. In particular, we implemented feature rescaling, missing value imputation and subsequent stepwise feature selection. We assessed the pairwise correlations between the predictors as orthogonal covariates are advantageous for building machine certain learning models. We also studied issues with multicollinearity as they have the potential of causing overfitting in machine learning models (Vatcheva et al., 2016; Atems and Bergtold, 2015, Zhang et al., 2018; Kroll and Song 2013)

Missing values imputation was handled by using the nearest neighbor and median values approach depending on the data kind (categorical or numerical features). Other discrepancies were handled using outlier detection techniques and min-max feature rescaling for data normalization while regular expression and stepwise Akaike information criterion (stepAIC) were used for feature extraction and feature selection respectively (Li et al., 2016; Zhang, 2016; Paja et al., 2017). Since the main objective of this study is to build a predictive model for survival status of acute lymphoblastic leukemia, glioblastoma, colorectal adenocarcinoma, and neuroblastoma cancer, we carried out a quantitative and qualitative comparative analysis of the overall survival status, sex, ethnicity/race, age at diagnosis, mutation count and survival event.

4.3.2 Proposed Models

4.3.2.1 Ensemble Classifiers

In this study, we used a comparative ensemble algorithms approach - extreme gradient boosting (xgboost) and gradient boosting machine(gbm) for survival status classification as described by Zheng et al, 2020; Chen and Guestrin,2016; Friedman,2001. We compared the performance of the classifiers with that of deep neural networks as described by Telenti et al.,

2018; Kalinin et al., 2018; Angermueller et al.2016. We adopted the D-CNN architecture used in the Telenti et al., 2019 for this comparative study. Also, we investigated variable importance to assess the most informative features in the survival ensemble models. To prevent overfitting of the classifiers and account for the skewness in the distribution of the overall survival status-deceased vs living, we oversampled the survival status using the synthetic minority over-sampling technique -SMOTE (Zheng et al.,2019). SMOTE theory assumes that the feature space of minority class instances is similar. We used F1-score, AUC, precision and recall metrics as described by for our model evaluation as described by Zheng et al., 2020.

4.3.2.2 Cox Proportional Hazard Model

In the survival analysis of the pediatric neuroblastoma and glioblastoma, we used the cox proportional hazards model as described by Wang and Albert, 2016; Meguid et al., 2010; Xintong Chen, 2014 for semiparametric modeling the hazards ratio of the covariates.

4.4 Results and Discussion

Our initial exploratory analysis showed that MSK-IMPACT, neuroblastoma (P-NB), glioblastoma (GLIO), colorectal adenocarcinoma (COAD) and acute lymphoid leukemia had 70.77 %, 62.72 %, 29.23%, 79.8% and 73.46% for the survival status distribution (Figure 4-1).

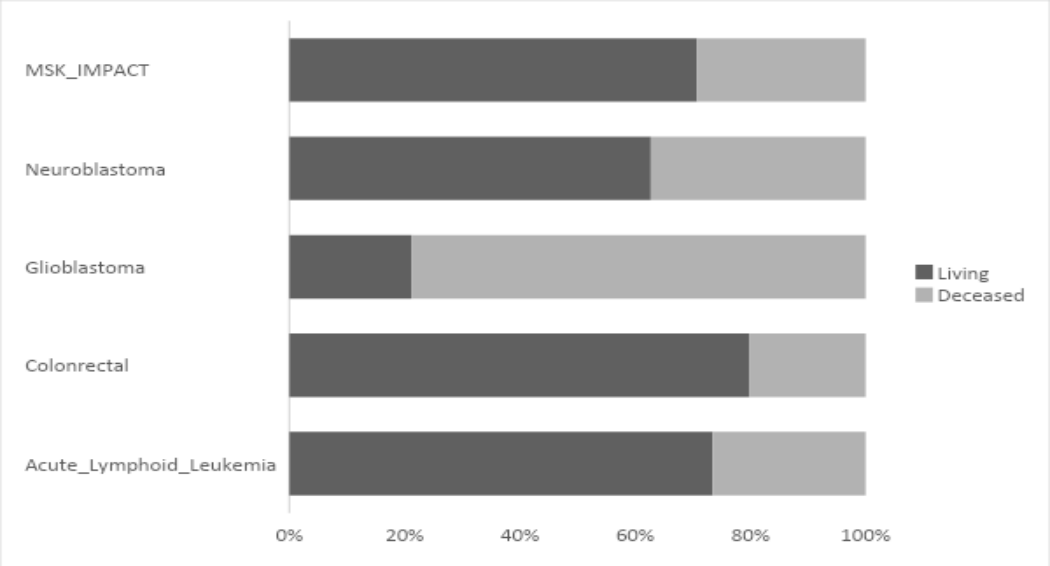


Figure 4-1. Bar plot showing the percentage bar graph of survival status (living versus deceased) of acute lymphoid leukemia (ALL), colorectal adenocarcinoma (COAD), glioblastoma (GLIO), neuroblastoma (P-NB) and MSK-IMPACT cancer studies

In the EDA, all 4-cancer types have a higher survival rate except for glioblastoma, this can be attributed to tumor location - glioblastoma is found in the brain's cerebral hemisphere which poses a challenge for chemotherapy and radiation. Age as a risk factor can also be attributed to the high mortality rate among glioblastoma patients, the average age at diagnosis of glioblastoma is sixty-four years and risk increases with an increase in age. Patients with glioblastoma face major mortality, with approximately 13000 deaths per year in the US. Also, of all 4 cancer types, colorectal had the highest survival rate of 79.8%. This aligns with NCI statistics which puts the 5-year survival rate of colorectal cancer at 71% (SEER stage: regional) and 90% (SEER stage: localized). Figure 2 shows the age at diagnosis of all four cancer types.

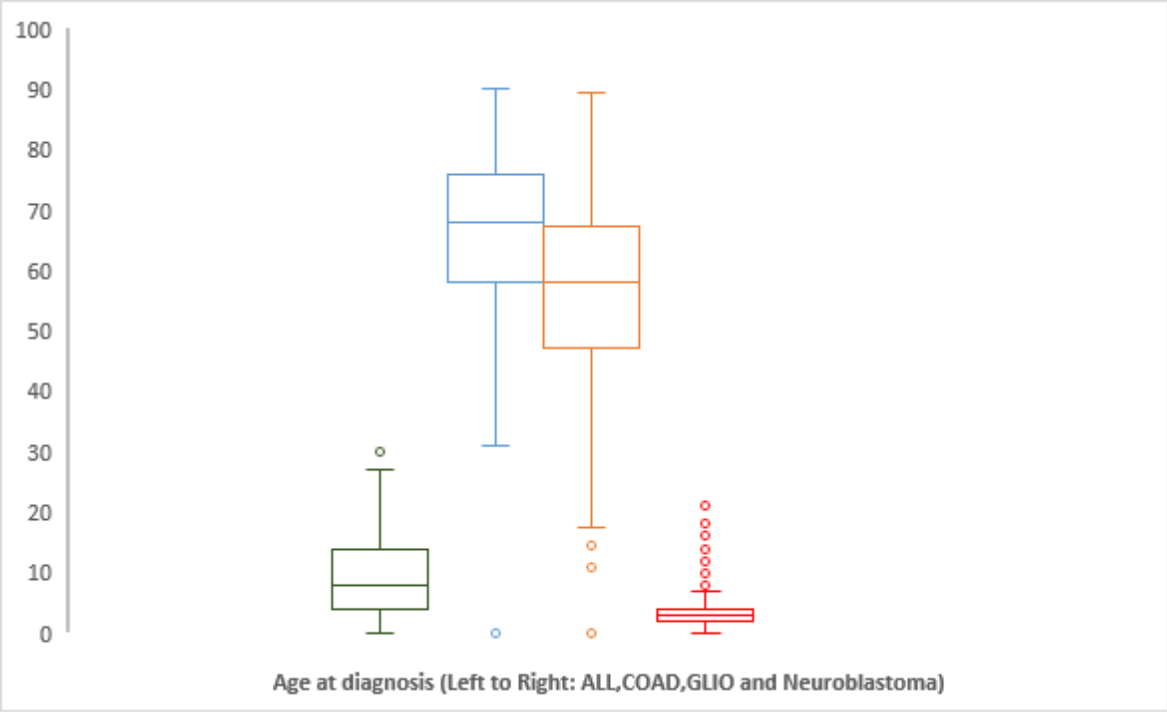


Figure 4-2. Summary statistics of age at diagnosis (left to right): ALL, COAD, GLIO and P-NB

The multicollinearity dependence test of the features indicated that most of the features have pairwise orthogonality except for sex (male vs female), tumor disease anatomic site (rectum vs colon) and tumor type (colon vs rectal) with -0.99, -0.97, -0.78 and -0.91 correlation coefficients. From the correlation matrix, diagnosis age, aneuploidy score, fraction genome altered, mutation count, overall survival in months, race and other features are good for the model (Figure 4-3)

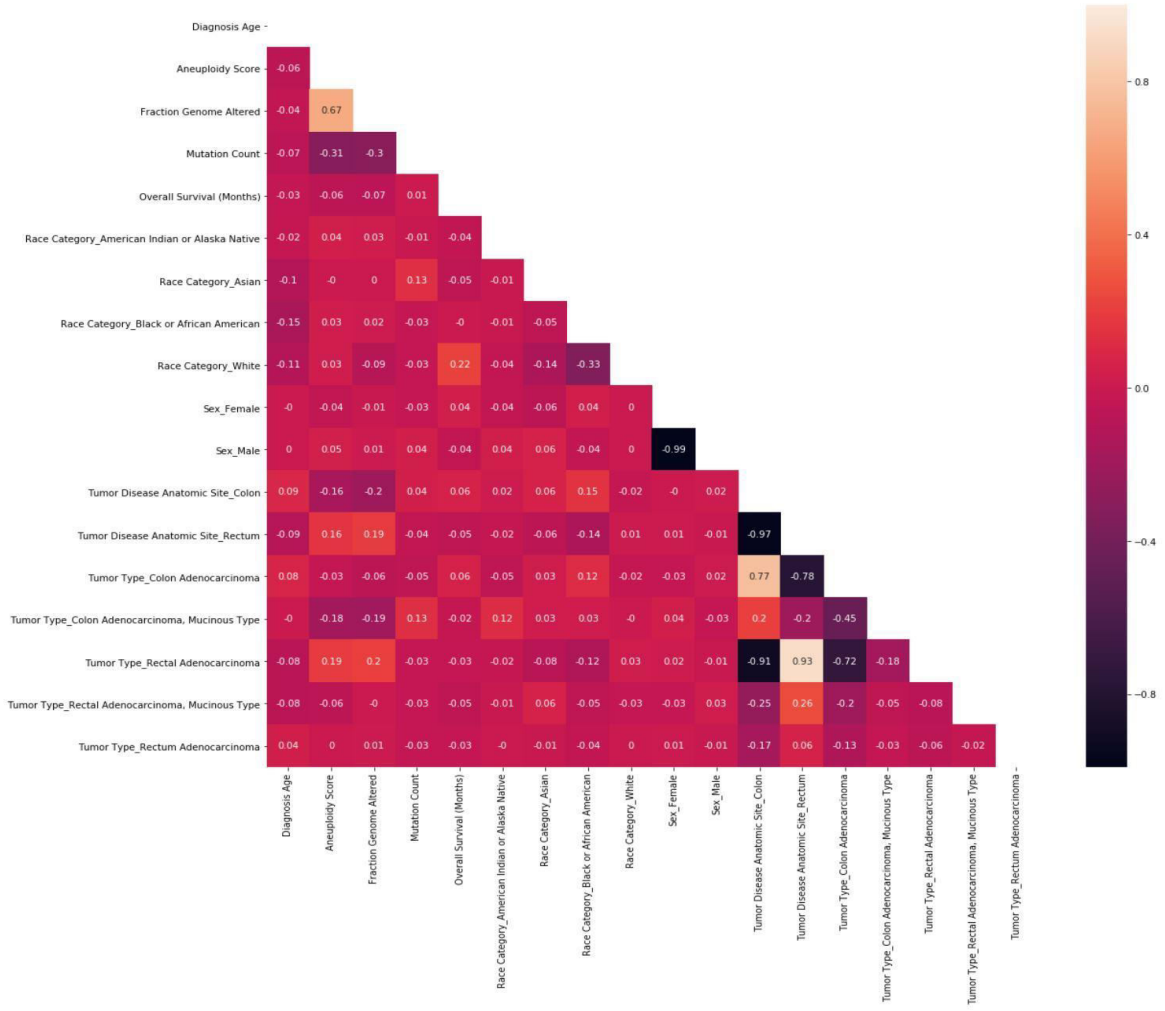


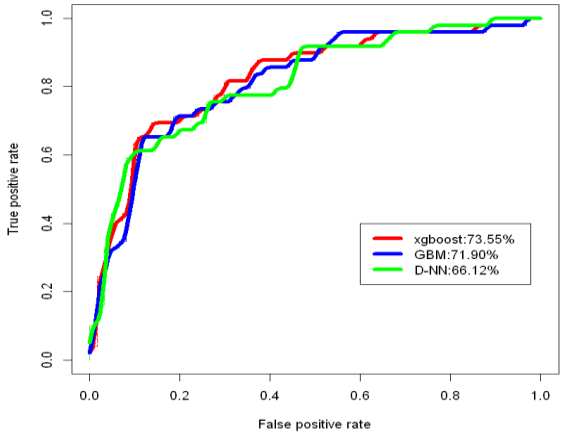
Figure 4-3. Multicollinearity dependence test for some of the clinical (sex, age, race, tumor type, overall survival in months) and genomic (fraction genome altered, aneuploidy score, mutation count) predictors

4.4.1 Comparative Analysis of Ensemble Classification of Cancer Survival Status

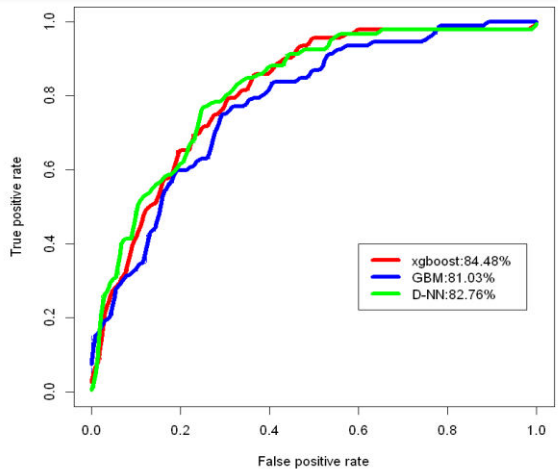
The classifiers performed fairly well at classifying the survival status of ALL, GLIO, COAD, P-NB and MSK- IMPACT models. The performance of the three classifiers (xgboost, GBM and D-NN) employed in the survival status model was examined using metrics including sensitivity(recall), specificity, precision, AUC and f-score (Table 4-2 and Figure 4.4A-E).

The xgboost classifier performed best at classifying the survival status of the cancer types and MSK-IMPACT except for the pediatric neuroblastoma model which recorded AUC and F-score of 0.9447 and0.9268. In the ALL model, the xgboost classifier performed best with a sensitivity score of 0.962, a precision score of 0.856, AUC score of 0.947 and an F score of 0.9058. The recall value tells us that of all the ALL patients that survived, 96.12% of them were correctly labeled as survived while the precision score tells us that of all the ALL patients that were labeled as survivors, 85.59% of them survived. In the GLIO model, the recall and precision values tell us that of all the diagnosed patients with glioblastoma that survived, 94.57% of them were aptly labeled as survivors while of the diagnosed GLIO patients that were labeled as survivors, 87% of them survived. Likewise, our COAD model indicated that of the diagnosed patients that were labeled as survivors, 79.17% were correctly labeled as survivors. Of the colorectal adenocarcinoma diagnosed patients that survived, 95% were correctly labeled.

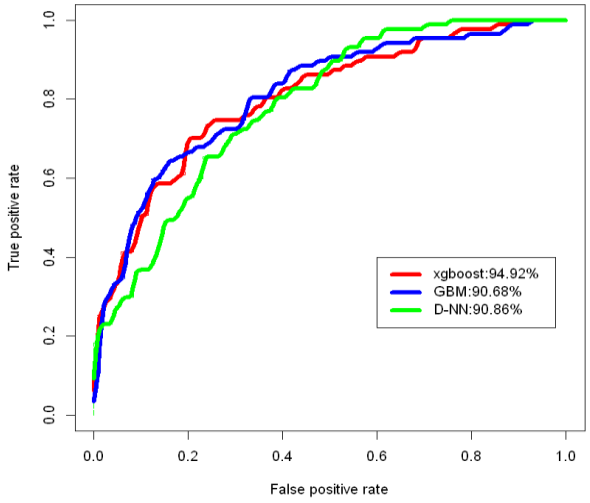
In the pediatric neuroblastoma model, the GBM classifier outperformed both xgboost and the D-NN classifiers – of the neuroblastoma survivors, 93.83% of them were correctly labeled as survivors whereas of the neuroblastoma patients that were labeled as survivors, 91.57% survived.



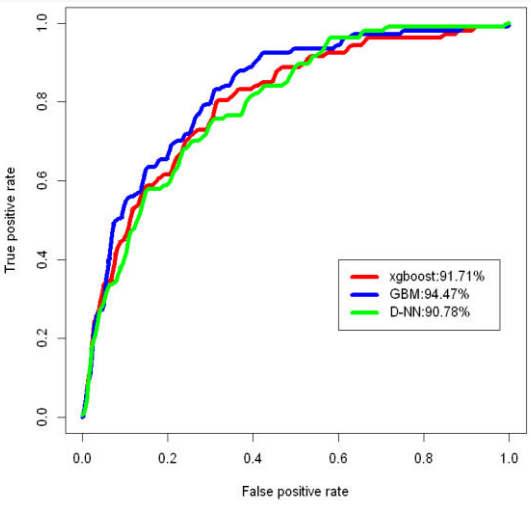
A


True positive rate
False positive rate
xgboost:84.48%
GBM:81.03%
D-NN:82.76%

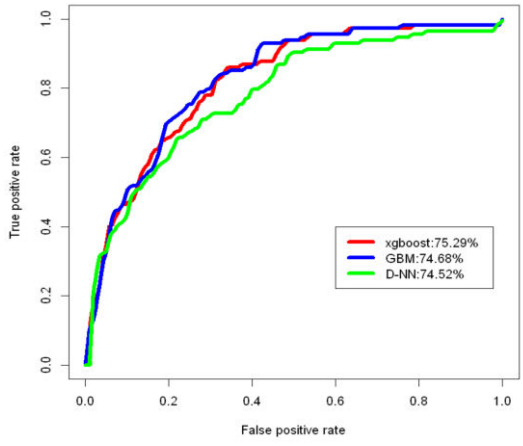
B


True positive rate
False positive rate
xgboost:94.92%
GBM:90.68%
D-NN:90.86%

C



D



E

Figure 4-4. ROC/AUC model evaluation for survival status classification in ALL, GLIO, COAD, P-NB and MSK-IMPACT data sets. The ROC plot of three classifiers – xgboost, GBM and D-NN for ALL. **A**: In the ALL models, xgboost classifier performed best at discriminating survival status (living v deceased) with an AUC score of 0.7355 whereas, the GBM with AUC 0.719 performed moderately well when compared with the D-NN with 0.6612 AUC score. **B**: In the GLIO model, the xgboost was the best classifier with AUC score of 0.8448.**C**: Xgboost classifier (0.9492) performed better at classifying survival status of COAD when compared with D-NN (AUC:0.9068) and GBM (AUC: 0.9086) classifiers. **D:** Interestingly, in the P-NB model, GBM (AUC: 0.9447) outperformed xgboost (AUC: 0.9171) and D-NN (AUC: 0.9078) at classifying survival status of pediatric neuroblastoma **E:** MSK-IMPACT survival status – the classifiers xgboost classifier with an AUC score of 0.7529 outperformed GBM and D-NN respectively

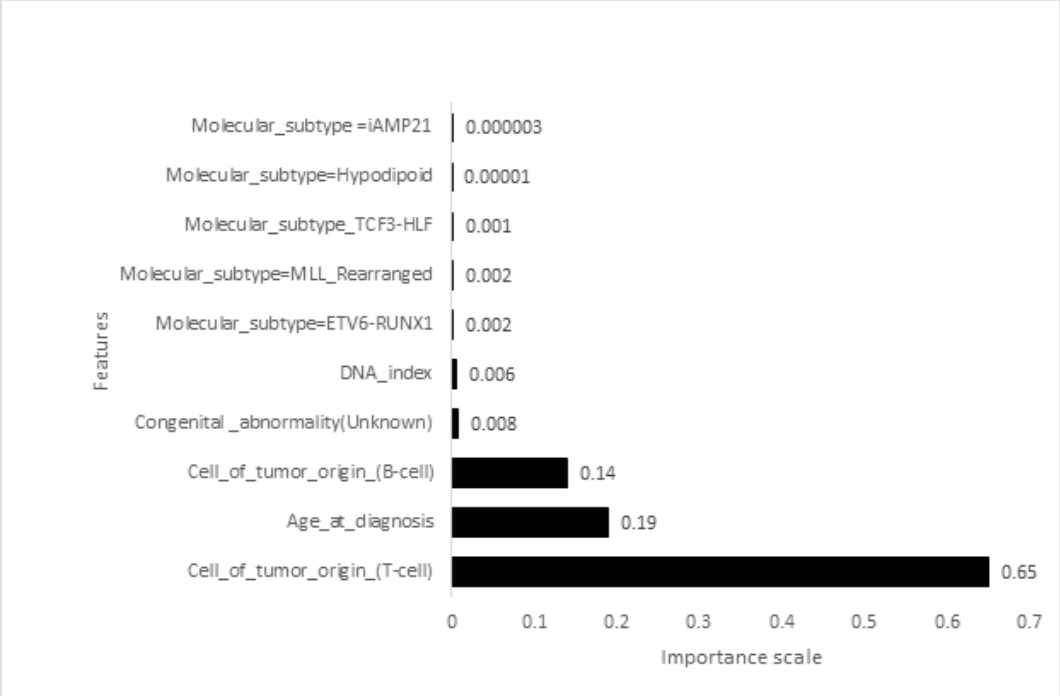
Table 4-2. Model evaluation result of the ensemble classifiers and deep neural network survival status of selected unresectable cancer types

CANCER TYPE	METRIC	CLASSIFIERS		
		D-Neural Network	Xgboost	GBM
ALL	Sensitivity (Recall)	0.9333	0.9619	0.8667
	Specificity	0.9448	0.9414	0.9517
	Precision	0.8596	0.8559	0.8667
	AUC	0.9418	**0.9468**	0.9291
	F-score	0.8949	0.9058	0.8667
GLIO	Sensitivity (Recall)	0.9674	0.9457	0.9348
	Specificity	0.2917	0.4583	0.3333
	Precision	0.8396	0.8700	0.8431
	AUC	0.8276	**0.8448**	0.8103
	F-score	0.8990	0.9063	0.8866
COAD	Sensitivity (Recall)	0.5417	0.7917	0.7083
	Specificity	1.0000	0.9894	0.9574
	Precision	1.0000	**0.9500**	0.8095
	AUC	0.9086	0.9492	0.9068
	F-score	0.7027	0.8637	0.7555
P-NB	Sensitivity (Recall)	0.9012	0.9136	0.9383
	Specificity	0.9118	0.9191	0.9485
	Precision	0.8588	0.8706	0.9157
	AUC	0.9078	0.9171	**0.9447**
	F-score	0.8795	0.8915	0.9268
MSK-IMPACT	Sensitivity	0.3290	0.3246	0.2894
	Specificity	0.9168	0.9295	0.9521
	Precision	0.6298	0.6549	0.6667
	AUC	0.7452	**0.7529**	0.7468
	F-score	0.4322	0.4341	0.4036

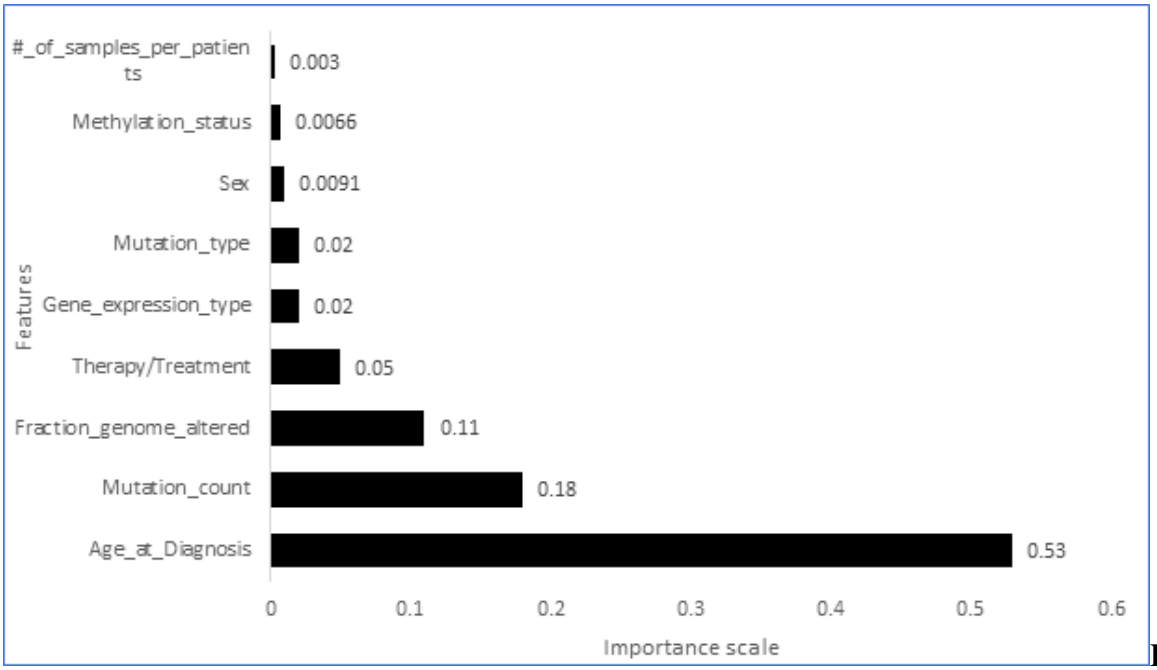
4.4.2 Informative Features and Feature Importance of Survival Status Classification Models

A combination of clinical and genomic features was used for the feature importance of all five models. Due to various cancer types having different number of input features, we narrowed the result of the feature features to the top-ten most informative features contributing to survival status classification of ALL, GLIO, COAD, P-NB and MSK-IMPACT models (Figure 4-5A-E).

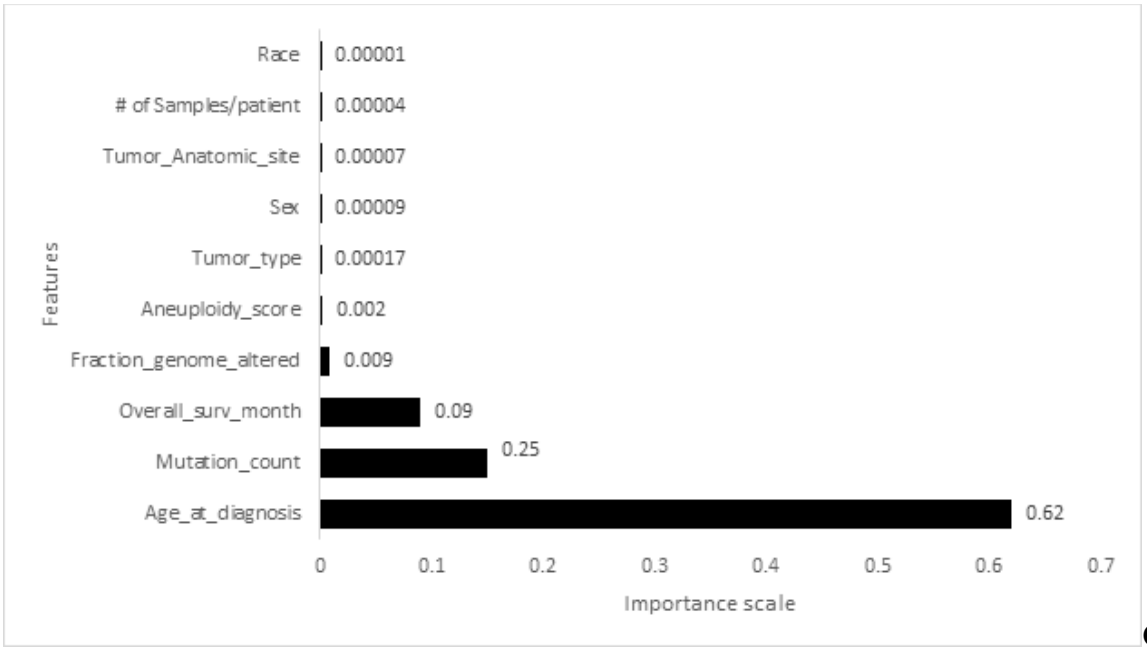
Cell of tumor (T-cell) origin, age at diagnosis, cell of tumor (B-cell) origin, Congenital abnormality, and the DNA Index were the most informative clinical features in the ALL model. Whereas the ETV6-RUNX1, MLL_Rearranged, TCF3-HLF, Hypodiploid and iAMP21 molecular subtypes were the most informative genomic predictors.



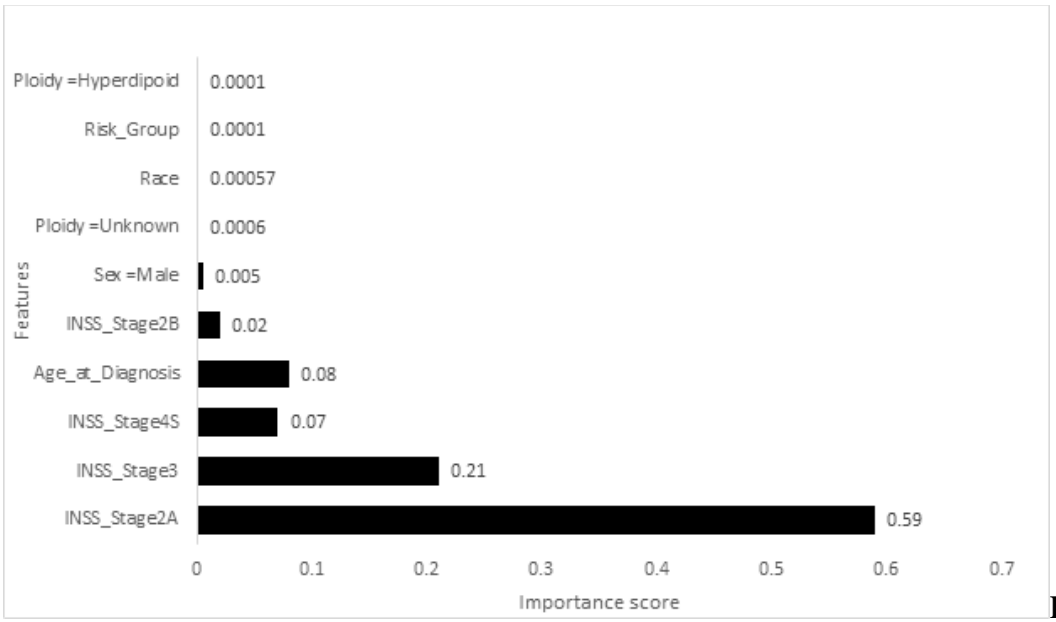
A


#_of_samples_per_patients
0.003
Methylation_status
0.0066
Sex
0.0091
Mutation_type
0.02
Gene_expression_type
0.02
Therapy/Treatment
0.05
Fraction_genome_altered
0.11
Mutation_count
0.18
Age_at_Diagnosis
0.53
Features
0
0.1
0.2
0.3
0.4
0.5
0.6
Importance scale

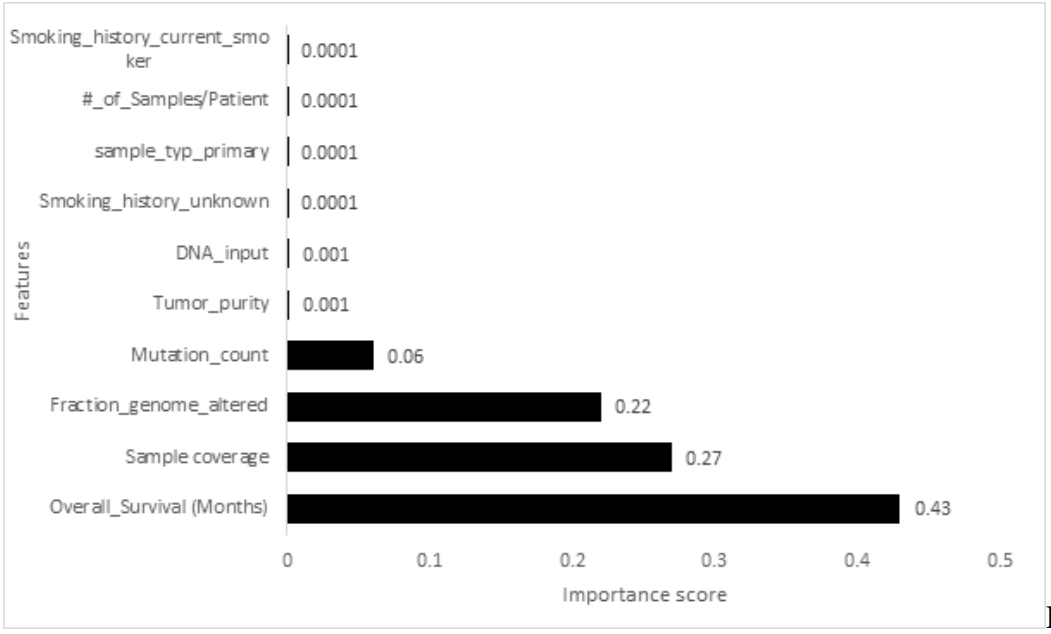
B


Race
0.00001
# of Samples/patient
0.00004
Tumor_Anatomic_site
0.00007
Sex
0.00009
Tumor_type
0.00017
Aneuploidy_score
0.002
Fraction_genome_altered
0.009
Overall_surv_month
0.09
Mutation_count
0.25
Age_at_diagnosis
0.62
Features
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Importance scale

C



D



E

Figure 4-5. Top 10 most informative features contributing to classification of survival status of ALL, GLIO, COAD, P-NB and MSK-IMPACT models. **A:** Clinical and genomic features such as cell of tumor origin, age at diagnosis, congenital abnormality, DNA index and Molecular subtypes were the most informative predictors in the ALL model. **B:** Age at diagnosis, mutation count, fraction genome altered, therapy, gene expression type, mutation type, sex, methylation status and number of samples per patients are the most informative predictors. **C:** In the COAD model, age at diagnosis, mutation count, overall survival in month, fraction genome altered were among the top 10 informative features. **D:** Clinical features including INSS stage, age at diagnosis, sex, race and risk group are predominantly the most informative predictor when compared with ploidy feature. **E:** Overall survival in months, fraction genome altered, mutation count, smoking history were among the most contributing features to the MSK-IMPACT model.

Interestingly, in this ALL model, the clinical features were more informative than genomic features (Figure 4-5A). In the GLIO model, age at diagnosis, mutation count, fraction genome altered, therapy type, gene expression type, mutation type, sex, methylation status and number of samples per patient contributed significantly to the model. Noteworthily, both clinical and genomic predictors contributed almost at an equal ratio to the model (Figure 4-5B).

In the COAD model, clinical features including age at diagnosis, mutation count, overall survival in months, tumor type, sex, tumor anatomic site, number of samples per patient and race were the most predominant predictors whereas fraction genome altered was the only contributing predictor in the model (Figure 4-5C). INSS_stage 2A, INSS_stage 3, INSS_stage 4S, age at diagnosis, INSS stage 2B, Sex (male), ploidy, race, risk group and ploidy (hyperdiploid) were the most statistical important features contributing to the P-NB model. Noteworthily, clinical features were predominantly the most informative predictors whereas the genomic features didn't contribute significantly to the P-NB model (Figure 4-5D). In the MSK-IMPACT model, overall survival (in months), sample coverage, fraction genome altered, mutation count, tumor purity, DNA input, smoking history (unknown), sample type (primary), number of samples per patient and smoking history (current smoker) were the most informative features in the model (Figure 4-5E). Strikingly, age at diagnosis and sex were the prominent clinical features in ALL, GLIO, COAD and P-NB models.

4.4.3 Survival Analysis

The median age at diagnosis for our acute lymphoblastic leukemia (ALL) and P-NB sample sizes are 8 and 3 years respectively (Figure 4-6). The mean value of the fraction genome altered (FGA) were 0.08 and 0.09 respectively even though a large number of the FGAs were unknown.

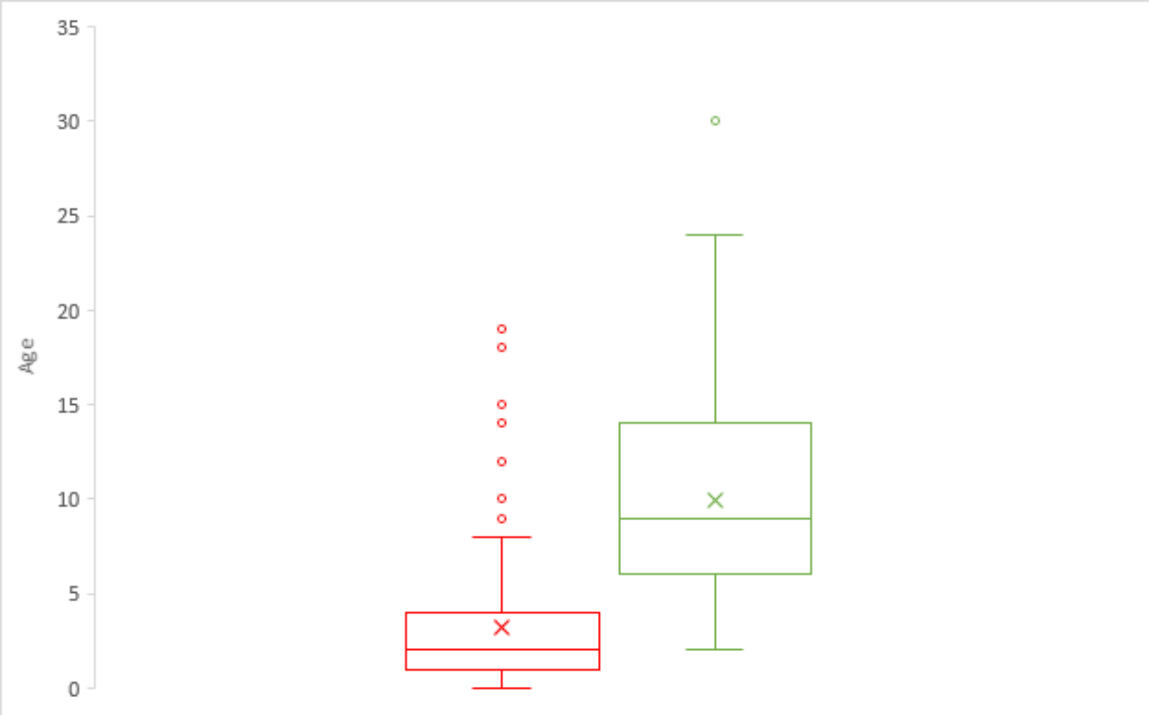


Figure 4-6. Summary statistics of age at diagnosis of ALL(left) and P-NB(right) with median ages of 3 and 8 respectively.

In this study, we observed the representation of American Indian, Asian, black, Pacific Islander, White (and White Hispanic) race/ ethnicity in the survival study (Figure 4-7). A significant number of patients' ethnic backgrounds were unreported. From our findings, children of Caucasian descent had the highest reported cases of ALL and P-NB followed by children of African descent. This confirms the findings of other studies such as Siegel et al., 2017 and Friedrich et al., 2017 that the risk of ALL and P-NB is slightly higher in white and Hispanic white children than in other races.

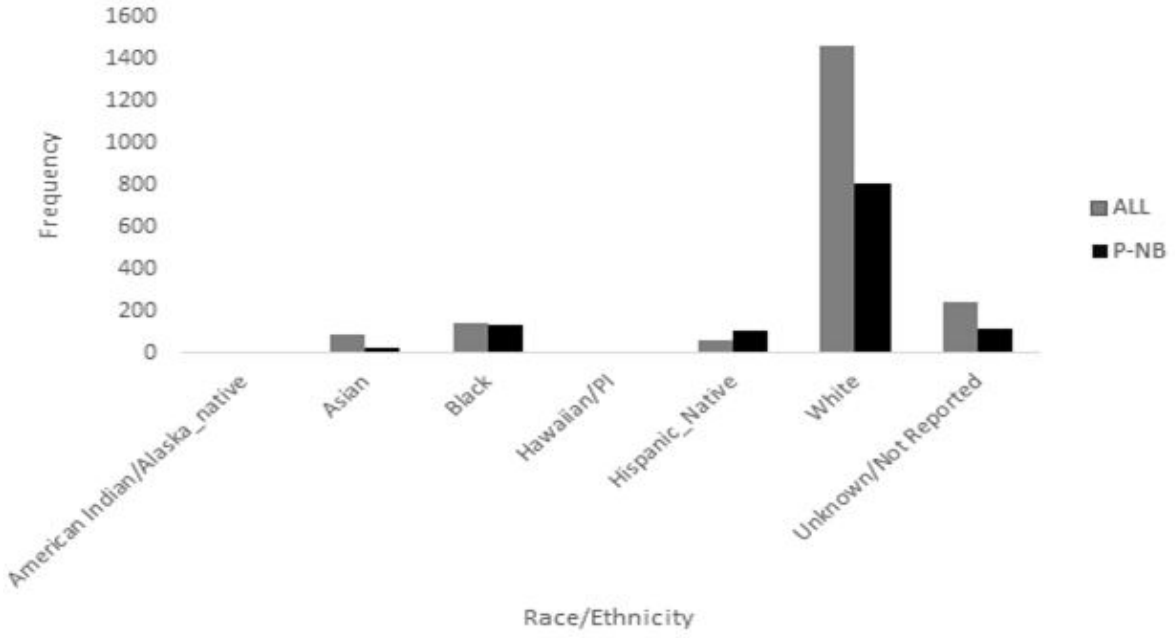


Figure 4-7. Distribution of ethnicity and race of ALL and pediatric neuroblastoma P-NB. The risk of ALL and neuroblastoma is slightly higher in white and Hispanic white children than in other races.

The best Cox model for ALL showed that the covariates age at diagnosis, cell of tumor origin (B and T-cells), congenital abnormality, DNA index, and MLL-rearranged molecular subtype were significant predictors of survival ($P < 0.05$). However, the other covariates such as molecular subtypes ETV6-RUNX1, hyperdiploid status 4 and 10 unknown, hyperdiploid without trisomy of both chromosomes 4 and 10, hypodiploid, TCF3-HLF, TCF3-PBX1, iAMP21 were not significant. The age at diagnosis with a hazard ratio of 1.03 shows that each additional year of age at diagnosis is associated with a 3% increase in the risk of dying. Similarly, the cell of tumor origin B-cell and T-cell with hazard ratios of 0.39 and 0.35 indicate that subjects with these genetic variations have 61 and 64% decrease in the risk of dying compared to the baseline group of B-cell precursor respectively. The congenital abnormality variable (No/Yes) with hazard ratios of 0.59 and 8.30 indicate a decrease of 41% and an increase of 730% in the risk of

dying compared to the baseline group with unknown congenital abnormality status. The molecular subtype had 12 levels, unknown, BCR-ABL1, ETV6-RUNX1, Hyperdiploid with a status of 4 and 10 unknown, Hyperdiploidy without trisomy of both chromosomes 4 and 10, Hypodiploid, iAMP21, MLL-Rearranged, None of above, TCF3-HLF, TCF3-PBX1, Trisomy of both chromosomes 4 and 10. Three of the categories, Hyperdiploidy without trisomy of both chromosomes 4 and 10, iAMP21, None of above with hazard ratios of 0.70, 0.68 and 1.67 show that the first two conditions decrease the hazard by 30 and 32% and the last increase the hazard by 67% compared to the baseline group with unknown molecular subtype. The TCF3-PBX1 biomarker had levels unknown, negative and positive. The negative category with a hazard ratio of 0.76 decreases the hazard by 24% compared to the baseline group of TCF3-PBX1 unknown. The trisomy 4 and 10 feature had levels, unknown, negative and positive. The negative category with a hazard ratio of 1.52 increases the hazard by 52% compared to the baseline group of trisomy 4 and 10 unknown (Table 4-3).

Table 4-3. Cox-proportional hazard results showing the effect size for pediatric acute lymphoblastic leukemia

Cancer type	Covariates	Coef	Exp(Coef)	S.E	Wald Z	Pr(>|z|)
ALL (Pediatric)	Age at diagnosis	0.03	1.03	0.005	6.66	2.75e-11
	BCR_ABL1_status(negative)	-0.33	0.72	0.60	-0.54	0.59
	BCR_ABL1_status(positive)	-0.27	0.76	0.68	-0.41	0.68
	Cell of tumor origin (B-cell)	-0.93	0.39	0.07	-13.29	<2e-16
	Cell of tumor origin (T-cell)	-1.04	0.35	0.08	-11.40	<2e-16
	Congenital abnormality (no)	-0.53	0.59	0.18	-3.59	<0.001
	Congenital abnormality (yes)	2.12	8.30	0.15	9.17	<0.001
	MS = BCR-ABL1	0.18	1.19	0.36	0.49	0.63
	MS = ETV6-RUNX1	0.01	1.00	0.11	0.01	0.99
	MS =Hyperdiploid with status of 4 and 10 unknown	-0.18	0.83	0.25	-0.75	0.46
	MS= Hyperdiploidy without trisomy of both chromosomes 4 and 10	-0.40	0.70	0.13	-2.73	0.01
	MS=Hypodiploid	0.39	1.47	0.31	1.27	0.20
	MS=iAMP21	0.51	1.67	0.26	1.96	0.05
	MS=MLL-Rearranged	0.14	1.15	0.19	0.73	0.47
	MS=None of above	-0.39	0.68	0.12	-3.19	0.001
	MS=TCF3-HLF	0.60	1.83	0.58	-1.04	0.30
	MS=TCF3-PBX1	0.15	1.17	0.29	0.52	0.60
	MS=Trisomy of both chromosomes 4 and 10	0.15	1.16	0.17	0.89	0.37
	TCF3_PBX1_status(negative)	-0.27	0.76	0.07	-3.92	<0.001
	TCF3_PBX1_status (positive)	-0.29	0.75	0.28	-1.06	0.29
	Trisomy 4 and 10 (negative)	0.42	1.52	0.07	5.23	<0.001
	Trisomy 4 and 10 (positive)	0.23	1.25	0.13	1.73	0.08

Similarly, the best Cox proportional hazards model for pediatric neuroblastoma showed that the covariates INSS stages (2a,2b,3,4,4s) and hyperdiploid were significant ($P < 0.05$). However, the covariate unknown ploidy was not significant with p values: 0.4884. The p-values of INSS stages- 2,2b,3,4 and 4s diagnosis with a hazard ratio HR exp(coef) = 6.2115, 4.3297, 3.5981, 9.4593 and 0.6928 indicating a strong relationship between the sex, INSS stage and the increased risk of mortality. Holding other covariates constant, a higher hyperdiploid is associated with poor survival. Relatedly, sex =male (p-value: 0.02) with a HR = 0.8918, indicating a very strong relationship between the male sex and increased risk of death. (Table 4-4).

Table 4-4. Cox-Proportional hazard results showing the effect size of pediatric neuroblastoma

	Covariates	**Coef**	**Exp(Coef)**	**S.E**	**Wald Z**	**Pr(>|z|)**
NB	Age at Diagnosis	-0.01	0.98	0.03	-0.83	0.40
	INSS Stage 1	-0.84	0.44	0.19	-4.31	1.60e-05
	INSS Stage 2a	-0.63	0.53	0.31	-1.98	0.04
	INSS Stage 2b	0.03	1.02	0.25	0.10	0.92
	INSS Stage 3	-0.04	0.96	0.19	-0.18	0.85
	INSS Stage 4	-0.67	0.51	0.18	-3.76	0.0017
	INSS Stage 4s	0.9916	2.6955	0.4747	2.09	0.0367
	MYCN = Amplified	0.28	1.33	0.40	0.71	0.48
	MYCN=Not Amplified	-0.11	0.90	0.11	-0.97	0.33
	Ploidy_Diploid(DI =1)	-0.367	0.6928	0.0952	-3.86	0.0001
	Ploidy = Hyperdiploid (DI >1)	0.87	2.39	0.21	4.14	3.42e-05
	Sex=Male	-0.12	0.89	0.05	-2.33	0.02
	Sex = Female	0.07	1.07	0.09	0.82	0.41

4.5 References

Allemani, C., Weir, H. K., Carreira, H., Harewood, R., Spika, D., Wang, X.-S., … Coleman, M. P. (2015). Global surveillance of cancer survival 1995–2009: analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2). The Lancet, 385(9972), 977–1010. doi: 10.1016/s0140-6736(14)62038-9

American Cancer Society. (n.d.). Cancer Facts & Figures 2020. Retrieved from https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2020.html

Angermueller, C., Parnamaa, T., Parts, L., & Stegle, O. (n.d.). Deep learning for computational biology. Molecular Systems Biology, 12(7), 878. doi: https://doi.org/10.15252/msb.20156651

Bielorai, B., Fisher, T., Waldman, D., Lerenthal, Y., Nissenkorn, A., Tohami, T., … Toren, A. (2013). Acute Lymphoblastic Leukemia in Early Childhood as the Presenting Sign of Ataxia-Telangiectasia Variant. Pediatric Hematology and Oncology, 30(6), 574–582. doi: 10.3109/08880018.2013.777949

Brenner, H., Rothenbacher, D., & Arndt, V. (2009). Epidemiology of Stomach Cancer. Methods in Molecular Biology Cancer Epidemiology, 467–477. doi: 10.1007/978-1-60327-492-0_23

Cancer Facts & Figures 2018. (n.d.). Retrieved from https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2018.html

Cancer Genome Atlas Research Network. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 455(7216), 1061–1068. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/18772890

Carey, R. M., Fathy, R., Shah, R. R., Rajasekaran, K., Cannady, S. B., Newman, J. G., … Brant, J. A. (2020). Association of Type of Treatment Facility With Overall Survival After a Diagnosis of Head and Neck Cancer. JAMA Network Open, 3(1). doi: 10.1001/jamanetworkopen.2019.19697

Cavalli, L. R., & Cavalli, I. J. (2013). Molecular Classification and Prognostic Signatures of Breast Tumors. Oncoplastic and Reconstructive Breast Surgery, 55–62. doi: 10.1007/978-88-470-2652-0_5

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16. doi: 10.1145/2939672.2939785

Chen, X., Sun, X., & Hoshida, Y. (2014). Survival analysis tools in genomics research. Human Genomics, 8(1). doi: 10.1186/s40246-014-0021-z

Chessells, J. M., Harrison, G., Richards, S., Hill, F., Bailey, C., & Gibson, B. (2001). Downs syndrome and acute lymphoblastic leukaemia: clinical features and response to treatment. Archives of Disease in Childhood, 85(4), 321–325. doi: 10.1136/adc.85.4.321

Cheung, N.-K. V., & Dyer, M. A. (2013). Neuroblastoma: developmental biology, cancer genomics and immunotherapy. Nature Reviews Cancer, 13(6), 397–411. doi: 10.1038/nrc3526

Coleman, E. A., Coon, S., Hall-Barrow, J., Richards, K., Gaylor, D., & Stewart, B. (2008). RE: Coleman et al. Feasibility of Exercise During Treatment for Multiple Myeloma. Cancer Nursing. 2003;26(5):410-419. Cancer Nursing, 31(4), 263–264. doi: 10.1097/01.ncc.0000305734.80592.40

Cronin, K. A., Lake, A. J., Scott, S., Sherman, R. L., Noone, A.-M., Howlader, N., … Jemal, A. (2018). Annual Report to the Nation on the Status of Cancer, part I: National cancer statistics. Cancer, 124(13), 2785–2800. doi: 10.1002/cncr.31551

Fabre, E., Rivera, C., Mordant, P., Gibault, L., Dujon, A., Foucault, C., … Riquet, M. (2015). Evolution of induction chemotherapy for non-small cell lung cancer over the last 30 years: A surgical appraisal. Thoracic Cancer, 6(6), 731–740. doi: 10.1111/1759-7714.12250

Friedman, J. (2001). GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE. Annals of Statistics, 29(5), 1189–1232.

Friedrich, P., Itriago, E., Rodriquez-Galindo, C. & Ribeiro, K. (2017). Racial and ethnic disparities in the incidence of pediatric extracranial embryonal tumors. JNCI, 109(10).

doi: https://doi.org/10.1093/jnci/djx050

German, J. (1997). Bloom's syndrome. XX. The first 100 cancers. Cancer Genetics and Cytogenetics, 93(1), 100–106. doi: 10.1016/s0165-4608(96)00336-6

Gorey, K. M. (2009). Breast cancer survival in Canada and the USA: meta-analytic evidence of a Canadian advantage in low-income areas. International Journal of Epidemiology, 38(6), 1543–1551. doi: 10.1093/ije/dyp193

Gérinière, L., Bastion, Y., Dumontet, C., Salles, G., Espinouse, D., & Coiffier, B. (1994). Heterogeneity of acute lymphoblastic leukemia in HIV-seropositive patients. Annals of Oncology, 5(5), 437–440. doi: 10.1093/oxfordjournals.annonc.a058876

Hoelzer, D., Bassan, R., Dombret, H., Fielding, A., Ribera, J., & Buske, C. (2016). Acute lymphoblastic leukaemia in adult patients: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of Oncology, 27, v69–v82. doi: 10.1093/annonc/mdw025

Howlader , N., Noone, A. M., Krapcho, M., Miller, D., Brest, A., Yu, M., … Cronin, K. A. (2019). SEER Cancer Statistics Review, 1975-2016, National Cancer Institute. Https://Seer.cancer.gov/Csr/1975_2016/, Based on November 2018 SEER Data Submission. Retrieved from https://seer.cancer.gov/statfacts/html/alyl.html

Hunger, S. P., & Mullighan, C. G. (2015). Acute Lymphoblastic Leukemia in Children. The New England Journal of Medicine, 373(16), 1541–1552. doi: 10.1056/NEJMra1400972

Imai, K., Allard, M.-A., Benitez, C. C., Vibert, E., Cunha, A. S., Cherqui, D., … Adam, R. (2016). Nomogram for prediction of prognosis in patients with initially unresectable colorectal liver metastases. British Journal of Surgery, 103(5), 590–599. doi: 10.1002/bjs.10073

Jabbour, E., Obrien, S., Konopleva, M., & Kantarjian, H. (2015). New insights into the pathophysiology and therapy of adult acute lymphoblastic leukemia. Cancer, 121(15), 2517–2528. doi: 10.1002/cncr.29383

Jain, N., Lamb, A. V., O'Brien, S., Ravandi, F., Konopleva, M., Jabbour, E., … Khoury, J. D. (2016). Early T-cell precursor acute lymphoblastic leukemia/lymphoma (ETP-ALL/LBL) in adolescents and adults: a high-risk subtype. Blood, 127(15), 1863–1869. doi: 10.1182/blood-2015-08-661702

Kalinin, A. A., Higgins, G. A., Reamaroon, N., Soroushmehr, S., Allyn-Feuer, A., Dinov, I. D., … Athey, B. D. (2018). Deep learning in pharmacogenomics: from gene regulation to patient stratification. Pharmacogenomics, 19(7), 629–650. doi: 10.2217/pgs-2018-0008

Kato, A., Shimizu, H., Ohtsuka, M., Yoshitomi, H., Furukawa, K., Takayashiki, T., … Miyazaki, M. (2015). Downsizing Chemotherapy for Initially Unresectable Locally Advanced Biliary Tract Cancer Patients Treated with Gemcitabine Plus Cisplatin Combination Therapy Followed by Radical Surgery. Annals of Surgical Oncology, 22(S3), 1093–1099. doi: 10.1245/s10434-015-4768-9

Kroll, C. N., & Song, P. (2013). Impact of multicollinearity on small sample hydrologic regression models. Water Resources Research, 49(6), 3756–3769. doi: 10.1002/wrcr.20315

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature Selection: A Data Perspective. ACM Comput Surv, 9(4), 1–44. Retrieved from https://arxiv.org/pdf/1601.07996.pdf

Lim, H.-S., Roychoudhuri, R., Peto, J., Schwartz, G., Baade, P., & Møller, H. (2006). Cancer survival is dependent on season of diagnosis and sunlight exposure. International Journal of Cancer, 119(7), 1530–1536. doi: 10.1002/ijc.22052

Meguid, R. A., Hooker, C. M., Harris, J., Xu, L., Westra, W. H., Sherwood, J. T., … Brock, M. V. (2010). Long-term Survival Outcomes by Smoking Status in Surgical and Nonsurgical Patients With Non-small Cell Lung Cancer. Chest, 138(3), 500–509. doi: 10.1378/chest.08-2991

Miller, K. D., Siegel, R. L., Khan, R., & Jemal, A. (2018). Cancer Statistics. Cancer Rehabilitation. doi: 10.1891/9780826121646.0002

Miyamoto, Y., Suyama, K., & Baba, H. (2017). Recent Advances in Targeting the EGFR Signaling Pathway for the Treatment of Metastatic Colorectal Cancer. International Journal of Molecular Sciences, 18(4), 752. doi: 10.3390/ijms18040752

Mohseni, M. M., Uludag, H. M., & Brandwein, J. M. (2018). Advances in Biology of Acute Lymphoblastic Leukemia (ALL) and Therapeutic Implications. American Journal Blood Research, 8(4), 29–56.

Monclair, T., Brodeur, G. M., Ambros, P. F., Brisse, H. J., Cecchetto, G., Holmes, K., … Pearson, A. D. (2009). The International Neuroblastoma Risk Group (INRG) Staging System: An INRG Task Force Report. Journal of Clinical Oncology, 27(2), 298–303. doi: 10.1200/jco.2008.16.6876

Mur, P., Mollejo, M., Ruano, Y., Lope, Á. R. D., Fiaño, C., García, J. F., … Meléndez, B. (2013). Codeletion of 1p and 19q determines distinct gene methylation and expression profiles in IDH-mutated oligodendroglial tumors. Acta Neuropathologica, 126(2), 277–289. doi: 10.1007/s00401-013-1130-9

National Cancer Intelligence Network Trends in incidence ... (n.d.). Retrieved from http://www.ncin.org.uk/view?rid=2818

Nitsche, U., Wenzel, P., Siveke, J. T., Braren, R., Holzapfel, K., Schlitter, A. M., … Kleeff, J. (2015). Resectability After First-Line FOLFIRINOX in Initially Unresectable Locally Advanced Pancreatic Cancer: A Single-Center Experience. Annals of Surgical Oncology, 22(S3), 1212–1220. doi: 10.1245/s10434-015-4851-2

Noushmehr, H., Weisenberger, D. J., Diefes, K., & Phillips, H. S. (n.d.). Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2872684/?report=reader#__ffn_sectitle

Paja, W., Pancerz, K., & Grochowalski, P. (2017). Generational Feature Elimination and Some Other Ranking Feature Selection Methods. Advances in Feature Selection for Data and Pattern Recognition Intelligent Systems Reference Library, 97–112. doi: 10.1007/978-3-319-67588-6_6

Pulte, D., Redaniel, M. T., Jansen, L., Brenner, H., & Jeffreys, M. (2013). Recent trends in survival of adult patients with acute leukemia: overall improvements, but persistent and partly increasing disparity in survival of patients from minority groups. Haematologica, 98(2), 222–229. doi: 10.3324/haematol.2012.063602

Sehgal, S., Mujtaba, S., Gupta, D., Aggarwal, R., & Marwaha, R. (2010). High incidence ofEpstein Barr virusinfection in childhood acute lymphocytic lukemia: A preliminary study. Indian Journal of Pathology and Microbiology, 53(1), 63. doi: 10.4103/0377-4929.59186

Shah, A., John, B. M., & Shondi, V. (2013). Acute lymphoblastic leukemia with treatment--naive Fanconi anemia. Indian Pediatr. , 50(5), 508–510. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/23778731

Siegel, R. L., Miller, K. D. & Jemal, A. (2017). Cancer Statistics, 2017. CA: A cancer Journal for Clinicians, 60(1),7-30. Doi: https://doi.org/10.3322/caac.21387

Spector, L. G., Ross, J. A., Robison, L. L., & Bhatia, S. (n.d.). Epidemiology and etiology. Childhood Leukemias, 48–66. doi: 10.1017/cbo9780511471001.004

Stewart, E., Federico, S., Karlstrom, A., Shelat, A., Sablauer, A., Pappo, A., & Dyer, M. A. (2016). The Childhood Solid Tumor Network: A new resource for the developmental biology and oncology research communities. Developmental Biology, 411(2), 287–293. doi: 10.1016/j.ydbio.2015.03.001

Telenti, A., Lippert, C., Chang, P.-C., & Depristo, M. (2018). Deep learning of genomic variation and regulatory network data. Human Molecular Genetics, 27(Supplement_R1). doi: 10.1093/hmg/ddy115

Terwilliger, T., & Abdul-Hay, M. (2017). Acute lymphoblastic leukemia: a comprehensive review and 2017 update. Blood Cancer Journal, 7(6). doi: 10.1038/bcj.2017.53

Tobias, J., & Hochhauser, D. (2015). Cancer and its management (7th edition) (7th ed.). Wiley Blackwell.

Understanding Statistics Used to Guide Prognosis and Evaluate Treatment. (2019, September 10). Retrieved from https://www.cancer.net/navigating-cancer-care/cancer-basics/understanding-statistics-used-guide-prognosis-and-evaluate-treatment

Vatcheva, K. P., & Lee, M. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. Epidemiology: Open Access, 06(02). doi: 10.4172/2161-1165.1000227

Wang, F., Liu, Z.-Y., Xia, Y.-Y., Zhou, C., Shen, X.-M., Li, X.-L., … Li, W. (2015). Changes in neutrophil/lymphocyte and platelet/lymphocyte ratios after chemotherapy correlate with chemotherapy response and prediction of prognosis in patients with unresectable gastric cancer. Oncology Letters, 10(6), 3411–3418. doi: 10.3892/ol.2015.3783

Wang, W., & Albert, J. M. (2016). Causal mediation analysis for the Cox proportional hazards model with a smooth baseline hazard estimator. Journal of the Royal Statistical Society: Series C (Applied Statistics), 66(4), 741–757. doi: 10.1111/rssc.12188

Yang, C.-F. J., Meyerhoff, R. R., Stephens, S. J., Singhapricha, T., Toomey, C. B., Anderson, K. L., … Berry, M. F. (2015). Long-Term Outcomes of Lobectomy for Non-Small Cell Lung

Cancer After Definitive Radiation Treatment. The Annals of Thoracic Surgery, 99(6), 1914–1920. doi: 10.1016/j.athoracsur.2015.01.064

Yang, S. W., Kim, M. G., Lee, J. H., & Kwon, S. J. (2013). Role of Metastasectomy on Overall Survival of Patients with Metastatic Gastric Cancer. Journal of Gastric Cancer, 13(4), 226. doi: 10.5230/jgc.2013.13.4.226

Zhang, C., Vinyals, O., Munos, R., & Bengio, S. (2018). A Study on Overfitting in Deep Reinforcement Learning. Cornell University, ArXiv:1804.06893. Retrieved from https://arxiv.org/pdf/1804.06893.pdf

Zheng, J., Chu, H., Struppa, D., Zhang, J., Yacoub, S. M., El-Askary, H., … Rakovski, C. (2020). Optimal Multi-Stage Arrhythmia Classification Approach. Scientific Reports, 10(1). doi: 10.1038/s41598-020-59821-7

Zheng, Y., Li, G., & Zhang, T. (2019). An Improved Over-sampling Algorithm based on iForest and SMOTE. Proceedings of the 2019 8th International Conference on Software and Computer Applications - ICSCA 19, 75–80. doi: 10.1145/3316615.3316641

5 Machine Learning Reclassification of Variants of Uncertain Significance

Oluyemi Oluyemi and Kristalee Lio

Author's Contribution

YO and KL contributed to the design of the study and interpretation of the results. YO and KL analyzed and wrote the manuscript. All the authors read and approved the final manuscript**.**

5.1 Abstract

Background: Variants of uncertain significance remain a major challenge in contemporary genetic variation screening methods. The advent of next-generation sequencing and other bioinformatics resources have uncovered a significant number of variants of uncertain significance which often hampers clinical decision-making.

Methods: We proposed an ensemble machine learning (ML) approach to classifying clinical implication value of variants of unknown significance (VUS) using known mutation scores (PolyPhen, SIFT and mutation assessor), allele frequency, mutation type (insertion, deletion, frame shift), gnomAD among other features. Our machine learning models were evaluated using precision, recall, f-measure and area under the curve metrics.

Result: Our results indicated that extreme gradient boosting and deep neural network classifiers had the best performance in the VUS and biological effect classification models respectively. Also, our results show that mutation predictive scores were the most informative features in the clinical implication classification of the VUS.

Conclusion: The combination of ML model, genetic profiling and clinical information can contribute significantly to our understanding and interpretation of genetically related cancers

5.2 Introduction

Recent breakthroughs in next-generation sequencing (NGS) have played significant roles in our understanding of human genetic variations. NGS technologies such as multigene panels, whole-exome sequencing (WES), whole-genome sequencing (WGS) and genome-wide association studies (GWAS) have paved the way for the synchronized analysis of many genes and the generation of millions of short nucleic acid sequences in parallel (Oulas et al.,2019; Li et al., 2017; Shendure and Ji, 2008; Bentley et al., 2008). Most variants identified by NGS in tumor cells can be traced to carcinogenic complexity (Ding et al., 2012). NGS results acquired using DNA or RNA extracted from tumor tissue often illustrate a convoluted molecular signature that varies from that of normal tissue for any given patient (Li et al., 2017).

More recently, there is a consensus to incorporate NGS methodologies in clinical diagnosis, prognosis, and overall patient counseling. This can be attributed to their cost-effectiveness and accuracy as evidenced by the increase in FDA approval of NGS methodologies (Green and Guyer 2011; Collins and Hamburg 2013). However, the overwhelming amount of variants shown in each patient has called into question the interpretation of clinically significant genetic variation (Yorczyk et al. 2015).

Cancer often arises from two to five mutations in various genes. These mutations are complex and with a combinatorial effect commonly lead to tumor proliferation and cancer progression (Levine et al.,2019). Oncogenes and tumor suppressor genes play prominent roles in cancer progression. An oncogene arises from mutation of proto-oncogenes which often leads to activation of uncontrolled cell growth. Most cancer-inducing mutations are acquired and activate oncogenes by gene duplication and chromosomal rearrangements. Tumor suppressors normally slow down cell division, regulate apoptosis and repair DNA errors. An anomaly in tumor

suppression mechanisms can lead to abnormal cell growth. The major difference between a tumor suppressor gene and an oncogene is that the former stems from the inactivation of proto-oncogenes while the latter causes cancer when activated. Studies have shown that most oncogenes and tumor suppressor genes are acquired and not inherited (Mair et al., 2016; Brognard et al., 2011).

Mutation can confer an enhanced activity on a protein or more often reduce the functionality of a protein. Loss of function mutations in tumor-suppressor genes is often oncogenic. Five main protein types are generally identified as tumor-suppressor gene encoders: enzymes that participate in DNA repairs; proteins that promote apoptosis; intracellular proteins that regulate progression through a specific stage of cell cycle; receptors for secreted hormones that function to inhibit cell proliferation; and checkpoint proteins that stop the cell cycle if DNA is damaged. A gain of function mutations is not as common as the loss of function. However, activation of a proto-oncogene into an oncogene involves a gain of function mutation. For example, *ras* gene a proto-oncogene that encodes an intracellular signal-transduction protein; *rasD* gene, a *ras* mutant is an oncogene, which encodes oncoprotein provides an uncontrolled growth-promoting signal (Lodish et al., 2000).

Clinical mutation screening (CMS) of cancer predisposition genes for the presence of the mutation is often used for identifying patients with an elevated risk of cancer (Lindor et al., 2012). The outcome from CMS is broadly categorized as pathogenic variation, benign mutation, and variants of unknown significance (VUS).

Pathogenic mutation is genetic alterations with sufficient evidence to classify as capable of causing disease. Well studied tumor suppressor genes such as APC, BRCA1, BRCA2, NF1, TP53 have been linked with pathogenic (or likely pathogenic) variation (He et al. ,2016). For

example, the APC gene encodes a protein that is prominent in tumor suppression in WNT signaling pathway. Studies have shown that the mutational loss of APC function contributes to tumor proliferation and may lead to prostate cancer and colorectal cancer (Andres et al.,2018; He et al. ,2016; Markowitz and Bertagnolli 2009: Chen et al., 2013). Benign tumors are generally non-life threatening, although there are exceptions to this assumption. Benign tumors, for example, may have fatal consequences in the brain. The skull is inadequately equipped to contain tumor growth in the brain. Most benign tumors do not progress to malignancy. Benign cells retain normal cell's functionality. The majority of benign tumors have fewer mutations and these mutations do not drive the more mutational proliferation (Wang et al., 2016). For example, Rutkowski et al., 2000 showed that identifying defective cell types in NF1-associated neurofibroma is important to understanding benign tumor mechanisms.

Variants of uncertain significance (VUS) are genetic mutations that have an unclear impact on protein function (Alosi et al., 2017 and Easton et al., 2007). VUS are often amino acid substitutions, intronic variants or small in-frame insertions or deletions. The actual effect of these variants on gene function is unknown, this makes the assessment of their clinical significance quite complex. About 50% of variants reported in well-studied genes such as p53, BRCA1 and BRCA2 are designated as VUS based on the level of clinical evidence (Larson et al. 2014).

The number of VUS in the residual set remains substantial. The degree of pathogenicity is often not trivial, considering that nearly fifty percent of the unique variants are novel, and cannot be resolved using published literature and variant databases (Foley et al., 2015). Loss-of-function variants represent a very small fraction of known variants. Other identified variants are missense and synonymous variants in the exon, single nucleotide changes, or in-frame insertions or deletions in intervening and intergenic regions (Mucaki et al., 2016). Several genetic

susceptibility testing cancer patients often receive inconclusive and uncertain results that can be attributed to inconsistent accuracy in *silico* protein-coding prediction tools which often becomes challenging for clinical risk diagnosis. Variants of uncertain significance such as insertions /deletions (INDELs) and single nucleotide polymorphisms (SNPs) cause dilemmas for clinicians and uncertainty on how to advise patients.

In this study, we propose a xgboost machine learning (ML) approach to reclassify the clinical significance value of VUS using a variety of prediction scores, mutation type and biological effect namely: total number of nonsynonymous mutations in the sample, variant allele frequency in the tumor sample, genome Aggregation Database (gnomAD), missense, fusion and in-frame deletion) features as predictors. Our ML methodology (1) utilizes probability for clinical significance class discrimination, (2) reclassifies a VUS' clinical implication based on the highest probability value, (3) Compares and evaluates the predictive power of various VUS classifiers using precision, recall and ROC metrics, and (4) ranks the features based on their level of informativeness .

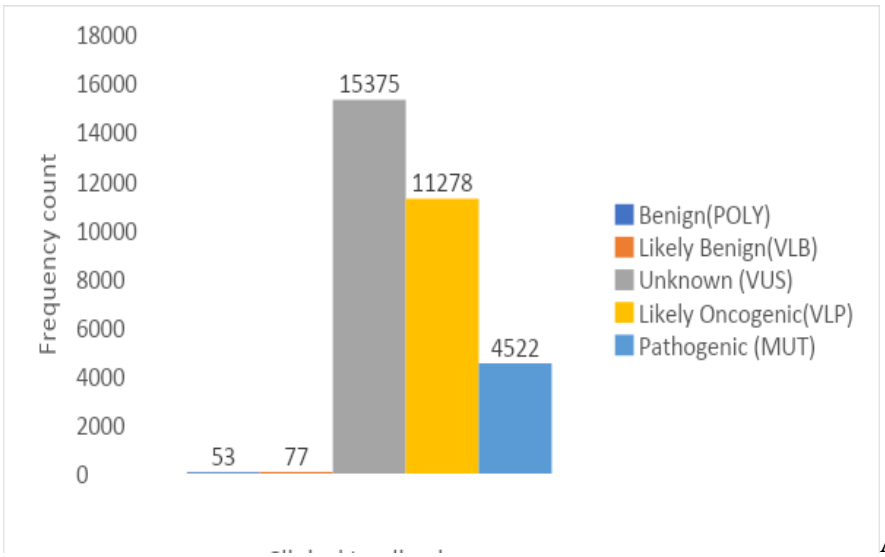
5.3 Methods

In this study, we examined the prediction capability of xgboost model at reclassifying VUS based on ACMG clinical values of cancer linked variants. The predictive performance of the xgboost model was compared with multilayer perceptron and polynomial support vector machines (SVM) models.

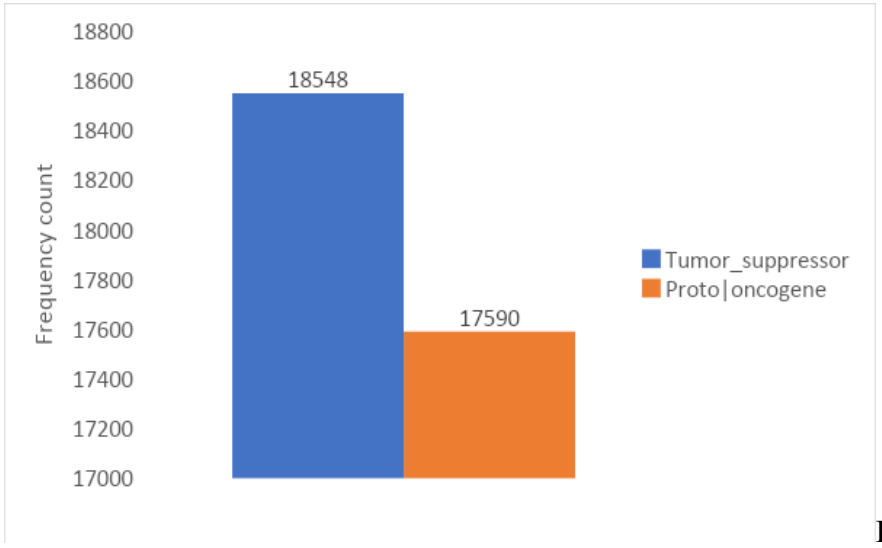
5.3.1 Genomic Variation Dataset

Open access CBioPortal and ClinVar databases were the primary sources of datasets (Cerami et al., 2012; Gao et al., 2013; Landrum et al., 2017). A total of distinct 145 oncogenes and tumor

suppressor genes (TSG) including *EGFR, KRAS, BRAF, MYC, BRCA (1 and 2), JAK2, IL2,* and *TP53* were used to aggregate the datasets for this study (Table 5-1). These cancer-linked genes are the genes that have been well studied and have enough data for genomic analysis. The dataset consisted of functionally validated 18548 tumor suppressor genes samples and 17590 oncogenic samples (Figure 5-2). A total of 43 functional and structural features including missense variant, upstream gene variant, regulatory region variant, allele frequency, mutation type, biological response(biotype) were used for the ML experiment. The classification models were subsequently used for reclassification of VUS by considering data from 18 primary cancer sites including liver, skin, ovary, prostate, kidney, lymphoid, stomach, bladder, cervix, pancreas, thyroid, uterus, myeloid, bowel.



A



B

Figure 5-1: Exploratory data analysis of the clinical implication of cancer-linked variations and cancer gene types. (A) Summary distribution of clinical significance of genetic variants. (B) Statistical distribution of cancer related genes- tumor suppressors and oncogenes

5.3.2 Clinical Significance Predictors: Feature Selection and Feature Importance Analysis

The features selected used for VUS reclassification were obtained from cbioportal for cancer genomics. The features used for the reclassification are broadly classified into (1) Mutation type (splice region, missense, in-frame insertion, in-frame deletion, fusion); (2)

Prediction scores (SIFT, PROVEAN, MutationAssessor, GERP, CADD, PolyPhen-2, FATHMM, MutationTaster); (3) Sequence sample features (allele frequency, copy deletion, copy gain, mutant in sample); (4) Biological effect impact (intron variant, splice donor variant, 3 prime UTR variants, missense variant, deleterious impact, chromosome location, cancer type, loss of function, gain of function); and (5) Biotype/biological response (nonstop decay, misc RNA, snoRNA, processed transcript, retained intron, CTCF binding site, processed pseudogene, polymorphic pseudogene and translated processed pseudogene).

Clinical significance (risk level) was used as the response/target feature. Multiple clinical and genetic evidence were combined to classify variants into 5 distinct risk levels: (1) benign variants, (2) likely benign (VLB), (3) VUS, (4) likely pathogenic (VLP) and (5) pathogenic variant (Table 5-2) (Richards et al., 2015; Karam et al., 2019). The benign variation of clinical value includes polymorphic sequence variants which normally do not lead to disease progression. A gene is considered polymorphic if more than one allele occupies the locus of a gene within a population and its alleles must have a frequency of one percent or more in the population. They are often found outside of genes and have neutral biological and function effects. Polymorphic variation affects drug responses and contributes to disease susceptibility as in the case of single nucleotide polymorphism. VLBs are alterations with significant evidence against pathogenicity, they are not expected to play a role in disease progression. However, additional evidence is usually needed to ascertain the clinical significance of these variants in variation disease paradigm. VUSs are variations in genetic sequence for which association with disease risk is unclear. VUS consists of a part of the gene that exhibits some alterations however there is no sufficient molecular information to determine if its benign or a risk factor for cancer. Over time, data is collected for reclassification of VUS. Generally, VUS are not used in clinical

decision making, therefore other factors are considered along with VUS to make any clinical diagnosis. VLPs have a high probability of pathogenicity. The identified variant is considered the likely cause of the cancer but with a degree of uncertainty. The variants often result in impetuous truncation in a gene where loss of function has been established as a mechanism of pathogenicity for cancer development. Pathogenic variation involves alterations that increase susceptibility and contribute to cancer development. Some pathogenic variants may not be wholly penetrant. Some of these variants may not be adequately equipped to cause cancer on their own such as in the case of recessive conditions. Usually, additional evidence has no impact on the classification of the variants.

We computed 243 features for each sample in the datasets. The VUS data was queried and set aside as unlabeled data. The remainder of the dataset (benign, VLB, VLP and pathogenic samples) was resampled and partitioned into train and test sets (70-30 percent ratio) so as to remove bias and prevent model overfitting. The train set consists of 70.80 % VLP, 28.39 % MUT, 0.48 % VLB and 0.33 % POLY samples (Figure 5-2A). The training set was subjected to outright removal of redundant features with high sparsity and recursive feature elimination process, where one by one a feature is removed, and the model trained on the resulting data set. If the accuracy of the model stays above a predefined threshold, the feature is removed permanently, and the process is repeated. If removal of any feature increases the accuracy, the threshold increases also. Single value elimination was used. We set a threshold of 0.001 less accuracy to declare a feature important to prediction, resulting in a final data set of 43 features. The test set contained 30% of the samples from the original data set, ensuring that the distribution of variants of known significance were equivalent to that of the original data set

A statistical test of dependence- *Pearson's correlation* was done to assess the pairwise relationships among the predictors. Ideally, orthogonal pairwise relationships are ideal and desirable for building machine learning models. The goal of the multicollinearity check is to identify pairwise relationship (high intercorrelations) that have potential of causing overfitting in the machine learning models (P. Vatcheva et al., 2016; Atems and Bergtold, 2015, Kroll and Song 2013). Data discrepancies- missing values imputation was handled by using nearest neighbor and median values approach depending on the data kind (categorical or numerical features).Other discrepancies were handled using outlier detection techniques and min-max feature rescaling for data normalization while regular expression for feature extraction of string features (Geng et al., 2018).

$$\widehat{\beta} \equiv argmin(\| y - X\beta \|^2 + \lambda_2 \| \beta \|^2 + \lambda_1 \| \beta \|_1 \qquad (5\text{-}1)$$

Where $\hat{\beta}\ is\ penalized\ parameter$, $\| \beta \|^2$ is the quadratic penalty term,

$\lambda \rightarrow tuning\ parameter, \beta \rightarrow parameter\ estimates$

Table 5-1. List of oncogenes and tumor suppressors used for data collection and aggregation

Tumor Suppressors			Oncogenes		
APC	IL2	TNFAIP3	ABL1	EVI1	MYC
ARHGEF12	JAK2	TP53	ABL2	EWSR1	MYCL1
ATM	MAP2K4	TSC1	AKT1	FEV	MYCN
BCL11B	MDM4	TSC2	AKT2	FGFR1	NCOA4
BLM	MEN1	VHL	ATF1	FGFR1OP	NKFB2
BMPR1A	MLH1	WRN	BCL11A	FGFR2	NRAS
BRCA1	MSH2	WT1	BCL2	FUS	NTRK1
BRCA2	NF1		BCL3	GOLGA5	NUP214
CARS	NF2		BCL6	GOPC	PAX8
CBFA2T3	NOTCH1		BCR	HMGA1	PDGFB
CDH1	NPM1		BRAF	HMGA2	PIK3CA
CDH11	NR4A3		CARD11	HRAS	PIM1
CDK6	NUP98		CBLB	IRF4	PLAG1
CDKN2C	PALB2		CBLC	JUN	PPARG
CEBPA	PML		CCND1	KIT	PTPN11
CHEK2	PTEN		CCND2	KRAS	RAF1
CREB1	RB1		CCND3	LCK	REL
CREBBP	RUNX1		CDX2	LMO2	RET
CYLD	SDHB		CTNNB1	MAF	ROS1
DDX5	SDHD		DDB2	MAFB	SMO
EXT1	SMARCA4		DDIT3	MAML2	SS18
EXT2	SMARCB1		DDX6	MDM2	TCL1A
FBXW7	SOCS1		DEK	MET	TET2
FH	STK11		EGFR	MITF	TFG
FLT3	SUFU		ELK4	MLL	TLX1
FOXP1	SUZ12		ERBB2	MPL	TPR
GPC3	SYK		ETV4	MYB	USP6
IDH1	TCF3		ETV6		

Table 5-2. The classification of genetic variants, based on the ACMG guidelines

Allelic Variants	clinical significance value	
Polymorphism	Benign (POLY)	Benign
Variant-likely benign	Likely benign (VLB)	
Uncertain significance Inconclusive	Variant of uncertain significance (VUS)	VUS
Variant-likely pathogenic	Likely pathogenic (VLP)	Pathogenic
Somatic not response	Pathogenic (MUT)	

For comparison purposes, we examined the result of binary VUS classification. In the binary models, the clinical significance values POLY and VLB were recategorized as "Benign" variants while the VLP and MUT were redesignated as "pathogenic" variants.

5.3.3 Proposed Models: Extreme Gradient Boosting (xgboost)

In this experiment, we used xgboost as the primary classifier and compared its performance with two other ML classifiers: polynomial support vector machines (poly-SVM) and deep-neural network (multilayer perceptron) for VUS reclassification (Chen and Guestrin, 2016; Narasimhan and Agarwal 2017; Zhang et al., 2019; Agajanian et al., 2019 and Sakellaropoulos et al., 2019). The choice of classifiers was chosen based on the following criteria: speed and run time on train set and statistical- probabilistic interpretability of the classification outcome. It is important to note that nominal features such as cancer gene type, mutation type, and variation copy were processed for modeling using dummy variable (one-hot encoding) technique. Dummy variables ensure a non-rank numerical representation of the categorical features. xgboost, SVM-Poly, and NN classifiers were used for the 4 class and binary

reclassification of the clinical implication of the VUS. To avoid the drawback of overfitting of the classifiers and account for the skewness in the distribution of the benign-pathogenic binary class, we oversampled the clinical implication using the synthetic minority over-sampling technique -SMOTE (Zheng et al.,2015). For the model evaluation, we used the f-measure, precision, recall, sensitivity, specificity and area under the curve (AUC) metrics as earlier described in chapter 2. After evaluating the model performance, feature importance was done to rank the most informative features used by the model for *VUS* reclassification.

5.4 Results and Discussion

5.4.1 Machine learning classification of cancer linked genes on variant datasets: Prediction scores, Mutation type and Sequence sample features outperform biological response predictors

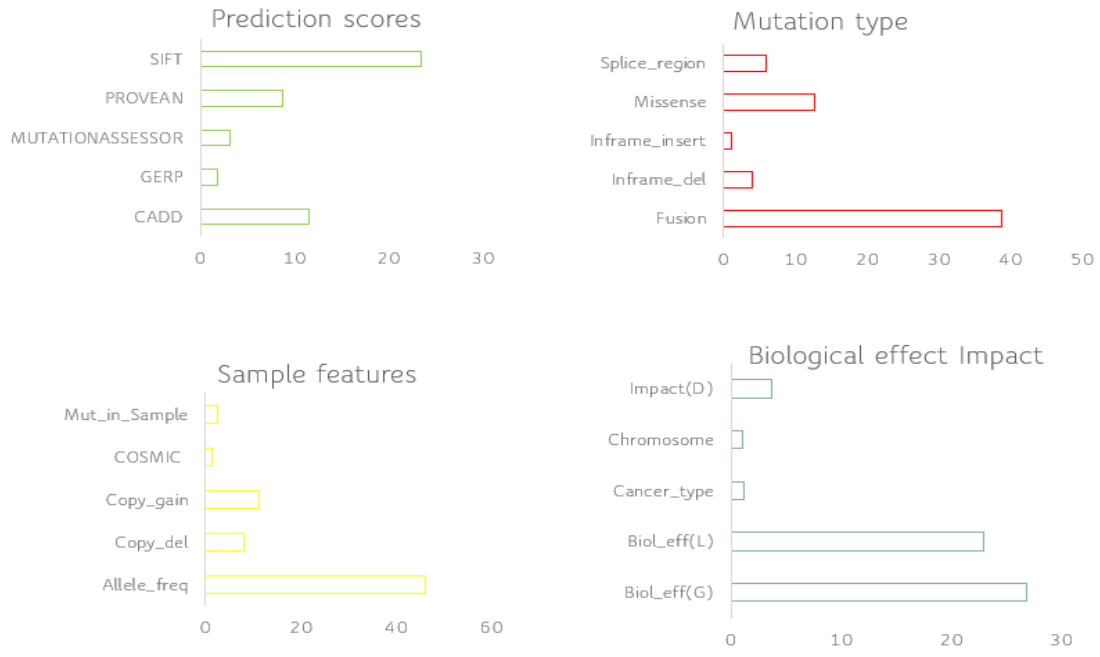
We trained the SVM, xgboost and MLP/D-NN models by considering cbioportal cancer-linked variants datasets and using a set of diverse features including mutation type features; biological response features; sequence sample features and prediction scores. In this analysis, we examined the performance of the three classifiers and focused on identifying the five most informative sets of features that contribute significantly to the VUS reclassification process (Figure 5-3). In our prediction score sets, SIFT was the most informative feature that contributed to the reclassification of VUS followed by PROVEAN, mutationassessor, Gerp and CADD. Whereas in the mutation_type set, fusion was the most informative predictor followed by missense, splice region, in-frame deletion and in-frame insertion (Figure 5-3A). In the sequence sample features set, allele frequency was the most significant feature that contributed to the VUS reclassification while copy gain, copy deletion, mutant in a sample and COSMIC occurrence came in second, third, fourth and fifth place respectively. Whereas gain of function (bio_eff (G))

and loss of function (biol_eff (L)) were the most prominent features in the biological response/biological effect set (Figure 5-3A). Interestingly, upon the combination of the four feature sets, allele frequency remained the most informative features in the ML reclassification of the VUS (Figure 5-3B). In addition, missense variation that was less prominent among the mutation_type set, is more informative upon the combination of all feature sets. Among the top ten combined features, biological response feature sets, biological were cumulatively the least informative features in the importance distribution in our model (Figure 5-2B)

5.4.2 Examination of Intercorrelation pairwise relationship among the prediction scores, mutation type, sequence sample features and biological response feature sets.

A key rationale behind consideration of the intercorrelation among the features sets is to examine the pairwise relationship among the predictors to identify and evaluate if there is codependency among the prediction scores, mutation type, sequence sample features, and biological features sets. Multicollinearity introduces bias and variance in a model and thereby compromising the predictive power of a model. To check for feature codependency, we computed and analyzed pairwise correlations between the feature sets with Pearson's pairwise correlation coefficient (Figure 5-4). From our findings, SIFT had an inverse relationship with MutationAssessor and PolyPhen scores -0.62 and -0.54. However, PolyPhen had a high positive correlation with mutation assessor 0.48 (Figure 5-4). This confirms the complementary nature of the mutation scores in predicting cancer pathogenicity. Overall, the correlation coefficient matrix indicated that the combination of these features is orthogonal and therefore not prone to model

overfitting



A

B

Figure 5-2: Feature importance analysis of xgboost ML model on cancer-linked dataset of genomic variation showing the most informative features in VUS reclassification. **(A)** Feature importance of prediction scores, mutation type, sequence sample and biological response feature sets showing top-five most informative features. **(B)** Feature importance of all four feature sets combination showing top ten most informative features

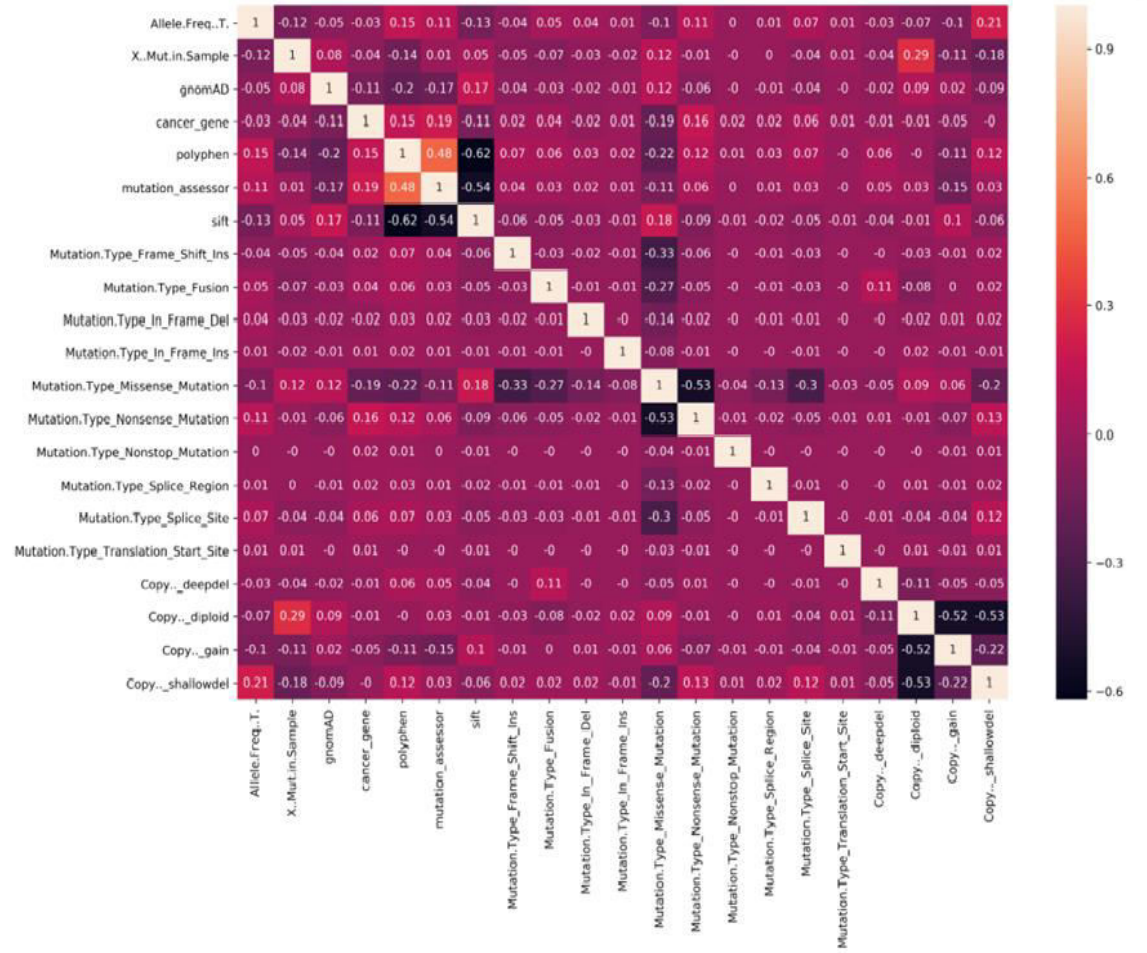


Figure 5-3. Pearson's pairwise correlation coefficients between the prediction scores, mutation type, sequence sample and biological response features sets.

5.4.3 Comparative Analysis of Machine Learning Reclassification of Clinical Implication of Variants of Uncertain Significance

We evaluated and compared the predictive performance of xgboost, SVM and D-NN/ MLP models using sensitivity(recall), specificity, precision and f-score metrics (Table 5-3).

The table showed that the xgboost model outperformed the SVM and D-NN/MLP models (f-score = 0.92) in the reclassification of VUS clinical implication. Xgboost had a sensitivity of 0.9177 and a precision of 0.9311.

Table 5-3. Comparative evaluation of classifiers for VUS clinical implication reclassification

Metric	Classifiers		
	SVM-POLY	Xgboost	Neural Network
Recall (Sensitivity)	0.8365	0.9177	0.8399
Precision	0.8411	0.9311	0.8147
F-score	0.8388	0.9244	0.8271

In the xgboost model, we recorded a recall, precision and f-measure scores of 0.9177, 0.9311 and 0.9244 respectively. This can be attributed to the relatively small sample size of the benign and VLB class. On the other hand, xgboost had a weighted true positive rate of 91.77% benign, VLB, VLP, and oncogenic(pathogenic) predictions respectively. In simpler terms, the percentage of VUS mutations that were correctly classified as benign, VLB, VLP and pathogenic while it had a true negative rate (specificity) of 84.25% (VLP) and 97.21%(pathogenic). Xgboost recorded a precision of 0.935 and 0.9272 this means that a VUS was predicted to be likely pathogenic (VLP) and pathogenic the predictions were correct 93.5% and 92.72% of the time respectively. In the SVM-poly model, we had recall and precision scores of 0.8365 and 0.8411 respectively. Of the benign mutations, VLB, VLP and pathogenic, 83.65% were classified as benign, VLP and pathogenic. Of the mutations classified as benign, VLP and pathogenic, 84.11% of them were correctly classified. In the deep neural network model, of the benign mutations, VLB, VLP, and MUT 83.99% of them were classified as benign, VLB, VLP and MUT while of the mutations classified as benign, VLB, VLP, and MUT 81.47% of them were correctly classified. Overall, the xgboost model performed best with an F-score of 0.9244 when compared with the SVM-poly and D-NN models.

In the binary classification model, the xgboost classifier (with AUC: 0.7844) outperformed SVM (AUC: 0.7365) and D-NN (AUC: 0.7404) classifiers at reclassifying VUS(Figure 5-5).

Evaluating the multiclass and binary models, we observed that the classifiers were better at reclassifying VUS based on ACMG 4 point scale (benign, VLB, VLP and pathogenic) when compared with the benign-pathogenic dichotomous class.

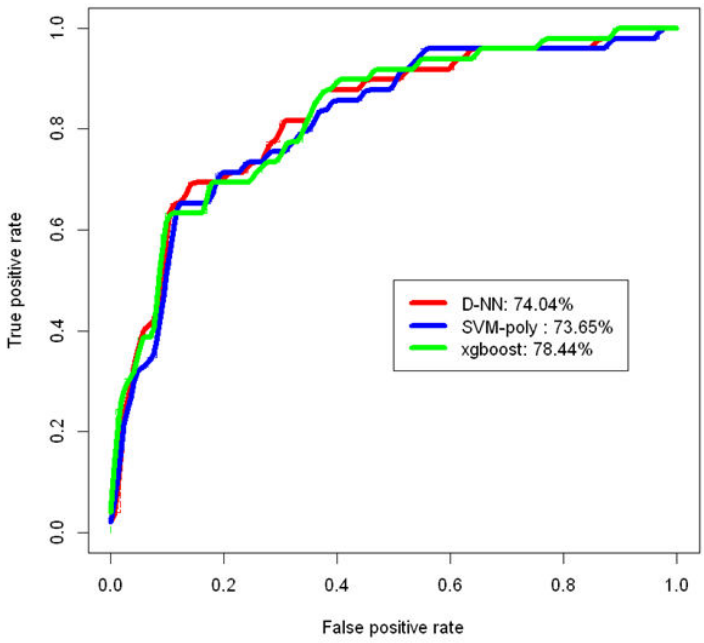


Figure 5-4: ROC/AUC plot for binary reclassification of VUS. Xgboost outperform SVM and D-NN at reclassifying VUS

5.5 Reference

Agajanian, S., Odeyemi, O. and Verkhivker, G. (2019). Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for Classification and Biomolecular Modeling of Cancer Driver Mutations. *Front Mol Biosci.* 2019 Jun 11;6:44. doi: 10.3389/fmolb.2019.00044. eCollection 2019.

Agajanian, S., Odeyemi, O., Bischoff, N., Ratra, S., and Verkhivker, G. M. (2018). Machine learning classification and structure-functional analysis of cancer mutations reveal unique dynamic and network signatures of driver sites in oncogenes and tumor suppressor genes. *J. Chem. Inf. Model*. 58, 2131–2150. doi: 10.1021/acs.jcim.8b00414

Alosi, D., Bisgaard, M., Hemmingsen, S., Krogh, L., Mikkelsen, H. and Binderup, M. (2016). Management of Gene Variants of Unknown Significance: Analysis Method and Risk Assessment of the VHL Mutation p.P81S (c.241C>T). *Current Genomics*, 18(1), pp.93-103.

Andres, S., Williams, K. and Rustgi, A. (2018). The Molecular Basis of Metastatic Colorectal Cancer. *Current Colorectal Cancer Reports*, 14(2), pp.69-79.

Atems, B. and Bergtold, J. (2015). Revisiting the statistical specification of near multicollinearity in the logistic regression model. *Studies in Nonlinear Dynamics & Econometrics*, 0(0).

Bentley, D., Balasubramanian, S., Swerdlow, H., Smith, G., Milton, J. and Brown, C. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, (456), pp.53–59.

Brognard, J., Zhang, Y., Puto, L. and Hunter, T. (2011). Cancer-Associated Loss-of-Function Mutations Implicate DAPK3 as a Tumor-Suppressing Kinase. *Cancer Research*, 71(8), pp.3152-3161.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B., Sumer, S., Aksoy, B., Jacobsen, A., Byrne, C., Heuer, M., Larsson, E., Antipin, Y., Reva, B., Goldberg, A., Sander, C. and Schultz, N. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. *Cancer Discovery*, 2(5), pp.401-404.

Chen, Y., Li, J., Yu, X., Li, S., Zhang, X., Mo, Z. and Hu, Y. (2013). APC gene hypermethylation and prostate cancer: a systematic review and meta-analysis. *European Journal of Human Genetics*, 21(9), pp.929-935.

Collins, F. and Hamburg, M. (2013). First FDA Authorization for Next-Generation Sequencer. *New England Journal of Medicine*, 369(25), pp.2369-2371.

Ding, L., Ley, T., Larson, D., Miller, C., Koboldt, D., Welch, J., Ritchey, J., Young, M., Lamprecht, T., McLellan, M., McMichael, J., Wallis, J., Lu, C., Shen, D., Harris, C., Dooling, D., Fulton, R., Fulton, L., Chen, K., Schmidt, H., Kalicki-Veizer, J., Magrini, V., Cook, L., McGrath, S., Vickery, T., Wendl, M., Heath, S., Watson, M., Link, D., Tomasson, M., Shannon, W., Payton, J., Kulkarni, S., Westervelt, P., Walter, M., Graubert, T., Mardis, E., Wilson, R. and DiPersio, J. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382), pp.506-510.

Foley, S., Rios, J., Mgbemena, V., Robinson, L., Hampel, H., Toland, A., Durham, L. and Ross, T. (2015). Use of Whole Genome Sequencing for Diagnosis and Discovery in the Cancer Genetics Clinic. *EBioMedicine*, 2(1), pp.74-81

Gao, J., Aksoy, B., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C. and Schultz, N. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Science Signaling*, 6(269), pp.pl1-pl1.

Geng, S., Kuang, Z., Liu, J., Wright, S. and Page, D. (2018). Stochastic Learning for Sparse Discrete Markov Random Fields with Controlled Gradient Approximation Error. *Uncertain Artif Intell*, [online] (Aug;2018), pp.155-165. Available at: https://www.ncbi.nlm.nih.gov/pubmed/30555282 [Accessed 13 Oct. 2019].

Green, E. and Guyer, M. (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470(7333), pp.204-213.

He, K., Zhao, Y., McPherson, E., Li, Q., Xia, F., Weng, C., Wang, K. and He, M. (2016). Pathogenic Mutations in Cancer-Predisposing Genes: A Survey of 300 Patients with Whole-Genome Sequencing and Lifetime Electronic Health Records. *PLOS ONE*, 11(12), p.e0167847.

Kroll, C. and Song, P. (2013). Impact of multicollinearity on small sample hydrologic regression models. *Water Resources Research*, 49(6), pp.3756-3769.

Landrum, M., Lee, J., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J., Kattman, B. and Maglott, D. (2017). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1), pp.D1062-D1067.

Levine, A., Jenkins, N. and Copeland, N. (2019). The Roles of Initiating Truncal Mutations in Human Cancers: The Order of Mutations and Tumor Cell Type Matters. *Cancer Cell*, 35(1), pp.10-15.

Li, M., Datto, M., Duncavage, E., Kulkarni, S., Lindeman, N., Roy, S., Tsimberidou, A., Vnencak-Jones, C., Wolff, D., Younes, A. and Nikiforova, M. (2017). Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer. *The Journal of Molecular Diagnostics*, 19(1), pp.4-23.

Lindor, N., Guidugli, L., Wang, X., Vallée, M., Monteiro, A., Tavtigian, S., Goldgar, D. and Couch, F. (2012). A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). *Human Mutation*, 33(5), pp.900-903.

Lodish, H. (2000). *Molecular cell biology*. New York: W.H. Freeman.

Lu, W., Cheng, Y., Xiao, C., Chang, S., Huang, S., Liang, B. and Huang, T. (2017). Unsupervised Sequential Outlier Detection with Deep Architectures. *IEEE Transactions on Image Processing*, 26(9), pp.4321-4330.

Mair, B., Konopka, T., Kerzendorfer, C., Sleiman, K., Salic, S., Serra, V., Muellner, M., Theodorou, V. and Nijman, S. (2016). Gain- and Loss-of-Function Mutations in the Breast Cancer Gene GATA3 Result in Differential Drug Sensitivity. *PLOS Genetics*, 12(9), p.e1006279.

Markowitz, S. and Bertagnolli, M. (2010). Molecular Basis of Colorectal Cancer. *New England Journal of Medicine*, 362(13), pp.1245-1247.

Mucaki, E., Caminsky, N., Perri, A., Lu, R., Laederach, A., Halvorsen, M., Knoll, J. and Rogan, P. (2016). A unified analytic framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer. *BMC Medical Genomics*, 9(1).

Oulas, A., Minadakis, G., Zachariou, M. and Spyrou, G. (2019). Selecting variants of unknown significance through network-based gene-association significantly improves risk prediction for disease-control cohorts. *Scientific Reports*, 9(1).

P. Vatcheva, K. and Lee, M. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology: Open Access*, 06(02).

Rutkowski, J. (2000). Genetic and cellular defects contributing to benign tumor formation in neurofibromatosis type 1. *Human Molecular Genetics*, 9(7), pp.1059-1066.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10), pp.1135-1145.

Wang, G., Chen, L., Yu, B., Zellmer, L., Xu, N. and Liao, D. (2016). Learning about the Importance of Mutation Prevention from Curable Cancers and Benign Tumors. *Journal of Cancer*, 7(4), pp.436-445.

Yorczyk, A., Robinson, L. and Ross, T. (2014). Use of panel tests in place of single gene tests in the cancer genetics clinic. *Clinical Genetics*, 88(3), pp.278-282.

6 Conclusion

Identification and characterization of cancer driving mutations remain a challenge in understanding tumor progression. Several bioinformatic strategies have been developed to examine the functional effects of variants that contribute to cancer. The incorporation of biological context in predictive models means more robust and biologically relevant predictions.

In this dissertation, we leveraged ML predictions to (i) identify and classify cancer-linked mutations, (ii) reclassify the clinical significance value of VUS and (iii) predict survival status of selected unresectable cancers.

In the ML classification and functional analysis of cancer mutations study (Chapter 2), we compared two cancer-specific machine learning classifiers (Logit vs. RF) for prediction of driver mutations in cancer-associated genes that were validated on canonical data sets of functionally validated mutations and applied to a large cancer genomics data set. We have found that evolutionary-based features to be the most informative in machine learning predictions and provide orthogonal information to the ensemble-based scores that dominate feature importance ranking. By utilizing a comparative analysis with various structure–functional experiments and multicenter mutational calling data from Pan-Cancer Atlas studies, we demonstrated that the robustness of machine learning models that capture ~90% of experimentally validated mutational hotspots among predicted driver mutations. The results of this study have also indicated that although the range of predicted driver mutations may be broader than the experimentally validated group of drivers, some of the predicted "false positive" variants can present viable candidates for further experimental testing and validation.

In the integration of RF and D-CNN for classification of cancer driver mutations study (chapter 3), we demonstrated that various ML approaches for prediction cancer driver mutations. We explored the D-CNN model's capability of classifying cancer driver mutations directly from raw nucleotide sequence information without depending on functional prediction and evolutionary scores. The results further demonstrate that though D-CNN models can learn high-level features from genomic information that has sufficiently high importance, accurate classification of cancer mutation driver phenotype using exclusively nucleotide data is not a practical approach.

In the survival status study (chapter 4), we used an xgboost model to predict the survival status of five cancer studies and found that genomic features (molecular subtypes and ploidy) and clinical features (age at diagnosis, race, sex) are critical components of survival status classification. In the feature importance ranking analysis, age at diagnosis was a dominant feature in ALL, COAD, GLIO and P-NB cancer studies. In the ALL Cox model, age at diagnosis, cell tumor origin (T and B cells), TCF3_PBX1_status(negative), congenital abnormality, MS= Hyperdiploidy without trisomy of both chromosomes 4 and 10 are statistically significant features. Likewise, the INSS stage and ploidy are key predictors for examining the survival rate of pediatric neuroblastoma. The results of this study suggest the further exploration of genomic features will help in the understanding of survival indicators of cancer diseases.

In the VUS reclassification study (chapter 5), we have developed an intuitive machine learning-based approach to classifying the clinical significance value of VUS. The major findings of the present study are as follows (i) xgboost classifier does a decent job at reclassifying clinical significance value of VUS with a precision score of 0.9177, recall score of 0.9311 and f-measure score of 0.9244 (ii) based on the analysis of our result, the three classifiers

reclassified most of the VUS as VLP,(iii) the classifiers are better at clinical significance discrimination for ACMG four-tiered scheme than the binary models and (iv) mutation scores and allele frequency are the most informative features need for VUS reclassification. Genomic data remains the deciding factor in the clinical significance value of a variant -the level of clinical evidence. To gather sufficient data, a collaboration between academia and industry collaboration has to be fostered. Machine learning in combination with clinical information about a patient (phenotype information and familial medical history) can help genetic counselors and clinicians in understanding clinical interpretation of genetically related cancers.

Contribution of This Work to Cancer Studies

Previous works use clinical indicators in the diagnosis of cancer. However, we used the combination of clinical and genomic features in the prediction and classification of cancer-driving mutations, VUS reclassification, and cancer survival classification. To evaluate the most informative features in the cancer classifications mentioned earlier, we ranked features based on their contribution to the models. This work has contributed to the identification and characterization of cancer-driving genes, which are vital for early cancer detection. We have demonstrated the robustness of the ML approach in the prediction and classification of cancer-driving mutations based on functional, evolutionary conservation, and integrated ensemble scores. Our ML approach has highlighted the role of SNVs in cancer diagnosis. Additionally, integration of ML approach with other computational strategies may improve the survival outcomes of unresectable cancers.