Considerations for a Machine Learning Approach to Classification of Cancer Driver Mutations

Considerations for a Machine Learning Approach to Classification of Cancer Driver Mutations

Daniel Robert Smith

A Thesis in the Field of Bioinformatics

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

May 2022

Abstract

Cancer is one of the leading causes of death for people worldwide. Since the completion of the Human Genome Project, Next-Generation Sequencing has made leaps in understanding of the cancer genome possible. Such a deep understanding has allowed researchers to develop novel targeted therapy options and improve survival rates. As the amount of complex genomic data increases, powerful tools are necessary to discern underlying genomic drivers and therapeutic targets in a patient's cancer. Machine learning has been an asset in the discovery of new relationships in cancer genomes and is explored in this research. Using publicly available genomic data from several databases, machine learning models were designed and implemented to classify variants as pathogenic or benign in APC, RB1, TP53, EGFR, ERBB2, and PIK3CA genes, all previously implicated in various cancers. The output of the classification experiments demonstrates the utility of random forest and extremely randomized trees classifiers and highlights the value of several key data features across these datasets. In addition, the implementations offer guidelines for future researchers by emphasizing reproducibility and generalizability of similar models. Through this framework, future machine learning research may be faster to implement using real-world data. By leveraging the power of machine learning, scientists can continue to expand the cancer genomics knowledgebase and take steps toward improved outcomes for patients.

Dedication

To my parents Mary Jo and Dave Smith for encouraging me and cheering me on from childhood through adulthood and beyond; To my fiancée Rachel for your support and reassurance throughout all my academic, professional, and personal pursuits; To my siblings Emily, Meghan, and Michael for your constant inspiration and laughter.

Acknowledgments

The completion of this research would not have been possible without the expertise of my advisor Dr. Garrett Frampton. Dr. Frampton's pioneering work in the field of cancer genomics research and his emphasis on producing quality work that I could be proud of helped to elevate my abilities as a writer, student, and researcher.

I thank James Pao and Nathanial Eddy for sharing their perspectives on machine learning performance evaluation and considerations for real-world applications of machine learning.

I am thankful for my professors and my academic advisory team at Harvard Extension School, especially Dr. Steven Denkin, Gail Dourian, Joan Short, and Trudi Goldberg Pires. This team was instrumental to my success as a student in this program and during this graduate research process.

I also thank my supervisor Caitlin Patriquin, who has been a mentor and friend throughout my professional and academic journey, having been a Harvard student herself. Through Caitlin I learned not only about this program, but important lessons in professionalism, leadership, and perseverance.

Finally, I thank my colleague and friend Dean Pavlick, who has been a motivator and source of inspiration to me for many years, not only throughout this thesis process, but for many of my academic and professional pursuits.

Table of Contents

List of Tables

List of Figures

Chapter I.

Introduction

The family of diseases known collectively as cancer have been a central focus of medical research for much of recorded history and continue to drive medical innovation today (Di Lonardo et al., 2015). A molecular understanding of cancer began in 1914 when Theodor Boveri proposed the theory that chromosomal abnormalities played an essential role in tumor development (Di Lonardo et al., 2015). In the 1970s and 1980s, scientists identified two important families of genes: oncogenes and tumor suppressors (Berry et al., 2019; Bister, 2015). In normal cells, these two classes of genes work in harmony, to initiate or inhibit cell growth, replication, rest, repair, and death at key points in the cell's lifecycle (Lee & Muller, 2010). Genomic mutations affecting oncogenes and tumor suppressors confer unique survival advantages to emerging cancer cells, allowing the disease to attain qualities central to its survival (Hanahan & Weinberg, 2011). The discovery of oncogenes and tumor suppressors further supported the hypothesis that cancer is a disease of the genome (Macconaill & Garraway, 2010). Later research would certify this hypothesis via study of HRAS genes in cancerous bladder tissue and in normal bladder cells, adding to a growing body of evidence demonstrating that mutated genes can enable cancer cell growth (Macconaill & Garraway, 2010).

The recognition of cancer as a molecular disease had profound implications for the study of cancer's survival mechanisms. As a disease of the genome, mutations in various genes can produce mutated proteins and enable tumor growth via acquisition of several key disease hallmarks (Hanahan & Weinberg, 2011). These hallmarks of cancer

include sustained proliferative signaling, evading growth suppressors, activating invasion and metastasis, enabling replicative immortality, inducing angiogenesis, and resisting cell death (Hanahan & Weinberg, 2011). Collectively, genomic mutations trigger changes to cellular behavior and capability, facilitating rapid expansion of cancerous cells (Hanahan & Weinberg, 2011).

Not all mutations in the genome are inherently pathogenic. Scientists estimate that about 1.1% of the human genome encodes functioning proteins, while the purpose of the remaining non-translated regions is less well understood (Ponting & Hardison, 2011). Beyond the scarcity of translated regions in the genome, acquisition of genomic mutations is also rare. The normal mutation rate can be approximated as 2.5×10^{-8} mutations per nucleotide site or about 175 mutations per diploid genome, per generation (Nachman & Crowell, 2000). In contrast, cancer genomes demonstrate both increased potential for mutation and more rapid positive selection for mutations conferring survival advantage (Temko et al., 2018). Reduction of cellular DNA repair capabilities may contribute in part to the elevated number of mutations acquired in cancer cell lineages (Hanahan & Weinberg, 2011).

Between 2011 and 2018, cancer disease was the 2nd highest cause of death worldwide (Rana et al., 2021). Many factors contribute to cancer's high rate of mortality. First, cancer is a diverse family of diseases with different origins and progression. Many cancers are classified based on the type of tissue they affect, however numerous unique subtypes have been described, each with various possible stages, progression patterns, and mortality trends (Carbone, 2020). Second, the genomic mechanisms of cancer progression are complex and difficult to generalize (Martin & Santaguida, 2020). Third,

there are an insufficient number of targeted treatment options available for cancer patients. Through the discovery that cancer is a disease of the genome, research which emphasizes discovery and characterization of cancer-relevant mutations will have the greatest opportunity to improve treatment options and outcomes for patients in a diverse set of cancer subtypes. The association of cancer mutations to changes in cellular behavior will enable clinical intervention at the molecular level, inhibiting the cells' ability to acquire disease hallmarks (Hanahan & Weinberg, 2011).

Sequencing Technology and Treatment Opportunities

Genomic sequencing has become an invaluable tool in the research and discovery of novel cancer mutations. Modern sequencing technology originated in 1977 and is credited to Frederick Sanger, who developed a chain termination method to sequence genetic code (Heather & Chain, 2016). Next-Generation Sequencing, or NGS, is a revolutionary tool for present day genomic research (Shyr & Liu, 2013). Commercially developed in 2004, NGS is defined by a high-throughput, massively parallel architecture, allowing researchers and clinicians to probe test specimens for comprehensive genomic profiling (Kamps et al., 2017).

The cost associated with modern NGS sequencing technology has decreased significantly in recent years. The Human Genome Project incurred an estimated cost of $3 billion USD and required 13 years to complete (Sboner et al., 2011). By 2019, the cost of sequencing a human genome had dropped to about $1,000 USD (NHGRI, 2019). The increased accessibility and affordability of sequencing technology has led to substantial growth in industry revenue overall. According to Phillips & Douglas (2018) the global

diagnostic NGS oncology market will realize $7.7 billion USD in revenue as of 2020, and the market is expected to experience continued growth.

The reduced cost of sequencing has enabled many different applications for the technology, including the study of cancer genomes. Sequencing of cancer genomes confers significant value in both diagnostic and prognostic applications (Mardis & Wilson, 2009), especially in the characterization of disease hallmarks (Hanahan & Weinberg, 2011). Today, the affordability of NGS sequencing enables individual patients to sequence their cancer genome to identify targeted therapy opportunities for treatment (Saito et al., 2018).

The growing accessibility of NGS technology throughout the 21st century has allowed investigation of the cancer genome at unprecedented scale. Such research has produced large databases, many of which are available to the public. The Cancer Genome Atlas (TCGA) is one such dataset, comprised of more than 20,000 different primary cancer and matched normal samples, together covering 33 different cancer types and more than 2.5 petabytes of data (NCI, 2019). The Catalogue of Somatic Mutations In Cancer (COSMIC) is another database experiencing substantial growth. Today, COSMIC comprises over 6 million coding mutations spanning more than 1.4 million tumor samples. Notably, COSMIC incorporates various publicly available data from both TCGA and the International Cancer Consortium (ICGC) (Tate et al., 2018). The growth of public repositories such as TCGA, COSMIC, and others is reflective of broader cancer NGS market trends. Collectively, these projects allow clinical utility for cancer genomics and diagnostics.

The effort to determine functional significance of genomic mutations is at the heart of modern cancer research and treatment advances. A functional variant is some mutation which leads to a change at the molecular level of a protein (Gonzalez-Perez et al., 2013). This change can be classified by a gain, loss, or change in function of the protein compared to wild-type (Gonzalez-Perez et al., 2013). Conversely, many mutations may be considered non-functional. These benign alterations can include germline mutations, passenger mutations, or mutations in non-coding regions of the genome (Stratton et al., 2009). Identifying a key genomic driver in a patient population can often lead to targeted therapy development, increased survival rates, and improved quality of life. There are many approaches to determine functional significance, and over time these methodologies have become more robust and complex. A promising approach to mutation classification involves the use of an artificial intelligence method known as machine learning.

Machine Learning and Cancer

Machine learning is a collection of different mathematical algorithms which can determine the relationship between features or characteristics of underlying data (Edwards, 2020). This capability is well suited for discovery of novel relationships between data and subsequent classification or prediction of future data against an existing model (Edwards, 2020). Machine learning has quickly become an attractive option for cancer genomics researchers, in part due to the complexity of highly dimensional genomics data.

Relevant genomic data features may include descriptions of sample type, the tissue of origin, the sample size, and other pathological characteristics. The specific

datapoints describing the observed mutation are recorded in detail, involving complex nomenclature to describe the change relative to the genome, the chromosome, the gene, the exon, and the nucleotide sequence itself. Other observations and contextual information can quickly expand the dimensionality of each entry in the dataset. Machine learning can navigate high dimensionality and peel away uninformative features to elucidate unseen relationships between key datapoints (Sidey-Gibbons & Sidey-Gibbons, 2019). The algorithms used in machine learning leverage a process known as training which uses real data to help the algorithm identify and learn about relationships (Edwards, 2020). As new data is interpreted, the model can improve its generalization (Edwards, 2020). This research will employ machine learning to explore best practices and key considerations in approaches to cancer mutation classification problems. Several computational methodologies have already been developed to assess driver status of mutations and a subset will be described herein: CanPredict, CHASM, and Cerebro.

CanPredict Classifier

The CanPredict method is a supervised machine learning algorithm developed by Kaminker et al. in 2007. The algorithm is based on the belief that cancer genome instability coupled with increased cellular divisions causes an explosion of passenger mutations which may obfuscate true somatic drivers (Kaminker et al., 2007). The researchers utilize a random forest algorithm to distinguish between mutations from the COSMIC database that are suspected cancer drivers versus non-synonymous single nucleotide polymorphisms (nsSNPs) with high mutant allele frequencies (MAFs) (Kaminker et al., 2007). To build their algorithm, the researchers gathered: (a) common variants from NCBI, which included overall minor allele frequencies; (b) cancer-

associated mutations, which were collected from the COSMIC database – further filtered based on analysis by Forbes et al. (2006) to include only variants likely to be involved in oncogenesis; (c) Mendelian disease-associated variants from Swiss-Prot (Kaminker et al., 2007). The researchers then segregated missense mutations and began by looking to the SIFT and LogR.E-value algorithms, two independent algorithms to predict the tolerability of different mutations and score protein products based on their distinction from wild type (Kaminker et al., 2007). The CanPredict random forest was built to incorporate output from SIFT, Pfam-based LogR.E-values, and GO (Gene Ontology) log-odds scores as part of the feature set for each variant (Kaminker et al., 2007). The research showed that many activating mutations impair protein function, positing that variants in kinases appear to affect amino acids involved in the control of enzymatic activity (Kaminker et al., 2007). The researchers also note several limitations, including their use of expressed sequence tags, or ESTs in the discovery phase for variant identification. Coverage and library bias were prevalent in the ESTs, leading to the exclusion of some well characterized somatic cancer drivers, such as BRAF V600E (Kaminker et al., 2007).

CHASM Classifier

CHASM is an alternate machine learning classification tool developed by Carter et al. in 2009. The CHASM classifier is built upon on the idea that methodologies which classify mutations must not be wholly dependent on mutation frequency, and their research thus comes as an expansion on the CanPredict method (Carter et al., 2009). The researchers hypothesize that a classifier can be trained with improved specificity if passenger mutations are represented with *in silico* simulations, incorporating mutation profiles reflective of tumor type and context (Carter et al., 2009). This model expanded

upon the CanPredict method by looking beyond high-MAF nsSNPs and made the classification of passenger mutations more complex. The researchers similarly trained a random forest classifier using 2,488 missense mutations previously identified as playing a functional role in oncogenesis from the COSMIC database, with samples sourced from various breast, colorectal, and pancreatic tumor studies (Carter et al., 2009). To assess their classifier, the researchers used two threshold-independent measures – Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves. Of note, the CHASM classifier identified several dominant features, such as SNP density, frequency of missense change type in COSMIC, and nucleotide-level conservation (Carter et al., 2009). The researchers found that there is a potential difference in distinguishing characteristics of neutral mutations in the cancer genome and the germline genome, and their CHASM classifier works to recognize the former, offering improved performance over prior methods (Carter et al., 2009). This work also highlights the importance of a null model, which in this research represents key assumptions about the nature of benign variants (Carter et al., 2009). At the time of publication, this research was perhaps the first to identify candidate driver mutations via control over false discovery rate (FDR) (Carter et al., 2009). By utilizing this method of FDR control, the researchers were able to achieve improved power in their model (Benjamini et al., 2001).

Cerebro Classifier

Another machine learning analysis performed by Wood et al. (2018) utilized data from The Cancer Genome Atlas (TCGA). Their machine learning implementation involved an extremely randomized trees classification model called Cerebro (Wood et al., 2018). The algorithm selected for dominant features such as allele frequency, nearby

sequence complexity, and presence of alterations in the matched normal specimen. The implementation of Cerebro was not well trained for tumors with low cellularity and low-frequency alterations, a shortcoming that the researchers predict could be alleviated with a more robust training set (Wood et al., 2018). Of note, the researchers found that Cerebro improved clinical outcomes for both melanoma and lung cancer patients when compared to a tumor mutation load assessment alone (Wood et al., 2018). Concordance calculated between paired tumor-normal exome data from 1,368 TCGA samples was only 74%, elucidating potential false positives in the core dataset (Wood et al., 2018). The Cerebro team's results identified an inaccuracy rate in TCGA datasets of about 16% which represents nearly 500,000 mutations incorrectly classified based on their modeling (Wood et al., 2018).

Limitations to Current Methods

Collectively, these machine learning implementations and others offer attractive options to researchers looking to discover and classify key biomarkers in the cancer genome. However, none of these approaches are without shortcomings. One shortcoming is the challenge of defining a universal set of important features that any successful machine learning implementation should contain. In some cases, feature selection may be directly influenced by the underlying data and the nature of each record in the dataset. In other situations, researchers may custom-fit features which are difficult to reproduce and may be dependent on proprietary methods.

Universal Features

To address this shortcoming, this work will define several suggested universal features which can be reproduced in future machine learning implementations. There are a growing number of valuable datasets describing cancer data which enable robust feature creation, such as COSMIC, dbSNP, Pfam, and others. The accessibility of this data will allow more rapid integration of independent sources to train intelligent machine learning models, a practice which will be performed in this research.

One such universal feature will be the inclusion mutational co-occurrence, which was not explicitly denoted in prior research studied, and which must be accounted for where possible (Temko et al., 2018). A subset of the reviewed studies considered the proximity of mutations to key conserved functional domains (Carter et al., 2009), which this research will also address. By defining a set of recommended universal features using data available to the public, faster implementation across different datasets and reduced risk of overfitting may be achieved.

Algorithm Selection

In addition to feature selection and reproducibility, a broader challenge of algorithm selection exists for researchers. Many mutation classification and discovery pipelines in the reviewed literature utilized a random forest algorithm, a classification algorithm which is often highly accurate and able to minimize error (Yiu, 2019). However, other machine learning algorithms may be viable alternatives in cancer genomics classification problems. The extremely randomized trees or extra-trees algorithm extends the randomness of a random forest classifier to further minimize variance (Geurts et al., 2006; Pedregosa et al., 2011). The adaboost classifier is another

popular ensemble-based machine learning algorithm. The adaboost algorithm utilizes continued boosting of so-called weak learners to vote on final prediction and classification, an alternative to the bagging approach employed by random forest and extra-trees classification (Freund & Schapire, 1997; Pedregosa et al., 2011).

This research will also explore the different performance characteristics of these classification algorithms. Beyond a comparison of performance outcomes, the key considerations for model design, dataset structure and training input will be evaluated. By comparing both the capabilities and relevant features of these models, this research will present important considerations and future direction for cancer driver classification using machine learning.

Chapter II.

Research Methods

This work explores considerations for *in silico* analysis of mutations in cancer. Throughout the research process, reproducibility and generalizability of the data was emphasized as a primary measure of success of the methods, in addition to key performance indicators following implementation. All data, software, and analyses were performed using public and open-source toolkits and genomic data. Where necessary, custom data features are described in greater detail to facilitate reproducibility.

This research was performed using the Python programming language and utilized a combination of Python libraries to achieve the desired machine learning outcomes. In addition to the base Python functionality, the Scikit-learn toolkit was utilized and provided the basis for each machine learning algorithm explored (Pedregosa et al., 2011). Supplemental Python libraries were incorporated for simplified data manipulation and to extend base capabilities, including Pandas (Team, 2020), NumPy (Harris et al., 2020), and Imbalanced-Learn (Lemaitre et al., 2017). The Matplotlib library was also used for data visualization (Hunter, 2007).

Collecting Genomic Data

The primary data source for this research was genomic data collected from The Catalogue of Somatic Mutations In Cancer, or COSMIC (Tate et al., 2018). Genomic information is taken directly from a filtered subset of COSMIC's mutation data file from

database release v94. For preliminary data aggregation, filters were applied on gene name alone. Due to the large number of genes implicated in cancer development, this research was focused on a selection of genes with some prior characterization in cancer. The utility of various cancer variant classifier implementations was tested on both oncogenes and tumor suppressors, including EGFR, PIK3CA, and ERBB2 (oncogenes) and TP53, APC, and RB1 (tumor suppressors). In addition to COSMIC mutation data collection, other sources of supplemental data were gathered from COSMIC, including FASTA files for both the coding sequence (CDS) and protein (amino acid) sequence for the genes studied, as well as COSMIC mutational signatures (Alexandrov et al., 2020).

Protein Domains

Protein domains are key regions in the protein sequence which can contribute both structurally and functionally to the encoded protein (Buljan & Bateman, 2009). Preservation of some domains may be essential to producing a functioning product and may be detrimental to cancer development, while mutations in other domains can have the opposite effect, removing important regulatory regions and conferring a survival advantage to cells expressing the mutated allele (Miller et al., 2015). Thus, the localization of mutations in the context of protein domains is predicted to be an important feature in the construction of any viable machine learning classification algorithm.

To describe mutations relative to protein domains, data from Pfam was incorporated into the COSMIC dataset for feature creation. Pfam is a resource dedicated to the collection and annotation of protein families and functional domain information. Such data is represented by a combination of multiple sequence alignments (MSAs) and hidden Markov models (HMMs) (Mistry et al., 2020). The integration of domain data

with these classifiers was performed to enrich the feature sets via additional datapoints of interest.

Single-Nucleotide Polymorphisms

Single-nucleotide polymorphisms (SNPs) are a type of heritable DNA mutation, estimated to occur at a frequency of about 1 in 1,000 base pairs throughout the genome (Shastry, 2009). While some SNPs are silent (synonymous or affecting non-coding regions), others may change amino acid composition and confer uniqueness unto an individual genome (Shastry, 2009). A causal relationship may exist between some SNPs and tumorigenesis based on prior research, but the extent of such a relationship more generally is yet unclear (Deng, et al., 2017). Denoting potential SNPs for the COSMIC entries studied is predicted to improve the quality of the feature sets for each classifier implementation.

Although COSMIC places emphasis on somatic mutations in cancer, information about single nucleotide polymorphisms (SNPs) is integrated into the default COSMIC mutation data. Discrepancies were observed in SNP designation between COSMIC and dbSNP for some mutations following a manual comparison of a random sampling of the data. To further enrich the information for labeled SNPs and ensure accurate representation, comprehensive SNP records were collected from dbSNP (Sherry et al., 2001). Data including classification of the SNP, as well as the validation method and a prediction of clinical significance were appended to the existing COSMIC dataset (Kitts, 2011; Sherry et al., 2001). The clinical significance attribute includes assertions made by the submitter, except where data from OMIM were utilized in assessment (Sherry et al., 2001).

Mutational Signatures

Genomic mutations are often attributable to various agents which damage DNA, such as carcinogens (Barnes et al., 2018). Through both exogenous and endogenous processes, different carcinogens and reactive species can influence DNA damage in cells, often triggering or accelerating cancer development (Barnes et al., 2018). Mutational signatures represent emergent patterns in mutated genomic data that are often directly linked to a specific agonist or phenotype and frequently observed in cancer genomes (Alexandrov et al., 2020). Signatures can be used to estimate a patient's age (SBS1), their tobacco smoking history (SBS4), and their exposure to ultraviolet light (SBS7a-d) (Alexandrov et al., 2020), all of which can influence the acquisition of new mutations in the genome. Several signatures point to specific breakdowns in cellular processes, such as a defective DNA mismatch repair process, while others may show the impact of prior treatment (Alexandrov et al., 2020).

COSMIC mutation signatures are captured as Single Base Substitutions (SBS), and 96 signature contexts are classified as of COSMIC Signatures v3.2, published in March 2021. The signatures are denoted by a reference and alternate nucleotide for a mutated base, flanked by the nucleotide immediately 5' and 3' in the coding sequence of the gene (Alexandrov et al., 2020). For each signature, prevalence of every combination of this 5'-[Ref/Alt]-3' sequence is computed based on data from SigProfiler, extracted from 2,780 whole-genome variant calls (Alexandrov et al., 2020; Bergstrom et al., 2020; Tate et al., 2018). For this research, a contribution fraction for each 3-nucleotide mutation context to a signature was gathered and compared against the raw COSMIC mutation data gathered for each gene studied (Tate et al., 2018). Each single nucleotide mutation

was extracted along with the 5' and 3' flanking bases, and the fractional contribution of that mutational context to each of the 96 signatures was recorded and added to the feature set for the mutation data (Alexandrov et al., 2020; Tate et al., 2018). Several COSMIC signatures listed as possible sequencing artifacts were excluded from the feature set as detection of sequencing artifacts is not a focus of this research.

Defining Custom Features

To improve the overall performance of the classifiers, custom features were designed and implemented. Design of custom features was deliberate to capture information predicted to be meaningful to any classification algorithm. Beyond COSMIC mutation data, several other descriptors were incorporated for each mutation in each of the genes studied.

Domain Mapping

Data from Pfam provides a mapping of functional domains onto the amino acid sequence of a protein (Mistry et al., 2020). This domain information was downloaded directly from the Pfam database. Using a combination of FASTA amino acid sequence, mutation annotation, and functional domain information, a domain mapping feature was developed. The novel feature examined whether a mutation is contained within a specific functional domain, the name of which was recorded. If a mutation was not contained within a defined functional domain, the closest characterized domain was recorded instead. Using the middle amino acid position (for mutations spanning multiple residues), the distance to the nearest characterized functional domain was calculated. This domain-based feature is predicted to elucidate genomic mutation hotspots within a specific gene

based on proximity to key functional regions within the gene. Beyond genomic coordinates, which are a default COSMIC dataset feature, such information may be helpful in extrapolating the model to other genes with shared superfamilies without needing to fully re-train the model.

Mutation Co-occurrence

For each gene-specific model generated, the mutational burden of tumor samples was measured and incorporated into the overall feature set. Mutational burden or co-occurrence in this research was defined simply as the total count of mutation records in COSMIC for each specific tumor by identifier in the dataset. Mutational burden can vary greatly between different tumor types and can be used clinically as an indicator of tumor genome instability, making some high-burden tumors excellent candidates for immunotherapy treatments (Sha et al., 2020). In this research, overall mutation co-occurrence from each tumor was used to inform pathogenicity prediction for each mutation record. In highly mutated tumors, it may be challenging to ascertain the true drivers of the tumor due to the accumulation of tumor-specific somatic alterations. While these accumulated alterations do often coincide with overall genome instability (Yao & Dai, 2014), they may be misclassified as pathogenic without the contextual information of mutational co-occurrence.

Exploration of the complete COSMIC mutation data containing all mutation records showed substantial variation in the number of co-occurring mutations across all recorded tumors. To reduce the degrees of freedom of this feature category, uniform bins were generated based on the distribution of mutation co-occurrence between each COSMIC tumor. Each mutation entry was then categorized into one of the resulting bins

describing a range of mutational burden to characterize a mutation's uniqueness in the context of the tested specimen.

Mutation Count

While mutational co-occurrence helps to assess tumor burden, a measure of mutation count was also included as an additional feature calculated from the dataset. In this research, mutation count represented the number of entries of a single mutation by protein effect in the COSMIC database across all tumors. Mutations which confer unique survival advantages to tumors will be inherently more likely to appear in different patients within the same cancer population (Vogelstein et al., 2013). Thus, a measurement of how many COSMIC records describe the same mutation is another feature included in the set. A similar system was used to produce several bins describing a range of the number of duplicate entries, allowing each mutation record to be categorized according to the number of occurrences in the dataset.

Data Normalization and Preparation

Once each feature was incorporated into the core dataset, the data required standardization before building the classifier. Several steps were taken to ensure execution of the classifier and to achieve high performance. The first action was to strip any leading or trailing characters from the feature names and from the data, as visual inspection of COSMIC data identified inconsistent formatting. Next, several features were dropped from the dataset. These dropped features included gene name and accession number, as well as gene CDS length, all of which were identical for each item in the set. Other identifiers were also removed, such as COSMIC mutation ID, sample

name, and tumor ID. This removal was performed after these features had been utilized to map tumor-specific mutation burden via total COSMIC mutation data (Tate et al., 2018).

Feature Encoding

An important consideration in machine learning implementations is feature encoding (Yu et al., 2018). The COSMIC mutation data is described with a combination of numerical and categorical features and modification to the representation of the features can be performed, both to allow the chosen machine learning algorithm to execute, as well as to optimize performance on the dataset. The scikit-learn (sklearn) package is equipped with a Label Encoder, which converts categorical variables into a numerical representation by assigning each a unique identifier (Pedregosa et al., 2011). However, label encoding may not be suitable for all features in a dataset. For categories with moderate to high numbers of different possible values, the numerical encoding led to a skewed weighting of features with a greater number of possible states or degrees of freedom (Shaikh, 2018).

Other scikit-learn packages are available to address this concern. Following label encoding, a One Hot Encoder was used to transform the data. A one hot encoding scheme can be visualized as a dimensional transform on the set of categorical states for a given feature. Rather than representing each state by a number under a single feature, a new feature is appended to the dataset for each possible state (Pedregosa et al., 2011). Now, a Boolean representation can be used for each entry in the dataset based on which of the new features are true or false for that entry, which may alleviate issues introduced by the Label Encoder alone (Shaikh, 2018).

Other features in the dataset required custom representation to improve performance of the classifier. Genomic coordinates represent the unique positions of each nucleotide within a given chromosome sequence (Dunnen & Hong, 2019). Because these positions can take on large values, dividing the gene into different balanced subsets or bins helps reduce the dimensionality of the feature. Genomic start and end positions were averaged for each alteration in the raw COSMIC output and placed into evenly distributed bins based on their position. While this representation may have reduced resolution, it was predicted to make the feature more useful for the model overall by reducing the cardinality or number of possible states.

Defining The Target Class

For each of the classification algorithms tested, a target class was defined. The target class represents the output class or category that each data point was classified into and which the models output when fitting a specific data point (Pedregosa et al., 2011). The models were designed to treat the FATHMM Prediction as the target class for each mutation. FATHMM, or Functional Analysis through Hidden Markov Models, is a methodology to predict the pathogenicity of coding and non-coding variants incorporated into the COSMIC data structure (Shihab et al., 2013). In addition to the categorical FATHMM prediction, COSMIC mutants also include a FATHMM probability score which was excluded from this classification approach. The models were trained to classify datapoints into pathogenic or benign FATHMM target classes. In this work, a mutation with a FATHMM prediction of pathogenic was considered a driver mutation, representing a mutation which would positively influence tumor development and survival. Conversely, mutations with a FATHMM prediction of unknown or benign were

considered non-functional variants which would have no meaningful impact on tumor development or survival. Other mutation classifications such as germline, passenger, or non-coding may be relevant, however these mutation types were not distinguished in this research for simplicity.

Random Forest Classifier

The random forest classifier was the first classifier approach implemented as part of this cancer mutation study. The random forest algorithm uses a consensus method, leveraging multiple estimators and averaging them together, which has the effect of reduced variance compared to any one estimator in most cases (Pedregosa et al., 2011). The feature set was segregated after incorporating new features and dropping features not explicitly used in modeling.

Model Parameter Customization

The implementation of the random forest algorithm was further customized through changes to default execution parameters enabled by the scikit-learn framework (Pedregosa et al., 2011). By varying different input parameters from their default states, the classifier implementation could be perturbed to evaluate impacts on performance. The goal of this customization was to establish a routine evaluation system to implement a model that maximizes performance for the data in question. Such a system is predicted to be necessary for any machine learning application to achieve robust, data-driven results without compromising efficiency.

Several parameters were adjusted for the random forest algorithm while keeping other parameters constant. Once favorable tuning for each parameter of interest was

determined, other parameters were changed from default values and the model was implemented once more in an iterative fashion. Upon the addition of each new custom parameter, the performance was re-evaluated. This process continued repeatedly until optimal combination of parameters was determined and was executed independently for each of the genes studied to further customize the performance of the model.

The first parameter evaluated was the number of estimators, which defaults to n = 100 for the scikit-learn RandomForestClassifier (Pedregosa et al., 2011). The number of estimators, or trees in the forest, was assigned an initial value of n_0 =10, and incremented such that $n_1 = n_0 * 2$ for each new value of N. This doubling of the number of trees in each forest was performed for each unique implementation until the total number of estimators reached n = 320 trees, with a maximum tree count of 400 evaluated for each implementation. The number of trees used can improve the convergence of a random forest system based on the Strong Law of Large Numbers, as described by Breiman in 2001. Breiman notes that while random forests do not overfit as more trees are added, there is a limiting value reached for the overall generalization error for the model (Breiman, 2001). Thus, the number of estimators were varied, the model was iteratively re-generated, and performance evaluated to determine the optimal number of trees based on the dataset and on the quality of the model and performance optimization.

In addition to the number of estimators, the bootstrapping parameter and related out-of-bag score parameter was utilized. The bootstrapping parameter improves the randomness of the trees in the forest by drawing samples with replacement, allowing duplicates (Breiman, 2021; Pedregosa et al., 2011). The scikit-learn random forest implementation performs bootstrapping by default. While the bootstrapping parameter

was enabled, out-of-bag (OOB) samples were also used to aid in calculation of the generalization score. The OOB method relies on some classifier h(x, T_k) built on a bootstrap training set, T_k. An OOB classifier can then be constructed by assessing each y, X in the training data and aggregating the votes from all trees within h(x, T_k) which do not include y, X themselves (Breiman, 2001). With this approach, Breiman (2001) argues that the error rate of the OOB classifier on the training data can act as a measure for generalization error more broadly.

Another parameter which was adjusted for random forest implementations was scikit-learn's criterion field, which represented a tree-specific split assessment (Pedregosa et al., 2011). The random forest classifier can be assessed using the Gini impurity (default method) and an entropy method, also called information gain (Pedregosa et al., 2011). The Gini impurity is a system to measure the overall quality of a split for a given tree in the forest (Nembrini et al., 2018). The entropy or information gain criterion for the random forest model also assesses split quality on a per-tree basis, relying on a logarithmic approach to assess uncertainty (Fan et al., 2011). Both algorithms were assessed in separate implementations of the random forest algorithm to determine overall impact to performance.

Finally, the proportion of test cases to use from the overall training and test subset of the data was modified across a range of values, beginning with a test size of 10% of the data and incrementing at 10% intervals until 90% of the data was segregated into the testing subset. A model built on a smaller training set (large test size) may be insufficient to classify most representative data from the broader dataset, especially if the training set is significantly unbalanced. Conversely, a model with a large training set (small test size)

may be more prone to overfitting. Thus, various train-test splits were evaluated for each random forest classifier generated.

The dataset for each gene was split into X and y subsets, where X represented each of the features used to make a prediction, and y represented the FATHMM_PREDICTION target class for each of the entries in the COSMIC dataset. These data were further segregated into training and test sets via the train_test_split sklearn function (Pedregosa et al., 2011). A series of test set sizes were evaluated before selecting the split to use on the feature set. The RandomForestClassifier was instantiated and fit on the X and y training sets, called X_train and y_train, respectively. Following model fitting, a set of predicted and actual target features were generated by mapping the X_test set onto the trained classifier. This predicted value, y_pred, was then compared against y_true. Such comparison was essential to ensure the same randomized subsets of the entire dataset were used following the train_test_split function. After defining y_pred and y_true, a classification report was generated to examine key performance metrics of the classifier. Additionally, a ranked list of features by importance was outputted and examined for each gene.

Extremely Randomized Trees Classifier

While the random forest classifier was used extensively in the reviewed literature (Carter et al., 2009; Kaminker et al., 2007), other ensemble-based classification methods were implemented to explore performance of different classification approaches on the studied genomic datasets. The Extremely Randomized Trees algorithm (Geurts et al., 2006; Pedregosa et al., 2011) is another algorithm available with scikit-learn and is an extension of the core random forest methodology, also seen in the reviewed literature

(Wood et al., 2018). The algorithm extends the randomness incorporated into the classification, specifically concerning different splits of the training set (Geurts et al., 2006; Pedregosa et al., 2011). Thresholds for splits are selected randomly for each of the candidate features. From this selection, the best threshold is selected for further splitting rules (Geurts et al., 2006; Pedregosa et al., 2011).

Model Parameter Customization

The extra-trees classifier has shared customization options with the random forest classifier in the scikit-learn toolkit. As such, the same model parameters were evaluated and adapted, beginning with the number of estimators or trees used. Following an assessment of the number of estimators, the extra-trees classifier bootstrapping was assessed in combination with the out-of-bag (OOB) sample inclusion for generalization error enhancement (Pedregosa et al., 2011). Finally, the criterion for split quality assessment was also modified from the default Gini impurity to assess the entropy or information gain approach as an alternative.

As performed for the random forest implementation, the ExtraTreesClassifier was instantiated and fit using training data from the X and y subsets of the COSMIC mutation data, generated via the same incrementally adjusted training subset to enable better cross-model comparison and to optimize the training and test set sizes on the dataset for each gene. Once fit, the y_pred target subset was generated by mapping the X_test data onto the trained model, and a classification report and feature importance list were generated.

AdaBoost Classifier

In addition to the random forest and extremely random trees classifiers, another ensemble classifier was applied to the datasets. The adaboost classifier is an ensemble boosting algorithm, first introduced by Freund and Schapire in 1997. The method of the adaboost classifier differs from both random forest and extremely randomized trees approaches. The AdaBoost algorithm works by fitting a sequence of so-called weak learners, described as models providing marginal improvement over random guessing (Freund & Schapire, 1997; Pedregosa et al., 2011). This process is performed on different versions and subsets of the data. The result combines a weighted majority vote/sum to produce a final prediction. Throughout the training process, weights are applied to different training samples and modified based on the correctness of the prediction at each boost step (Pedregosa et al., 2011)

Model Parameter Customization

The adaboost classifier takes fewer customization parameters which predominantly differ from both the random forest and extra-trees model parameters. The first parameter modified was the number of estimators. This parameter is like the number of estimators or trees utilized in both the random forest and extra-trees classifiers, but instead represents the number of estimators that are used for boosting before termination (Pedregosa et al., 2011). Should a perfect fit be achieved, this process is stopped early. The parameter was incremented from n = 10 estimators to n = 320 estimators, doubling the number of estimators at each increment, as was the case for both random forest and extra-trees models. A maximum n = 400 estimators was also assessed.

In addition to the number of estimators, the learning rate parameter was also tuned. This parameter is a measure of the weight that is applied to each classifier for each iteration of boosting (Freund & Schapire, 1997; Pedregosa et al., 2011). As learning rate increases, fewer iterations may be needed as defined by the number of estimators used (Pedregosa et al., 2011). Conversely, reducing the learning rate may cause the algorithm to regress toward a weak learner, which in turn would require a greater number of estimators for boosting to achieve improved accuracy and performance (Freund & Schapire, 1997). Thus, learning rate will also be examined at several values and compared against the number of estimators to identify the optimal tuning for the dataset at hand.

Finally, the last parameter adjusted for the adaboost classifier was the algorithm used for implementation. The scikit-learn toolkit offers two options, the SAMME (Stagewise Additive Modeling using a Multi-class Exponential loss function) and real SAMME.R boosting algorithms (Hastie et al., 2009; Pedregosa et al., 2011). The SAMME algorithm was developed as a multi-class extension of the base adaboost algorithm and may reduce test error overall for binary classification problems (Hastie et al., 2009). The real SAMME.R algorithm is the default choice, and uses the predicted class probabilities to boost, whereas the discrete SAMME alternative uses errors in the predicted class labels to adapt (Pedregosa et al., 2011). While real SAMME.R is expected to reduce both train and test error compared to the discrete alternative (Pedregosa et al., 2011), both were examined to confirm this assumption in the examined datasets.

Similar steps were taken for model training and testing, including developing a training subset using a split of the total dataset data, and mapping the test data for the

independent features onto the model to yield a predicted target feature set, y_pred. Once generated, a classification report was constructed using the predicted vs true target features on the test set. Additionally, feature importance rankings were exported for further comparison. Precision-recall (PR) and receiver operating characteristic (ROC) curves were generated by the final implementation.

Chapter III.

Results

The dataset of COSMIC mutation data supplemented with other customized features was used to train three different ensemble-based machine learning classifiers, including the random forest, extremely randomized trees, and adaboost classifiers. Each classifier model was trained and tested for each of the genes of interest to compare behaviors between models and between gene-specific feature sets.

APC Analysis and Model Evaluations

The Adenomatous Polyposis Coli or APC gene serves various cellular functions, including *Wnt* signaling pathway antagonism and secondary functions such as facilitating cellular migration and adhesion, among others (Hanson & Miller, 2005). Research has also elucidated APC functionality independent of *Wnt* signaling, demonstrating further tumor suppressive capabilities via mitotic spindle regulation and DNA replication inhibition (Hankey et al., 2018). As such, this gene plays an important role in normal tumor suppressive behaviors. The APC gene is the most frequently mutated gene in colorectal cancers and other epithelial cancer syndromes (Lesko et al., 2014). As of 2018, colorectal cancers are estimated to be the third most deadly and fourth most diagnosed worldwide and pose a significant threat to patient health and survival. (Rawla et al., 2019), making the study and classification of APC mutations important to improve treatment options for patients. Three machine learning classifiers were constructed using

ensemble-based methods to define an approach for classification of APC genomic mutations in cancer.

Random Forest

The random forest classifier was the first classifier implemented for this research. Following the training of the classifier using a subset of the input data for each gene independently, predictions on the test data were made using the model. To simplify performance measurement, the problem was reduced to a binary classification problem to describe each entry in the dataset as either pathogenic or benign.

The APC random forest tuning was assessed using various parameter combinations (See Chapter II, Research Methods). In each instance of the APC random forest classifier, the square root method for determining the maximum number of features used to train each tree matched or improved performance over the alternative log2 method. When other parameters were constant, the differences in precision and recall between the two methods typically stayed within 0.2%. The number of estimators used for each implementation did contribute to changes in overall performances, though with less consistency as compared to the changing method for determining the maximum number of features. As the number of estimators rose, so too did precision, recall, and F1-score for most classifier instances. The model using 400 total estimators achieved the highest performance and minimized out-of-bag error compared to other classifiers (OOB error = 3.5%). However, improvements in these output measurements were marginal compared to the implementation using only 160 estimators in model fitting. In the 160-estimator model, OOB error marginally increased by ~0.1% while other metrics remained consistent. From a computational resource efficiency perspective, the lower estimator

model was the more practical choice for this application. Finally, the use of Gini impurity to assess the quality of each split was superior for each estimator level, which was apparent in the consistent reduction of OOB error where the Gini impurity was used compared to the entropy-based alternative.

The data for the APC gene was used to train this optimized random forest classifier using the parameter tunings described previously (See Chapter II, Methods). The test set included 5125 variant entries from the COSMIC Mutation Data set. A classification report was generated using scikit-learn's classification_report function (Pedregosa et al., 2011). The classifier achieved a positive predictive value (PPV) of 0.95 for classification of pathogenic alterations, and a PPV of 0.98 for benign alterations. The recall (sensitivity) of the classifier was 0.97 for pathogenic alterations and 0.96 for benign alterations. For pathogenic and benign, a F1-score of 0.97 and 0.96 was achieved, respectively. The out-of-bag error for the implemented random forest algorithm was 0.041, which was an improvement over the other random forest iterations examined for APC data. A plot of the receiver operating characteristic (ROC) curve and precision-recall (PR) curve is included in Figure 1**.** In addition to ROC and PR curves for the classifier, a list of the top weighted features was also gathered and can be found in Table 1.

Extremely Randomized Trees

The extremely randomized trees (extra-trees) classifier was the next ensemble-based method implemented on the APC dataset. The extra-trees classifier can be considered an extension of the random forest classifier in that a new layer of randomness is introduced (Pedregosa et al., 2011). While a random subset of candidate features is

utilized in each tree, thresholds are taken randomly for each candidate feature, allowing the model to select the best threshold to define a new splitting rule (Pedregosa et al., 2011). For each extra-tree's implementation, the same training split size (0.33) was used from the sample data.

The performance distribution observed for the extra-trees classifier was comparable to the random forest implementations, which is not unsurprising given the relatedness of these algorithms (Pedregosa et al., 2011). Review of performance indicators again demonstrated that 160 estimators was most appropriate for the size and complexity of the dataset for APC. Additionally, a square root max feature method and use of Gini impurity for split quality assessment were selected as other tuning parameters. Except for models with a low (10) or high (400) number of estimators, precision and recall calculations were quite similar for many of the different models. Thus, out-of-bag error and number of estimators were the primary parameters leveraged to achieve high modeling performance and to optimize computational efficiency.

The APC data generated using the extra-trees classifier produced a PPV of 0.95 for pathogenic variants and 0.98 for benign calls. The algorithm also produced a recall of 0.98 and 0.94, pathogenic and benign variants, respectively. The F1-score was 0.97 for the pathogenic class and 0.96 for the benign class. These values were notably similar to the random forest implementation output, albeit with a slightly elevated out-of-bag error of 0.042. The ROC and PR curves for the APC extra-trees classifier are displayed in Figure 2. Additionally, the top weighted features by importance for this classifier are listed in Table 2.

AdaBoost

Finally, the adaboost classifier was also executed on the APC dataset. The adaboost method differs from both random forest and extra-trees through boosting, whereby a combined weighted majority vote of predictions is used to produce a final prediction on modified versions of the data (Freund & Schapire, 1997; Pedregosa et al., 2011).

The parameter tuning for the adaboost classifier presented a more extreme variation in performance between different iterations. The learning rate had a significant impact on precision and recall. Increasing the learning rate to 2.0 from the default 1.0 rate caused a substantial drop in performance, with precision and recall falling to roughly half their achieved levels when using the SAMME algorithm, regardless of the number of estimators used. The default SAMME.R algorithm was more heavily affected by the change in learning rate, causing precision and recall falling below 0.1 in some cases for benign alterations, and below 0.4 for pathogenic cases. Thus, it was clear that a default learning rate and default algorithm selection would be optimal for this algorithm. With learning rate and algorithm constant, the selection of number of estimators had a marginal impact on the performance of the classifier. Performance generally improved as the number of estimators was increased but stabilized between 40 and 80 estimators. Furthermore, the parameter began to negatively affect some metrics as the number of estimators approached 400. Thus, 40 estimators were selected as the optimal target, which maximized key performance metrics while also maintaining computational efficiency. The adaboost algorithm cannot calculate an out-of-bag error (Pedregosa et al., 2011) so no OOB error rate was used to directly inform optimal parameter tuning.

The performance of the adaboost classifier was varied overall as compared to the random trees or extra trees ensemble methods, producing a PPV of 0.94 and 0.97 for pathogenic and benign variants, respectively. Similarly, recall showed slightly altered performance for both pathogenic (0.98) and benign (0.93) variants as compared to sensitivity achieved by either random forest or extra-trees approaches. F1-score for both pathogenic and benign alterations was 0.96 and 0.95, respectively. The ROC and PR curves for the APC adaboost classifier are shown in Figure 3, and the top feature rankings are listed in Table 3.

Feature Comparisons

Each of the models identified different feature rankings and relative importance scores when using the same train-test split. The top two features for the random forest model were related to mutation descriptions (nonsense, frameshift deletion), while feature three was a feature describing mutations without a SNP identifier. Features four and five for the random forest were a mutation description of missense, and a feature describing variants with unknown clinical significance as determined by dbSNP.

For the extra-trees model, the top four features were identical to the random forest, including mutation descriptions of nonsense, frameshift, and missense, as well as a feature describing unknown SNPs. Although relative importance of these features was slightly different, their inclusion at the top of the list is not unexpected. As a tumor suppressor, inactivation of APC is known to play a role in tumorigenesis (Zhang & Shay, 2017), and the importance of frameshift and nonsense events is understood. The fifth highest ranked feature by importance described variants that are confirmed SNPs

according to dbSNP, suggesting that the SNP designation of a mutation was a valuable descriptor to predict variant status.

The adaboost classifier selected a different list of top features when constructing the model on the same training data. The highest ranked feature for this model denoted mutations with a nonsense functional effect, again highlighting the relevance of truncating alterations in tumor suppressors. Beyond the nonsense designation, adaboost classifier ranked a dbSNP prediction of likely pathogenic to be the second highest feature by importance. The third feature ranking described a genome position region, spanning coordinates chr5:112839769-112839942. Visual inspection of the APC genomic mutation landscape shows a clear hotspot region based on localization of mutations in this region. Interestingly, the fifth through seventh feature also identified genomic coordinate bins, suggesting the presence of key genomic hotspot regions in APC. The fourth highest ranked feature described mutations in tumors with a primary site of large intestine, which is relevant given APC's prevalence in colorectal cancer syndromes (Zhang & Shay, 2017). The fifth highest ranked feature denoted mutations within the range of chr5: 112839769-112839942.

TP53 Analysis and Model Evaluations

The TP53 gene was the next tumor suppressor gene examined in this research. Discovered in 1979, TP53 performs essential tumor suppressor activities via cell cycle arrest and induced apoptosis, responding to numerous stressors faced during a cell's lifetime (Brady & Attardi, 2010). Research has also demonstrated TP53 mutation is a common aspect of many cancers, including those related to inherited TP53 deficiency and Li-Fraumeni Syndrome (Guha & Malkin, 2017). When altered, TP53 can impact a

host of different processes, from transcription and DNA damage response to reduction in both apoptosis signaling and cell proliferation control (Kastenhuber & Lowe, 2017). Thus, TP53's status as an established cancer-related gene makes it an ideal target for study in the context of mutation classification.

Random Forest

The TP53 COSMIC data supplemented with custom features was also modeled using a random forest classifier. Compared to other genes studied, the test set included a large fraction of samples, totaling 11,588 unique COSMIC mutation data entries. As observed for other genes, the out-of-bag error was inversely related to the number of estimators, stabilizing once at least 40 estimators were used in the modeling. Overall precision (PPV) and recall (sensitivity) values were relatively stable throughout different iterations of the parameter tuning. Using the square root maximum feature determination proved more successful than the log2 method, the latter reducing performance overall. Optimum performance was achieved when 80 estimators were used by the classifier, especially given the size of the dataset and the computational time needed for a greater number of estimators. The classification report for this model achieved a PPV of 0.95 for both pathogenic and benign target classes. The recall (sensitivity) was calculated at 0.97 for pathogenic alterations, while the sensitivity to benign classification suffered and was reduced to 0.89. Pathogenic and benign target classes achieved F1-scores of 0.97 and 0.89, respectively. Class splits were not perfectly balanced (8,922 pathogenic and 2,666 benign), but synthetic re-balancing of classes did not meaningfully improve measures of performance and thus was not used (Lemaitre et al., 2017). The ROC and PR curves for the TP53 random forest implementation can be found in Figure 4. Additionally, the top

features by importance were extracted from the model for comparison to the other algorithms implemented and are listed in Table 4.

Extremely Randomized Trees

For TP53 data, changes in performance were observed for the extra-trees classifier compared to random forest, with slight reductions in performance overall for each of the key evaluated metrics. A total of 80 estimators were once again used, striking an ideal balance between suppression of out-of-bag error, and maximizing both PPV and sensitivity for each of the target classes. The Gini impurity for split assessments and square root determination of maximum features used in training once more proved superior over the entropy or log2 alternatives, respectively. Compared to the random forest, similar performance was observed for PPV (0.95 and 0.95) and recall (0.99 and 0.84) for pathogenic and benign variants, respectively. Additionally, a slight improvement was measured for the F1-score for benign alterations (0.89) compared to the random forest classifier. The use of synthetic re-balancing of classes was once again ineffective in improving performance for the classifier and was excluded from model training (Lemaitre et al., 2017). The achieved OOB error was 0.049. The TP53 extra-trees ROC and PR curves can be found in Figure 5, while the top ranked features can be found in Table 5.

AdaBoost

The adaboost classifier again demonstrated greater variability in performance based on the parameters tuned throughout each iteration of the TP53 models. The default learning rate of 1.0 was ideal, whereas a learning rate of 2.0 significantly impacted both

precision and recall for each target class, regardless of either the number of estimators or the algorithm of choice. The algorithm choice was still impactful, with the default SAMME.R algorithm proving far superior over the alternative SAMME option. The use of 40 estimators was optimal based on a balance between key performance indicators and computational speed of the model training and execution. For the optimal adaboost classifier for TP53 data, a PPV of 0.91 was achieved for pathogenic alterations, and a PPV of 0.88 was achieved for benign. Sensitivity was high for the pathogenic class (0.97), although substantially reduced for the benign class (0.69). Although the target classes did differ in size (8,922 for pathogenic and 2,666 records for benign), synthetic balancing during model training did not measurably improve performance for any key indicators examined (Lemaitre et al., 2017). The TP53 adaboost ROC and PR curves are displayed in Figure 6. Additionally, the top features were exported from the adaboost classifier and are listed in Table 6.

Feature Comparisons

The top TP53 features were compared between the three algorithms implemented. For the random forest model, the top feature by importance described missense mutations, followed closely by frameshift deletion mutations. Interestingly, the feature for nonsense mutations ranked 13th, significantly lower than in the APC data. The third and fourth highest ranked features described mutations designated as non-SNPs and those for which SNP status was unknown, suggesting that a lack of assigned rsID was highly informative to the classifier. Finally, the fifth highest ranked feature by importance describes a specific mutation genome position range, spanning from chr17:7675221.0-

7687510. The high importance ranking of this feature suggests that key mutation hotspots may be relevant in TP53 variant classification.

The extra-trees feature list demonstrated exact overlap with the random forest for the top five entries reviewed. The top features once again highlighted missense mutations, deleterious frameshifts, non-SNP or unknown SNP status, and finally a mutation genome region of chr17:7675221.0-7687510. The classifier feature rankings between random forest and extra-trees began to differ at about the 10th feature ranking. Beyond the 10th highest ranked feature, random forest and extra-trees highlighted additional mutation descriptions, genome position ranges, and bins of duplicate entries with different priority rankings compared to each other.

The adaboost classifier feature list differed from both the random forest and extra-trees models overall, however several top features did overlap. The top feature by importance was a mutation genome region defined as spanning chr17:7675221.0-7687510, the same mutation genome region identified in both random forest and extra-trees top five. The second highest ranked feature denoted mutations that were not near any characterized domain according to Pfam data, while the third entry described TP53 mutations with between 294 and 743 supporting records in COSMIC describing the same mutation. This ranking suggests the adaboost classifier weighted classification differently depending on the level of redundancy of a given mutation in the COSMIC database. The remaining features demonstrated identical ranking, and the next several features included mutations with unknown origin (according to COSMIC records), as well as several other mutation genome bins spanning different regions of the TP53 gene.

RB1 Analysis and Model Evaluations

The RB1 gene is another important tumor suppressor in the context of normal cellular function and study of cancer progression. The RB1 gene was notably the first tumor suppressor identified when researchers recognized that inactivation of the protein directly caused pediatric retinoblastoma (Chinnam & Goodrich, 2013). Prior study also points to RB1's pivotal role in various transcription regulation and in controlling the assembly or disassembly of different proteins via complex signaling interactions (Chinnam & Goodrich, 2013). As mentioned, deleterious mutations in RB1 may trigger tumorigenesis in various cell types throughout the body. Based on RB1's demonstrated role in cancer development and in the history of tumor suppressor research it was included in the study of machine learning classification in this work.

Random Forest

The third tumor suppressor gene, RB1, was also run through a random forest classifier using the same data preparation approach described (see Chapter II, Research Methods). Compared to both APC and TP53, RB1 used only 1,262 variants in the classifier's test set due to a significantly smaller collection of mutations available in the COSMIC database (Tate et al., 2018). The random forest implementations varied the number of features, the criterion to assess splits, and the max feature determination to evaluate the optimal tuning settings. In general, the out-of-bag error decreased as the number of estimators went up, a trend observed previously for both APC and TP53. The OOB error was minimized when 400 estimators were used, however the implementation performance was inefficient, even with a relatively small test set to evaluate. The OOB error did begin to stabilize once the number of estimators reached 80, which was then

used as a starting point for evaluation of other model parameters. Precision and recall for each classifier were assessed alongside the F1-score, and optimal performance was achieved with 160 total estimators and with the square root method of determining the max number of features. The difference between Gini and entropy-based split criteria was negligible, so the default Gini impurity was selected for convenience. With this tuning, the RB1 classifier achieved a PPV (precision) of 0.91 and 0.93 for pathogenic and benign classes, respectively. Additionally, sensitivity for pathogenic variants was 0.94, while sensitivity for benign alterations was calculated at a lower level of 0.90. The F1-scores for pathogenic and benign classes were 0.92 and 0.91, respectively. The calculated OOB error was calculated as 0.074. The ROC and PR curves for the RB1 random forest implementation are displayed in Figure 7 and the top features by importance can be found in Table 7.

Extremely Randomized Trees

The RB1 extra-trees implementations followed a parameter tuning pattern like those used in the random forest implementations. Out-of-bag error remained relatively high regardless of the number of estimators used when the number of estimators exceeded 40. The performance of the extra-trees implementations mirrored that of the random forests, with 80 and 160 estimator models yielding similar results. Because the time efficiency of both 80 and 160 were comparable, the 160 estimator models were preferred given slightly improved behavior on the RB1 dataset. Once more, the default Gini impurity proved superior when assessing the quality of each split, and the default square root calculation for the maximum number of features also exceeded the alternative log2 method. With the optimal model parameters, a PPV of 0.91 was achieved for

pathogenic alterations, and PPV for benign alterations was higher at 0.93. The model was more sensitive to pathogenic alterations (0.94) than benign (0.90). The F1-scores were identical to the random forest at 0.92 and 0.91 for the two classes. Additionally, the OOB error was identical at 0.074. The ROC and PR curves for the extra-trees implementation are displayed in Figure 8. Additionally, the top feature rankings extracted from the model are listed in Table 8.

AdaBoost

The adaboost classifiers demonstrated greater extremes in performance as compared to either random forest or extra-trees approaches based on the parameter tuning. Models built on the default SAMME.R algorithm consistently outperformed the alternative SAMME algorithm selection. A learning rate of 2.0 proved more unstable, maximizing sensitivity for pathogenic alterations at the expense of precision, which would be unacceptable a clinical setting. Using the default 1.0 learning rate and 80 total estimators, optimal performance balance was achieved for both target classes. With such an implementation, pathogenic and benign classes had calculated PPVs of 0.89 for both classes. Sensitivity was different for each class, with a calculated recall of 0.90 for pathogenic and 0.87 for benign. The F1-scores were 0.89 and 0.88 for pathogenic and benign classes, respectively. The RB1 adaboost ROC and PR curves are shown in Figure 9. Additionally, feature rankings by importance were exported and can be visualized in Table 9.

Feature Comparisons

Feature importance was extracted from each of the three algorithms implemented and compared. The random forest classifier ranked the feature describing variants with unknown SNP status as the highest by importance. The second through the fourth highest ranked feature captured mutation descriptions of missense, deleterious frameshifts, and nonsense mutations. Again, the role of truncating short variant alterations was reinforced in the context of tumor suppressor genes, as seen in both APC and TP53 data. The fifth highest ranked feature described variants with a clinical significance ranking of pathogenic, a ranking predicted by the integrated dbSNP data.

The extra-trees classifier shared identical features with the random forest classifier. The top five features once more described SNP unknown status, followed by missense, frameshift (deletion), and nonsense functional statuses. The fifth feature once again denoted a pathogenic prediction from the dbSNP dataset. The feature set rankings for both random forest and extra-trees classifiers remained nearly identical until the ninth ranked feature was reached, at which point the features sets diverged slightly in terms of rank order.

The adaboost implementation highlighted several other features as being important for model training and performance. The top ranked feature by importance for adaboost was the feature describing a clinical significance of pathogenic, as predicted by the integrated dbSNP dataset. Interestingly, the second highest weighted feature denoted mutation from tumors with between 12,308.5 and 219,194 other mutation records in the COSMIC database, identified by matching the tumor ID of the variant in question against the entirety of the COSMIC mutation database. Indeed, mosaicism and hypermutation

has been characterized in some inherited cancer syndromes such as retinoblastoma (Rodriguez-Martin, et al., 2019). Such a weighting suggests a perceived relationship between overall tumor burden and pathogenicity of RB1 alterations, a feature which was not ranked highly in either random forest or extra-trees classifiers. Beyond these two features, the third feature listed was COSMIC signature SBS87. The COMSIC dataset asserts the SBS87 signature originates from Thiopurine chemotherapy treatment based on experimental evidence (Alexandrov et al., 2020). Another clinical study further supports the ototoxicity of this treatment regimen in children treated for retinoblastoma (Soliman et al., 2018). Beyond these features, the adaboost classifier ranked another tumor mutation burden-related feature highly, denoting mutation records from tumors with between 6,059 and 12,308.5 other detected mutations present in the COSMIC database. Finally, the fifth feature listed describes a mutation genome position span, ranging from chr13:48376917-48380217. Visual inspection of the RB1 COSMIC landscape shows relatively consistent distribution of mutations across the span of the gene. However, the region of chr13:48376917-48380217 does show some clustering of mutations, many of which are nonsense or frameshift deletions/insertions.

EGFR Analysis and Model Evaluations

The epidermal growth factor receptor (EGFR) is an important protein involved in a multitude of cellular functions. The EGFR protein localizes on the cell surface, facilitating cellular differentiation and proliferation to further cell growth (Voldborg et al., 1997). Modern cancer research overwhelmingly recognizes EGFR's role in tumorigenesis in several cancer subtypes, whereby upregulation and over expression fuel unrestricted cell growth (Sigismund & Lanzetti, 2018). Several targeted EGFR therapies

have been developed, improving survival rates of patients with EGFR-driven disease in several different tumor types (Xu et al., 2017). Thus, a machine learning implementation to categorize somatic EGFR alterations in a clinical setting would prove valuable in determining prognosis and developing future treatment options for various patient populations.

Random Forest

The COSMIC mutation data for EGFR was the first of the oncogenes to be modeled using the random forest classifier. In total, 1,432 variants were included as part of the test set once the model was trained, with 671 variant records being classified as pathogenic, and 761 being classified as benign. The model was assessed by manually adjusting different parameters as performed for other random forest implementations. As seen previously, the Gini impurity proved to be the best assessment of split quality, and the max features was set to the default square root value. A model using these parameters along with n = 80 estimators produced the best balance of OOB score minimization, maximization of precision and recall, and computational efficiency. This optimized model recorded a PPV of 0.94 for pathogenic variants, and a PPV of 0.98 for benign alterations. For sensitivity, pathogenic and benign variants each scored identically with a calculated recall of 0.96. The F1-score was also calculated as 0.96 for both target classes. The out-of-bag error for the random forest implementation was 0.03. The top features identified by the model can be found in Table 10. Additionally, ROC and PR curves for the EGFR random forest implementation are displayed in Figure 10.

Extremely Randomized Trees

The extra-trees implementation for EGFR showed similar performance distribution to the random forest implementation, where standard parameter tuning was performed in accordance with the research methods. When trained using a total of 80 estimators, performance metrics were maximized while the out-of-bag error remained relatively low. While little difference was observed between the Gini and entropy split assessment methods, the Gini method was selected as the default value for simplicity. As with the random forest algorithm, the use of a log2 method to determine the maximum features negatively impacted performance, so the default square root method was utilized. With these parameters, PPV of 0.94 and 0.98 was achieved for pathogenic and benign variants respectively, while sensitivity for both classes remained high (0.98 for pathogenic alterations, 0.95 for benign). The F1-scores were once again 0.96 for both target classes, with an OOB error consistent with that achieved by the random forest (0.031). The top features identified by the model can be seen in Table 11, and ROC and PR curves for EGFR extra-trees modeling displayed in Figure 11.

AdaBoost

The adaboost classifier was the third classifier trained using the EGFR training data. The adaboost classifier had reduced performance output overall compared to the random forest and extra-trees algorithms, though the reduction was not as extreme as that observed for classifiers trained using other genomic data. Based on performance output and computational time, ideal performance was achieved when 40 total estimators were used to train the model. Additionally, the learning weight of 1.0 proved superior, as did the SAMME.R algorithm over the SAMME alternative. With these parameters, a PPV of

0.95 was achieved for pathogenic variants, and a PPV of 0.96 was achieved for benign alterations. Sensitivity for pathogenic and benign alterations was also high, calculated as 0.96 and 0.95, respectively. The F1-scores were like both random forest and extra-trees models, with an F1-score of 0.95 for pathogenic alterations and 0.96 for benign. The ROC and PR curves for the EGFR adaboost implementation are displayed in Figure 12. The top features from the adaboost classifier can also be found in Table 12.

Feature Comparisons

Feature comparisons for each of the three EGFR models was also performed. For the random forest, the top five features were ranked with similar feature weights. The highest ranked feature described mutations with a missense functional effect, while the second ranked feature described mutations with an unknown SNP status according to either COSMIC or dbSNP. The third through fifth features were all COSMIC signatures; SBS40, SBS2, and SBS16, respectively. SBS40 has an unknown signature designation in COSMIC and the signature composition is generally balanced in terms of percentage of single base substitutions contributing to the signal itself. COSMIC also notes that SBS40 was described by validated evidence, and additionally mutations contributing this signature may correlate with patient age for some cancers (Alexandrov et al., 2020). The SBS2 signature is attributed to the activity of the AID/APOBEC family of cytidine deaminases, according to COSMIC (Alexandrov et al., 2020). Research suggests that this signature may manifest directly via DNA replication errors across uracil, or via error-prone polymerases replicating in the presence of sites generated by excision repair removal of uracil (Alexandrov et al., 2020). The fifth-ranked SBS16 signature also has unknown significance in the COSMIC database. The signature is notably defined by a

high percentage of T>C single base substitutions. Preliminary data suggests that low levels of nucleotide excision repair on untranscribed strands and elevated levels of DNA damage on untranscribed strands may contribute to the signature (Alexandrov et al., 2020).

The extra-trees features were notably different than the random forest features, which was not usually the case for other genomic data studied. The top ranked feature according to this classifier described mutations with a missense functional effect. The second and third highest ranked features by importance described variants with an unknown SNP status or with a non-SNP designation. The fourth ranked feature described mutations near a disordered domain, based on incorporated Pfam data. The Pfam database describes disordered regions as being conserved, however such regions often demonstrate biased sequence composition (Mistry et al., 2020). The fifth highest ranked feature is a genome region, spanning chr7:5519182-55191822. This 41-base pair span covers a region of EGFR intron 2 (NM_005228.5). While the COSMIC data examined does not explicitly contain intronic variants, a complex deletion with a protein effect of p.V30_R297>G spans the intron 2 region. Because of the position averaging approach to simplify description for complex alterations, this mutation was localized to the region described, thus suggesting the relevance of this 801 base pair deletion and others in the same region.

Finally, the adaboost classifier features were examined and compared to the other algorithm implementations. The top ranked feature was COSMIC signature SBS89. Although SBS89 has an unknown etiology, it is predicted to be active in the first decade of life, according to COSMIC records (Alexandrov et al., 2020). The signature shows

similar percent contribution of different base substitutions, though C>T and T>G appear to be enriched for some combinations of flanking bases. The next highest feature denoted mutations with a coding silent status, which may include synonymous or non-protein changing mutations which conserve the overall amino acid sequence of the protein. The third highest ranked feature described COSMIC signature SBS17b. According to the COSMIC database, SBS17b has some prior association to a form of fluorouracil chemotherapy treatment, characteristic of damage caused by reactive oxygen species resulting from treatment (Alexandrov et al., 2020). A causal link has been established between the impact of reactive oxygen species on EGFR-mediated cancer progression, especially where EGFR resistance is concerned in the context of tyrosine kinase inhibitors or TKIs (Weng et al., 2018). This signature may describe patterns in the dataset indicative of some tumors developing resistance alterations conferring survival advantage to their tumors. The fourth highest ranked feature described mutations with missense functional status, while the fifth highest ranked feature described mutations near low complexity domains as modeled by Pfam. Although difficult to study evolutionarily, some researchers suggest low complexity domains play a role in mediating types of protein-protein interaction (Kastano et al., 2021).

ERBB2 Analysis and Model Evaluations

The ERBB2 gene was the next oncogene tested in these experiments. A relative of EGFR in the epidermal growth factor receptor family (Negro et al., 2004), the ERBB2 gene encodes a transmembrane tyrosine kinase receptor or TKR (Bertucci et al., 2004). Notably, the receptor can heterodimerize to promote neuregulin signaling (Negro et al., 2004). In humans, this complex network of signals can play a role in neural circuitry,

myelination, and neurotransmission/plasticity (Mei & Nave, 2014). Deregulation and overexpression of ERBB2 in humans is a known driver of oncogenesis, frequently observed in human breast cancers (Bertucci et al., 2004). Notably, the monoclonal antibody Trastuzumab was one of the first developed for treatment of ERBB2-positive breast cancer (Dean & Kane, 2015). Because pathogenic ERBB2 genotypes typically manifest as over-expression via amplification (Bertucci et al., 2004), the gene is an interesting target for machine learning-based classification of short variant alterations in this research to elucidate possible drivers in cancer.

Random Forest

COSMIC mutation data for ERBB2 was split into training and testing sets and executed on several random forest model implementations as was done previously. With a variant test set of only 885 total records in the test set, performance was notably reduced compared to the previously trained classifiers, though category splits were comparable in size (339 benign alterations versus 546 pathogenic alterations). As the number of estimators increased, the out-of-bag error decreased, however above 160 estimators the out-of-bag error began to increase once more. This is predicted to be a consequence of the relatively small test set size. Overall, the classifier with 160 estimators demonstrated the best balance of different performance indicators when using both the default Gini impurity and the square root max features method. While the out-of-bag error of 0.112 was elevated, other precision and recall metrics were improved. A PPV of 0.91 was achieved for the pathogenic class, while a PPV of 0.89 was achieved for the benign class. The sensitivity of pathogenic alterations was higher at 0.94, while benign variant sensitivity suffered at a calculated level of 0.85. The F1-scores registered

at 0.92 and 0.87 for pathogenic and benign, respectively. The ROC and PR curves for ERBB2 random forest performance can be found in Figure 13. Additionally, the top features were explored from the model and are listed in Table 13.

Extremely Randomized Trees

Performance characteristics for the extra-trees implementation on ERBB2 once more mirrored those of the random forest due to the relatedness of the two algorithms. The classifier instance using 160 estimators, default Gini impurity, and a square root max features cap exceled in each of the performance categories. Worth noting is that nearly identical performance was achieved using only 80 estimators and using the entropy-based split quality assessment. If performance considerations were of high importance, this implementation may reduce computational time of the algorithm. However, given the small test set size for this data, the 160-estimator model with default settings was chosen for simplicity. The out-of-bag error for this implementation was marginally reduced compared to the random forest at an error fraction of 0.107. Precision (PPV) for the pathogenic class was 0.91, while the benign class had a PPV of 0.89. The sensitivity registered at 0.94 and 0.85 for pathogenic and benign, respectively. The F1-scores were identical to the random forest, at 0.92 and 0.87 for pathogenic and benign classes, respectively. The ROC and PR curves for the ERBB2 extra-trees model are displayed in Figure 14. The top features were also exported from the model and are listed in Table 14.

AdaBoost

The adaboost implementations for ERBB2 were varied according to the number of estimators, the learning rate, and the algorithm selection. As seen in prior

implementations, the SAMME.R algorithm proved superior to the SAMME alternative, stabilizing both PPV and sensitivity between classes compared to the more extreme differences in calculated values observed for the SAMME implementations. An acceptable balance of performance was achieved with 40 estimators used and a learning rate of 1.0. While fewer estimators (20) did yield promising performance, improvements were at the expense of PPV for pathogenic alterations. With 40 estimators, PPV was 0.89 for pathogenic alterations and just 0.83 for benign. Sensitivity was higher for pathogenic alterations (0.90) than benign (0.83). The F1-scores showed a similar disparity, calculated at 0.89 for pathogenic and just 0.83 for benign. Displays of the ROC And PR curves for this adaboost implementation can be found in Figure 15. Additionally, the top features were examined and are listed in Table 14.

Feature Comparisons

A comparison was performed between the top features of each of the algorithms implemented for ERBB2. The random forest algorithm ranked the missense functional effect feature highest by importance. The second highest ranked feature described mutations near unknown functional domains, while the third feature denoted mutations with an unknown SNP status according to both COSMIC and dbSNP. The fourth highest ranked feature described in-frame insertion mutations, and the fifth highest ranked feature denoted mutations predicted to be pathogenic according to dbSNP.

The extra-trees model shared the same top five features with the random forest classifier, however with slightly modified rankings by importance. The top three features were missense functional designation, followed by unknown nearby domains and a SNP unknown designation. The fourth feature was a clinical significance prediction of

pathogenic based on data from dbSNP. The fifth ranked feature described in-frame insertion mutations, as was seen in the random forest feature list.

The adaboost features differed from both the random forest and the extra-trees classifiers. The top adaboost feature by importance was COSMIC signature SBS9, followed by COSMIC signature SBS8 as the second highest ranked feature. The SBS9 signature characterizes mutations arising from replication by polymerase eta, specifically as part of somatic hypermutation in lymphoid cells (Alexandrov et al., 2020). However, COSMIC notes that this etiology is made by statistical association alone, so the signature may have different origins and be relevant in ERBB2-mediated cancers as well. The SBS8 signature has an unknown etiology according to COSMIC. The signature composition is heavily weighted toward C>A and T>A mutations based on experimental data (Alexandrov et al., 2020). The third highest ranked feature identified mutations with a primary tumor site of prostate. Research suggests that signaling from ERBB2 can increase the expression of the Androgen Receptor, a common driver in some prostate cancer patients (Gao et al., 2016). The fourth listed feature denoted mutations with a deleterious frameshift functional effect, while the fifth listed feature describes mutations with a malignant melanoma tumor type. While deleterious protein effects may be detrimental to translation of oncogenic proteins, the inclusion in the list may be to further separate benign or non-activating alterations in the dataset. Although the role of ERBB2 mutation in melanoma is not well characterized, recent research suggests it can represent a targetable option for some patients with BRAF V600 wild type melanoma (Gottesdiener et al., 2018).

PIK3CA Analysis and Model Evaluations

The PIK3CA gene is the third and final oncogene examined as part of this research. The PIK3CA protein product is a phosphoinositide kinase (PIK), involved in phosphorylation and signal transduction (Samuels & Waldman, 2010). As with other proto-oncogenes, PIKICA supports a host of cellular activities including proliferation, survival, motility, and general cell growth upon activation (Karakas et al., 2006). While somatic alterations were initially discovered in patients with colorectal cancers, PIK3CA has widespread prevalence in various human tumors including breast, brain, skin, and ovarian cancers, among others (Samuels & Waldman, 2010). Thus, PIK3CA was selected for study to explore more robust machine learning variant classification approaches.

Random Forest

The random forest classifier was implemented by assessing different combinations of parameters to tune the model for optimal performance. The split for PIK3CA test data included 3,853 pathogenic variants and 515 benign alterations. Given the highly unbalanced test set, the Imbalanced-learn library was included to synthetically boost the minority class in the model training (Lemaitre et al., 2017). A comparison of the raw unbalanced data and the synthetically balanced data showed no meaningful change in key performance metrics, so the balancing library was not used. Overall, each model exhibited high quality performance, and improvements were incremental for each combination of parameters. Once again, out-of-bag error decreased sharply at about 40 estimators and remained low throughout the remaining implementations. Beyond 40 estimators used, all models performed similarly regardless of the use of Gini or entropy split assessments and regardless of the max features used in training. As a result, a model

with 40 estimators was used with default settings of Gini impurity and square root max feature determination. With this tuning, a PPV of >0.99 was achieved for the pathogenic target class and a PPV of 0.97 for benign. Sensitivity and F1-scores for both target classes were the same as their respective PPVs (>0.99 for pathogenic and 0.97 for benign). For this implementation, OOB error was calculated at 0.009. The ROC And PR curves for PIK3CA are shown in Figure 16, while the top features extracted from the model are listed in Table 16.

Extremely Randomized Trees

The extra-trees implementations followed similar patterns to the random forest models tested. As seen in the random forest models, the out-of-bag error declined sharply for models using at least 40 estimators, which was again selected as the ideal number of estimators based on performance metrics. Within the implementations using 40 estimators, the default square root max feature value was preferred. While the entropy method for split assessment did reduce the OOB error for all implementations on the dataset over the default Gini impurity, the Gini impurity showed consistent performance and was selected for simplicity. Using this optimal tuning, the same values for PPV, sensitivity, and F1-score for both pathogenic and benign classes were achieved when compared to the random forest implementation (>0.99 for the pathogenic class and 0.97 for benign for each calculation). The ROC and PR curves for the PIK3CA extra-trees implementation are shown in Figure 17. Additionally, the top features were extracted from the model and can be found in Table 17.

AdaBoost

Finally, the adaboost classifier was deployed on the PIK3CA dataset. As observed previously, elevated learning rate (2.0) created more extreme separation of both PPV and sensitivity calculations between the two classes and was non-viable. Using a default learning rate of 1.0, the SAMME.R algorithm improved performance over the SAMME alternative. Using 40 estimators balanced performance output and computational time. With these parameters, the adaboost classifier achieved a PPV of 0.99 for the pathogenic class and 0.96 for benign. The same values (0.99 and 0.96) were calculated for sensitivity and F1-scores for the pathogenic and benign classes, respectively. The ROC and PR curves for PIK3CA adaboost modeling are displayed in Figure 18 while the top feature rankings are listed in Table 18.

Feature Comparisons

The top PIK3CA features for each classifier were exported and compared. For the random forest classifier, the top ranked feature by importance was the mutation description of missense. The second highest ranked feature described mutations that were not localized near a characterized domain according to Pfam data. The third and fourth highest ranked features by importance described variants with predictions of likely pathogenic and pathogenic according to dbSNP prediction algorithms incorporated into the dataset. Finally, the fifth highest ranked feature described mutations in tumors with a primary site of meninges. The role of PIK3CA is well studied in ovarian, breast, and colorectal cancer syndromes (Ligresti et al., 2009). However, clinical evidence also characterizes activation of PIKCA in some meningioma cancers, representing a possible target for neoadjuvant therapy for patients (Zadeh et al., 2016).

The extra-trees model demonstrated some variation in top features used to train the model compared to the random forest. The top ranked domain described mutations that were not near a known functional domain based on Pfam data, followed closely by a mutation description of missense as the second highest ranked feature by importance. The third feature was not listed in the random forest top five, describing mutations with 1.0 – 10.0 duplicate entries present in the COSMIC database. This finding suggests that mutation prevalence in the sample population informed classification behavior during training of the model. The fourth highest ranked feature described variants predicted to be likely pathogenic according to dbSNP data. The fifth feature described variants with an unknown SNP status according to dbSNP and COSMIC designations.

The adaboost classifier highlighted several other features ranked uniquely compared to either random forest or extra-trees models. The top ranked feature by importance denoted mutations without a known functional domain nearby according to Pfam data. The second and third top features listed were COSMIC signatures SBS88 and SBS39. The SBS88 signature is characterized by strong presence of T>C and to a lesser extent, T>G single base mutations with various flanking base combinations (Alexandrov et al., 2020). The signature is believed to originate from exposure to *E. coli* bacteria carrying a pks pathogenicity island. The result of such exposure is production of a genotoxic compound called colibactin (Alexandrov et al., 2020). The SBS39 signature is characterized by a strong percentage contribution of C>G single base substitutions, with the other changes of C>A/T and T>A/C/G showing similar contribution with various flanking bases for each. The etiology for this signature remains unknown, however evidence shows the signature is commonly observed (in terms of mutations per

megabase) in head small cell carcinoma and breast cancers (Alexandrov et al., 2020). Clinical evidence supports PIK3CA's involvement in both breast and head small cell carcinomas, among others (Ligresti et al., 2009). The fourth and fifth highest features listed correspond to mutation descriptions of coding silent mutations and nonsense mutations, respectively.

Chapter IV.

Discussion

In the rapidly expanding field of cancer genomics, mutation classification is essential for the diagnosis of cancer, as well as the discovery and prescription of personalized therapy options for patients living with the disease. In this research, machine learning was shown to be a viable, efficient, and effective tool when used to address the complex problem of variant classification in cancer. Based on the observed results of this research, several factors should be considered to improve the efficacy of future machine learning applications in cancer.

Feature Selection and Creation

For supervised machine learning implementations, feature selection is an essential step toward developing a robust model. The need for thoughtful feature creation was clear in this research, and enrichment of the feature set was essential for each of the classifiers constructed. In addition to the default COSMIC mutation records, information from other reputable public databases such as NCBI's dbSNP database, the Pfam database, and COSMIC mutation signatures was incorporated. By extracting key attributes from these external datasets and supplementing the existing records, new relationships between the data could be discovered during model training. The use of external data also enabled further extrapolation to create linkage between datasets, allowing features to be transformed or encoded more thoughtfully. Once the features

predicted to be of high importance could be structured and encoded systematically, the entire aggregate dataset could be constructed in real-time to train the classifier of interest. This templatizing of features based on available data may enable greater reproducibility in development and refinement of a machine learning approach to cancer variant classification.

Mutation Functional Effect

This research showed that functional effects are often a strong predictor of pathogenic or benign status for a variant with unknown clinical significance. For tumor suppressor genes, all of implementations studied reinforced the relevance of truncating alterations, whether nonsense point mutations or deleterious frameshift truncations. Conversely, a functional status of missense was repeatedly ranked higher for oncogenes. Taken together, these observations suggest that missense mutations may be a more common mechanism of activation for oncogenes, while deleterious mutations are generally expected to hinder the expression potential for genes which harbor them. The functional effect feature can be further refined in future adaptations to predict the tolerability of different amino acid changes in the protein.

Relevance of dbSNP Data

A notable finding was the relevance of dbSNP data and its impact on model performance. Data incorporated from dbSNP included not only dbSNP identifiers (which denote reported germline mutations), but also dbSNP pathogenicity prediction. This extra information proved valuable for each machine learning classifier, signifying that somatic-germline information from dbSNP is an essential component of any machine learning

implementation. As sequencing becomes more accessible to patients and providers, this database is expected to grow. Discovery of new SNPs is expected to improve understanding of the role of SNPs in cancer development and cancer predisposition.

Genomic Positions and Recurrence

The features for genome position and the prevalence of each mutation (via the duplicate entry feature mapping) were informative in several machine learning model implementations. The genome position of each mutation was simplified and divided into one of 10 bins for each gene to prevent any classifier from over-weighting this feature. The combination of mutation genome position and presence of duplicate entries strongly support the notion of genomic hotspots in genes, a proposition which can enable rapid estimations of variant classification when few other features or descriptors are available. Additionally, the presence of duplicate entries suggests that population prevalence is a meaningful way to estimate relevance of a mutation in cancer, either for a common germline SNP or for a canonical disease-specific mutation. The emergence of such common somatic alterations in different patient populations and different disease subtypes may readily lend itself toward classification of otherwise uncharacterized mutations.

Utility of COSMIC Signatures

The various COSMIC signatures incorporated were ranked highly in different classification approaches for both tumor suppressors and oncogenes. Although widespread discussion of genomic mutation signatures began in the last few decades (Brash et al., 1991), the growing list of recognized mutation signatures continues to

provide insight into the influences of environmental conditions and of treatment on patient tumor progression. The genomic signatures weighted highly by the various machine learning implementations often had a clear link to diseases in which the gene in question was frequently mutated or treatment side effects related to standard of care for that disease or gene. Other genomic signatures represented a novel finding with unclear implications, possibly due to low evidence to support the impetus of the signature. Such signatures with unknown etiology were especially prevalent in the oncogenes studied which represent future research opportunities. Still, genomic signatures provided unique perspective on the predicted relevance of certain mutations in cancer, and for some genes may be highly important considerations for an effective machine learning classifier.

Selecting the Optimal Algorithm

In this research, three different ensemble-based learning approaches were implemented, including the random forest classifier, the extremely randomized trees classifier, and the adaboost classifier. As demonstrated by this work, the choice of classifier can vary greatly depending on the application at hand and the data available for training and classification.

The random forest classifier was consistently high performing for each of the different genes tested, a testament to the robustness and reliability of this algorithm. This algorithm has proven utility in the field of cancer variant classification as evident by prior research (Carter et al., 2009; Kaminker et al., 2007). The extra-trees classifier operated as an extension of the random forest, introducing additional randomness in the training process (Geurts et al., 2006). Despite this increased randomness, performance remained

nearly identical to the random forest classifier for each gene, both in terms of standard performance measures and when examining the top features selected. In some instances, the extended randomness appeared to worsen performance, albeit only slightly. It may be valuable to compare the random forest and extra-trees performance with the dataset being studied to determine the necessity of increased randomness.

The adaboost classifier employed a different method than either random forest or extra-trees models, using a weighted boosting approach instead of a set of random decision trees (Freund & Schapire, 1997). Overall, the adaboost classifier demonstrated reduced precision, recall, and F1-score when compared to both random forest and extra-trees models trained on the same dataset. Despite the lower performance metrics overall, the adaboost models consistently identified novel features not ranked as favorably in the other models used. In some implementations, these extra features described COSMIC mutational signatures ranked comparatively lower in the other implementations, though with etiology often supporting the ranking which they were assigned and their perceived relevance in cancer. The uniqueness of adaboost feature lists often compensated for the sub-par performance of the model, offering new perspectives not captured in either of the other implementations. For this reason, the adaboost classifier provided additional areas of interest that future research may find valuable, especially in genes not well studied.

Research Limitations

Several limitations influenced the scope of this research and are necessary acknowledgements for future study. The databases used may have limited the overall utility of the algorithms developed. Aggregate public databases such as COSMIC and

dbSNP provide valuable information, and accessibility was a primary focus in this research. However, larger uniformly curated datasets may be optimal for such variant analyses. Curated databases are expected to control for artifacts more robustly and will also enable more rapid feature design and streamlined implementation.

The design of features in this study was also a limiting factor. The feature set was designed based on publicly available databases and the use of proprietary algorithms or expanded feature sets was not within scope. Additionally, institutions gathering their own data directly from sequenced tumor tissue are expected to have access to various other performance metrics such as sequencing depth, variant allele frequency, and matched normal tissue data, among others. The feature set of future implementations can then be enriched with additional relevant descriptors to further strengthen the classification performance.

The methods used emphasized single gene datasets, with independent classifiers developed for each gene. The decision to utilize focused classifiers based on single gene data was made to highlight the variability in variant classification based on the gene and other contributing factors. Due to resource constraints, a more robust universal classifier on the entire COSMIC dataset was not possible in this research. However, a classifier that can evaluate different mutations for various genes will likely identify different feature importance rankings which can be informative to researchers examining specific problems.

Another limitation was the challenge of determining a ground truth for the classifiers designed and implemented. Information about the samples from which the mutation records were gathered was not known, and thus clinical context about the

patients in question was unknown at the time of research. Information related to patient disease history and other comorbidities will further inform pathogenicity prediction, though this was not possible in this research. In a real-world setting, such information will be essential to evaluate the business use case and approved use of a machine learning classifier where direct patient care or clinical trial enrollment is concerned.

Finally, simplification of the classification problem also had implications for clinical utility of the research outcomes. Several mutation subtypes have been studied and provide unique perspectives on the role of mutations in the cancer genome (Stratton et al., 2009). In this research, the target class was simplified into a binary classification problem, characterizing pathogenic driver mutations and benign mutations which were predicted to be non-functional. This classification does not allow for more nuanced reporting of mutation subtypes, such as passenger mutations, germline mutations, or silent non-coding mutations (Stratton et al., 2009).

Future Research Opportunities

Future research can expand upon the methodologies described herein in several ways. Consensus building from a multi-model approach to machine learning has shown promise (Xiao et al., 2018) and may improve outcomes in real-world applications. Additionally, advancements in the field of three-dimensional protein modeling have enormous potential for improved accuracy in machine learning models. Such predictions of protein structure can enable more complete understanding of the tolerability of mutations in cancer and the impact on protein-protein interactions. (Chevalier et al.,

2017). The integration of protein structure and tolerability prediction into machine learning classifiers may therefore strengthen their performance and application.

The inclusion of genomic data from various sources, including single cell sequencing, is also predicted to enhance future research endeavors. Single-cell sequencing has the potential to resolve observed cellular heterogeneity in sequenced data (Kim et al., 2021). Incorporation of single-cell data may also offer additional insight into expression patterns, providing a more complete understanding of tolerability and expression of mutated proteins (Kim et al., 2021). Exploration of machine learning in the context of single-cell data and other sources of genomic data is expected to enrich training sets and provide better real-world applications, especially in the context of therapy development.

Research design to enable more robust target classification of mutations is also predicted to improve the utility of classification algorithms in cancer. Clonal hematopoiesis of indeterminant potential, or CHIP, is proving to be an important consideration in modern cancer diagnostics and study (Marnell et al., 2022). Research has demonstrated that CHIP mutations arise from a clonally expanded hematopoietic stem cell and may obfuscate overly simplistic classification frameworks (Marnell et al., 2022). Future research which applies more rigorous classification to such nuanced genomic contexts is predicted to identify therapeutic targets with higher confidence compared to the binary classification approach utilized in this research.

Conclusions

It is apparent that machine learning is a powerful tool for bioinformaticians navigating the growing collections of complex cancer genomics data. The ability to rapidly instantiate new learning algorithms and continue to train and improve upon the models used provides longevity to machine learning solutions as the collective understanding of cancer increases. The random forest algorithm represents a consistently serviceable option for variant classification applications, both for its ease of interpretability and study as well as the results it produces for different genes and different datasets. An approach utilizing different modeling workflows may elucidate meaningful features and directions for future research, as demonstrated herein using the extra-trees and adaboost classifier alternatives. Beyond these classifiers, several others may be used to adequate effect depending on the data and the hypothesis being tested. By incorporating custom features, protein domain conservation, dbSNP data, and mutational signatures, meaningful predictions can be made about the pathogenic or benign status of a novel mutation. Through use of machine learning to predict pathogenicity of a mutation in cancer, new mutations of interest can be discovered, and new treatments may emerge for patients living with the disease. Such research can continue to provide new opportunities for improved patient health and survival and will lead to improved patient quality of life.

Appendix 1.

Tables

The following tables describe the top features for each of the models implemented. Review and comparison of the top five features between models for each gene can be found in Chapter III: Results.

Table 1. APC Random Forest Feature Rankings

Rank	Features	Relative Importance
1	mutation_description_substitution – nonsense	0.079
2	mutation_description_deletion – frameshift	0.058
3	snp_unknown	0.058
4	mutation_description_substitution – missense	0.054
5	dbsnp_clin_sig_uncertain-significance	0.021
6	snp_n	0.020
7	sbs94	0.019
8	sbs19	0.017
9	snp_y	0.015
10	mutation_description_insertion – frameshift	0.014
11	sbs30	0.014
12	sbs16	0.013
13	sbs89	0.012
14	dbsnp_clin_sig_pathogenic	0.012
15	sbs86	0.011

This table lists the feature rankings for the random forest classifier trained on APC genomic data. Mutation description features describe different functional effects of the mutations. Additionally, each SBS## feature represents a COSMIC mutation signature.

Table 2. APC Extra-Trees Feature Rankings

Rank	Features	Relative Importance
1	mutation_description_substitution - nonsense	0.098
2	mutation_description_deletion - frameshift	0.085
3	snp_unknown	0.066
4	mutation_description_substitution - missense	0.053
5	snp_y	0.028
6	dbsnp_clin_sig_uncertain-significance	0.022
7	snp_n	0.019
8	mutation_description_insertion - frameshift	0.019
9	dbsnp_clin_sig_pathogenic	0.018
10	dbsnp_clin_sig_likely-pathogenic	0.014
11	nearest_domain_unknown	0.012
12	sbs40	0.012
13	sbs6	0.010
14	mutation_genome_bins_0:112737925.0-112802235.0	0.010
15	sbs39	0.010

This table lists feature rankings for the extra-trees classifier trained on APC data. Mutation description features describe different functional effects of the mutations. Additionally, each SBS## feature represents a COSMIC mutation signature. Mutation genome bins represent a specific range of the gene studied on the appropriate chromosome.

Table 3. APC AdaBoost Feature Rankings

Rank	Features	Relative Importance
1	mutation_description_substitution - nonsense	0.075
2	dbsnp_clin_sig_likely-pathogenic	0.050
3	mutation_genome_bins_0:112737925.0-112802235.0	0.025
4	primary_site_large_intestine	0.025
5	mutation_genome_bins_6:>112839769.0-112839942.0	0.025
6	mutation_genome_bins_5:>112839550.0-112839769.0	0.025
7	mutation_genome_bins_4:>112839294.0-112839550.0	0.025
8	mutation_zygosity_hom	0.025
9	mutation_description_substitution - missense	0.025
10	dbsnp_clin_sig_other	0.025
11	primary_site_bone	0.025
12	nearest_domain_low_complexity	0.025
13	mutation_description_substitution - coding silent	0.025
14	nearest_domain_coiled_coil	0.025
15	sbs88	0.025

This table shows feature rankings for the adaboost classifier trained on APC data. Mutation description features describe different functional effects of the mutations. Additionally, each SBS## feature represents a COSMIC mutation signature. Mutation genome bins represent a specific range of the gene studied on the appropriate chromosome. Primary site features describe the tissue of origin of the tumor sample as per COSMIC records.

Table 4. TP53 Random Forest Feature Rankings

Rank	Features	Relative Importance
1	mutation_description_substitution - missense	0.079
2	mutation_description_deletion - frameshift	0.069
3	snp_n	0.063
4	snp_unknown	0.063
5	mutation_genome_bins_9:>7675221.0-7687510.0	0.035
6	duplicate_entries_bins_0:1.0-2.0	0.024
7	mutation_somatic_status_variant of unknown origin	0.022
8	duplicate_entries_bins_7:>294.0-743.0	0.018
9	duplicate_entries_bins_1:>2.0-9.0	0.018
10	mutation_genome_bins_0:7668450.0-7673764.0	0.014
11	mutation_genome_bins_8:>7675137.0-7675221.0	0.013
12	mutation_genome_bins_3:>7674191.0-7674222.0	0.013
13	mutation_description_substitution - nonsense	0.012
14	mutation_description_insertion - frameshift	0.012
15	duplicate_entries_bins_4:>46.0-92.0	0.009

This table shows feature rankings for the random forest classifier trained on TP53 genomic data. Mutation description features describe different functional effects of the mutations. Mutation genome bins represent a specific range of the gene studied on the appropriate chromosome. The duplicate entries bins represent balanced bins listing the number of times a specific mutation appears in the COSMIC database.

Table 5. TP53 Extra-Trees Feature Rankings

Rank	Features	Relative Importance
1	mutation_description_substitution - missense	0.094
2	mutation_description_deletion - frameshift	0.069
3	snp_n	0.061
4	snp_unknown	0.052
5	mutation_genome_bins_9:>7675221.0-7687510.0	0.036
6	duplicate_entries_bins_0:1.0-2.0	0.030
7	mutation_somatic_status_variant of unknown origin	0.026
8	duplicate_entries_bins_1:>2.0-9.0	0.020
9	duplicate_entries_bins_7:>294.0-743.0	0.019
10	mutation_description_insertion - frameshift	0.016
11	mutation_genome_bins_3:>7674191.0-7674222.0	0.013
12	mutation_genome_bins_8:>7675137.0-7675221.0	0.012
13	mutation_description_deletion - in frame	0.012
14	mutation_genome_bins_0:7668450.0-7673764.0	0.012
15	mutation_description_substitution - nonsense	0.011

This table shows feature rankings for the extra-trees classifier trained on TP53 genomic data. Mutation description features describe different functional effects of the mutations. Mutation genome bins represent a specific range of the gene studied on the appropriate

chromosome. The duplicate entries bins represent balanced bins listing the number of times a specific mutation appears in the COSMIC database. Additionally, mutation somatic status is a somatic-germline prediction made by COSMIC.

Table 6. TP53 AdaBoost Feature Rankings

Rank	Features	Relative Importance
1	mutation_genome_bins_9:>7675221.0-7687510.0	0.050
2	nearest_domain_unknown	0.050
3	duplicate_entries_bins_7:>294.0-743.0	0.050
4	mutation_somatic_status_variant of unknown origin	0.025
5	mutation_genome_bins_4:>7674222.0-7674263.2	0.025
6	mutation_genome_bins_3:>7674191.0-7674222.0	0.025
7	mutation_genome_bins_2:>7673802.0-7674191.0	0.025
8	mutation_genome_bins_1:>7673764.0-7673802.0	0.025
9	mutation_genome_bins_0:7668450.0-7673764.0	0.025
10	mutation_somatic_status_reported in another cancer sample as somatic	0.025
11	age_66.0	0.025
12	primary_site_meninges	0.025
13	mutation_genome_bins_7:>7675085.0-7675137.0	0.025
14	primary_site_liver	0.025
15	mutation_description_deletion - frameshift	0.025

This table shows feature rankings for the adaboost classifier trained on TP53 genomic data. Mutation description features describe different functional effects of the mutations. Mutation genome bins represent a specific range of the gene studied on the appropriate chromosome. The duplicate entries bins represent balanced bins listing the number of times a specific mutation appears in the COSMIC database. Additionally, mutation somatic status is a somatic-germline prediction made by COSMIC.

Table 7. RB1 Random Forest Feature Rankings

Rank	Features	Relative Importance
1	snp_unknown	0.056
2	mutation_description_substitution - missense	0.054
3	mutation_description_deletion - frameshift	0.042
4	mutation_description_substitution - nonsense	0.036
5	dbsnp_clin_sig_pathogenic	0.026
6	snp_n	0.021
7	nearest_domain_disorder	0.017
8	nearest_domain_unknown	0.016
9	mutation_genome_bins_0:48303732.0-48340625.4	0.013
10	sample_type_ns	0.012
11	snp_y	0.012
12	site_subtype_1_ns	0.011
13	co-occurring_bins_8:>6059.0-12308.5	0.011
14	primary_site_lung	0.010
15	mutation_genome_bins_6:>48411476.0-48453057.2	0.009

This table shows the top 15 feature rankings for the random forest classifier trained on RB1 genomic data. Mutation description features describe different functional effects of the mutations. Mutation genome bins represent a specific range of the gene studied on the appropriate chromosome. The co-occurring bins are balanced bins indicating the total number of co-occurring mutations in COSMIC for a specific tumor.

Table 8. RB1 Extra-Trees Feature Rankings

Rank	Features	Relative Importance
1	snp_unknown	0.062
2	mutation_description_substitution - missense	0.055
3	mutation_description_deletion - frameshift	0.048
4	mutation_description_substitution - nonsense	0.040
5	dbsnp_clin_sig_pathogenic	0.040
6	nearest_domain_unknown	0.025
7	snp_n	0.024
8	nearest_domain_disorder	0.019
9	duplicate_entries_bins_0:1.0-2.0	0.018
10	snp_y	0.014
11	mutation_genome_bins_0:48303732.0-48340625.4	0.013
12	sample_type_ns	0.013
13	mutation_genome_bins_6:>48411476.0-48453057.2	0.011
14	site_subtype_1_ns	0.011
15	co-occurring_bins_8:>6059.0-12308.5	0.011

This table shows the top 15 feature rankings for the extra-trees classifier trained on RB1 genomic data. Mutation description features describe different functional effects of the mutations. Mutation genome bins represent a specific range of the gene studied on the appropriate chromosome. The co-occurring bins are balanced bins indicating the total number of co-occurring mutations in COSMIC for a specific tumor. Additionally, the

duplicate entries bins represent balanced bins listing the number of times a specific mutation appears in the COSMIC database.

Table 9. RB1 AdaBoost Feature Rankings

Rank	Features	Relative Importance
1	dbsnp_clin_sig_pathogenic	0.038
2	co-occurring_bins_9:>12308.5-219194.0	0.025
3	sbs87	0.025
4	co-occurring_bins_8:>6059.0-12308.5	0.025
5	mutation_genome_bins_4:>48376917.0-48380217.0	0.025
6	sbs18	0.025
7	duplicate entries bins 0:1.0-2.0	0.025
8	age_52.58	0.013
9	histology_subtype_1_chronic_lymphocytic_leukaemia-small_lymphocytic_lymphoma	0.013
10	tumour_origin_primary	0.013
11	dbsnp_val_type_by-alfa	0.013
12	age_83.0	0.013
13	age_0.0	0.013
14	age_0.7	0.013
15	histology_subtype_1_sebaceous_carcinoma	0.013

This table shows the top 15 feature rankings for the adaboost classifier trained on RB1 genomic data. Mutation description features describe different functional effects of the mutations. Mutation genome bins represent a specific range of the gene studied on the appropriate chromosome. The co-occurring bins are balanced bins indicating the total number of co-occurring mutations in COSMIC for a specific tumor. Additionally, the duplicate entries bins represent balanced bins listing the number of times a specific mutation appears in the COSMIC database. Features with SBS## nomenclature represent unique COSMIC mutation signatures.

Table 10: EGFR Random Forest Feature Rankings

Rank	Features	Relative Importance
1	mutation_description_substitution - missense	0.110
2	snp_unknown	0.034
3	sbs40	0.033
4	sbs2	0.032
5	sbs16	0.026
6	sbs89	0.026
7	nearest domain unknown	0.023
8	sbs7c	0.022
9	sbs90	0.020
10	sbs7b	0.020
11	sbs93	0.020
12	sbs3	0.019
13	sbs41	0.019
14	sbs33	0.019
15	sbs32	0.018

This table shows the top 15 feature rankings for the random forest classifier trained on EGFR genomic data. Mutation description features describe different functional effects of the mutations. Features with SBS## nomenclature represent unique COSMIC mutation signatures.

Table 11. EGFR Extra-Trees Feature Rankings

Rank	Features	Relative Importance
1	mutation_description_substitution - missense	0.141
2	snp_unknown	0.055
3	snp_n	0.054
4	nearest_domain_disorder	0.038
5	mutation_genome_bins_7:>55191821.0-55191822.0	0.033
6	dbsnp_clin_sig_likely-pathogenic	0.031
7	duplicate_entries_bins_4:>3854.0-10574.0	0.029
8	nearest_domain_unknown	0.022
9	mutation_description_deletion - in frame	0.022
10	sbs5	0.021
11	sbs3	0.017
12	dbsnp_clin_sig_drug-response	0.015
13	duplicate_entries_bins_0:1.0-7.0	0.014
14	nearest_domain_low_complexity	0.014
15	snp_y	0.014

This table shows the top 15 feature rankings for the extra-trees classifier trained on EGFR genomic data. Mutation description features describe different functional effects of the mutations. Mutation genome bins represent a specific range of the gene studied on the appropriate chromosome. Additionally, the duplicate entries bins represent balanced bins listing the number of times a specific mutation appears in the COSMIC database. Features with SBS## nomenclature represent unique COSMIC mutation signatures.

Table 12: EGFR AdaBoost Feature Rankings

Rank	Features	Relative Importance
1	sbs89	0.050
2	mutation_description_substitution - coding silent	0.050
3	sbs17b	0.050
4	mutation_description_substitution - missense	0.025
5	nearest_domain_low_complexity	0.025
6	primary_site_lung	0.025
7	sbs14	0.025
8	histology_subtype_1_astrocytoma_grade_iv	0.025
9	sbs90	0.025
10	age_69.0	0.025
11	histology_subtype_1_small_cell_carcinoma	0.025
12	histology_subtype_1_large_cell_neuroendocrine_carcinoma	0.025
13	dbsnp_clin_sig_benign	0.025
14	sbs32	0.025
15	sample_type_surgery-fixed	0.025

This table shows the top 15 feature rankings for the adaboost classifier trained on EGFR genomic data. Mutation description features describe different functional effects of the mutations. Features with SBS## nomenclature represent unique COSMIC mutation signatures.

Table 13. ERBB2 Random Forest Feature Rankings

Rank	Features	Relative Importance
1	mutation_description_substitution - missense	0.095
2	nearest_domain_unknown	0.051
3	snp_unknown	0.042
4	mutation_description_insertion - in frame	0.024
5	dbsnp_clin_sig_likely-pathogenic	0.024
6	duplicate_entries_bins_0:1.0-2.0	0.021
7	duplicate_entries_bins_4:>47.0-134.0	0.018
8	primary_site_lung	0.017
9	nearest_domain_disorder	0.017
10	nearest_domain_transmembrane	0.016
11	snp_n	0.015
12	dbsnp_clin_sig_pathogenic	0.013
13	mutation_description_substitution - coding silent	0.012
14	mutation_genome_bins_9:>39726903.0-39728659.0	0.012
15	dbsnp_clin_sig_unknown	0.010

This table shows the top 15 feature rankings for the random forest classifier trained on ERBB2 genomic data. Mutation description features describe different functional effects of the mutations. Mutation genome bins represent a specific range of the gene studied on the appropriate chromosome. Additionally, the duplicate entries bins represent balanced bins listing the number of times a specific mutation appears in the COSMIC database.

Table 14. ERBB2 Extra-Trees Feature Rankings

Rank	Features	Relative Importance
1	mutation_description_substitution - missense	0.090
2	nearest_domain_unknown	0.058
3	snp_unknown	0.045
4	dbsnp_clin_sig_likely-pathogenic	0.028
5	mutation_description_insertion - in frame	0.026
6	duplicate_entries_bins_0:1.0-2.0	0.025
7	primary_site_lung	0.020
8	nearest_domain_disorder	0.020
9	nearest_domain_transmembrane	0.019
10	dbsnp_clin_sig_pathogenic	0.016
11	snp_n	0.016
12	duplicate_entries_bins_4:>47.0-134.0	0.016
13	mutation_genome_bins_9:>39726903.0-39728659.0	0.011
14	dbsnp_clin_sig_unknown	0.011
15	mutation_description_substitution - coding silent	0.010

This table shows the top 15 feature rankings for the extra-trees classifier trained on ERBB2 genomic data. Mutation description features describe different functional effects of the mutations. Mutation genome bins represent a specific range of the gene studied on the appropriate chromosome. Additionally, the duplicate entries bins represent balanced bins listing the number of times a specific mutation appears in the COSMIC database.

Table 15. ERBB2 AdaBoost Feature Rankings

Rank	Features	Relative Importance
1	sbs9	0.050
2	sbs8	0.050
3	primary_site_prostate	0.025
4	mutation_description_deletion - frameshift	0.025
5	primary_histology_malignant_melanoma	0.025
6	mutation_genome_bins_5:>39723635.0-39724008.0	0.025
7	age_29.0	0.025
8	age_72.0	0.025
9	age_69.0	0.025
10	nearest_domain_transmembrane	0.025
11	sbs17b	0.025
12	mutation_genome_bins_8:>39725079.0-39726903.0	0.025
13	snp_unknown	0.025
14	nearest_domain_unknown	0.025
15	sample_type_surgery-fixed	0.025

This table shows the top 15 feature rankings for the adaboost classifier trained on ERBB2 genomic data. Mutation description features describe different functional effects of the mutations. Mutation genome bins represent a specific range of the gene studied on the appropriate chromosome. Additionally, features with SBS## nomenclature represent unique COSMIC mutation signatures.

Table 16. PIK3CA Random Forest Feature Rankings

Rank	Features	Relative Importance
1	mutation_description_substitution - missense	0.162
2	nearest_domain_unknown	0.121
3	dbsnp_clin_sig_likely-pathogenic	0.064
4	dbsnp_clin_sig_pathogenic	0.056
5	primary_site_meninges	0.037
6	snp_unknown	0.036
7	duplicate_entries_bins_0:1.0-10.0	0.036
8	dbsnp_clin_sig_unknown	0.028
9	nearest_domain_disorder	0.017
10	co-occurring_bins_6:>2387.0-202524.0	0.017
11	primary_histology_meningioma	0.015
12	primary_histology_carcinoma	0.015
13	dbsnp_clin_sig_benign	0.015
14	snp_y	0.013
15	mutation_somatic_status_variant of unknown origin	0.013

This table shows the top 15 feature rankings for the random forest classifier trained on PIK3CA genomic data. Mutation description features describe different functional effects of the mutations. Mutation genome bins represent a specific range of the gene studied on the appropriate chromosome. The co-occurring bins are balanced bins indicating the total number of co-occurring mutations in COSMIC for a specific tumor. Additionally, the duplicate entries bins represent balanced bins listing the number of times a specific mutation appears in the COSMIC database.

Table 17. PIK3CA Extra-Trees Feature Rankings

Rank	Features	Relative Importance
1	nearest_domain_unknown	0.147
2	mutation_description_substitution - missense	0.132
3	duplicate_entries_bins_0:1.0-10.0	0.065
4	dbsnp_clin_sig_likely-pathogenic	0.056
5	snp_unknown	0.039
6	nearest_domain_coiled_coil	0.032
7	dbsnp_clin_sig_unknown	0.029
8	dbsnp clin sig benign	0.029
9	nearest_domain_disorder	0.024
10	mutation_somatic_status_variant of unknown origin	0.017
11	co-occurring_bins_6:>2387.0-202524.0	0.016
12	mutation_description_deletion - in frame	0.016
13	primary_site_meninges	0.014
14	dbsnp_clin_sig_pathogenic-likely-pathogenic	0.012
15	mutation_genome_bins_1:>179199707.2-179210285.0	0.012

This table shows the top 15 feature rankings for the extra-trees classifier trained on PIK3CA genomic data. Mutation description features describe different functional effects of the mutations. Mutation genome bins represent a specific range of the gene studied on the appropriate chromosome. The co-occurring bins are balanced bins indicating the total number of co-occurring mutations in COSMIC for a specific tumor. Additionally,

the duplicate entries bins represent balanced bins listing the number of times a specific mutation appears in the COSMIC database.

Table 18. PIK3CA AdaBoost Feature Rankings

Rank	Features	Relative Importance
1	nearest_domain_unknown	0.075
2	sbs88	0.050
3	sbs39	0.050
4	mutation_description_substitution - coding silent	0.050
5	mutation_description_substitution - nonsense	0.050
6	sbs26	0.050
7	dbsnp_clin_sig_benign	0.050
8	mutation_description_deletion - frameshift	0.025
9	primary_site_skin	0.025
10	mutation_description_substitution - missense	0.025
11	primary_histology_meningioma	0.025
12	dbsnp_clin_sig_likely-pathogenic	0.025
13	histology_subtype_1_ns	0.025
14	dbsnp_clin_sig_pathogenic	0.025
15	dbsnp_val_type_by-alfa	0.025

This table shows the top 15 feature rankings for the adaboost classifier trained on PIK3CA genomic data. Mutation description features describe different functional effects of the mutations. Additionally, features with SBS## nomenclature represent unique COSMIC mutation signatures.

Appendix 2.

Figures

The following figures show receiver operating characteristic (ROC) and precision-recall (PR) curves for each of the models and genes studied in this research. The ROC and PR curves included represent performance output for the optimal model hyperparameter tunings determined during this research.

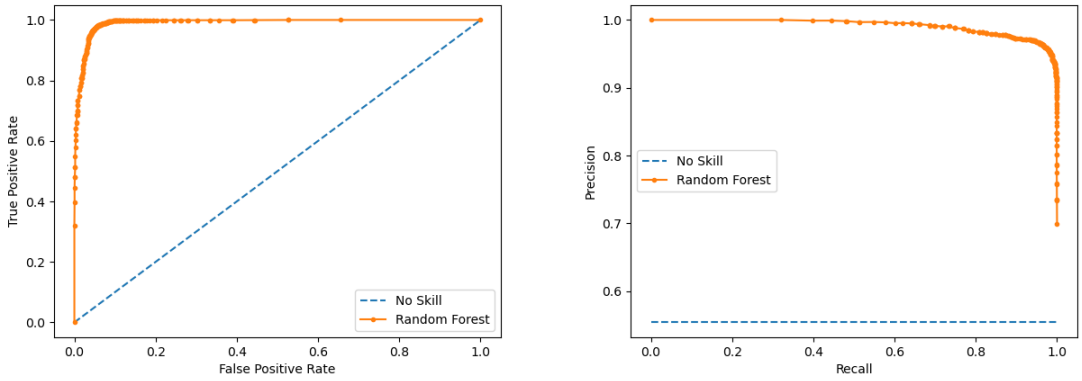

Figure 1. Random Forest ROC and PR Curves - APC

The ROC curve (left) for APC shows the relationship between True Positive and False Positive rates for the pathogenic target class using the random forest algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for APC shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

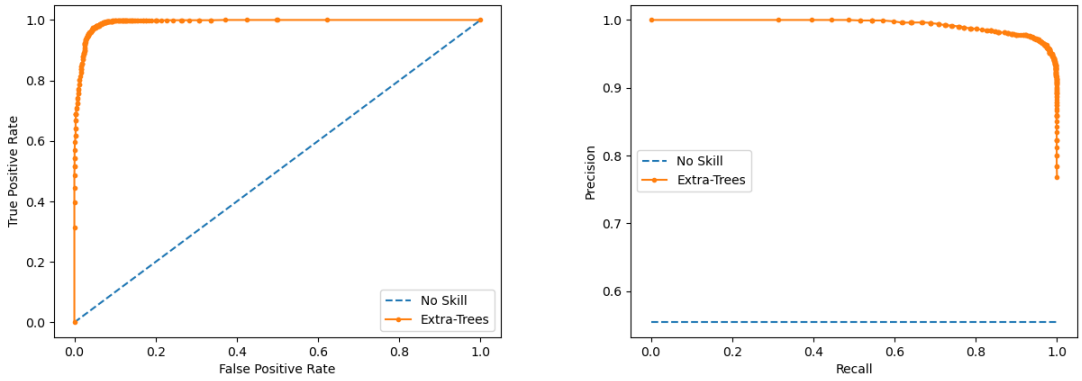

Figure 2. Extra-Trees ROC and PR Curves - APC

The ROC curve (left) for APC shows the relationship between True Positive and False Positive rates for the pathogenic target class using the extremely randomized trees (extra-trees) algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for APC shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

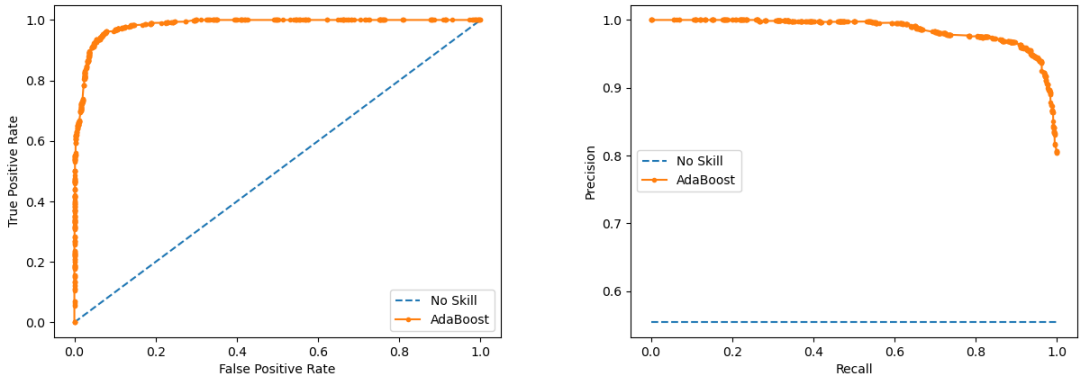

Figure 3. AdaBoost ROC and PR Curves - APC

The ROC curve (left) for APC shows the relationship between True Positive and False Positive rates for the pathogenic target class using the adaboost algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for APC shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

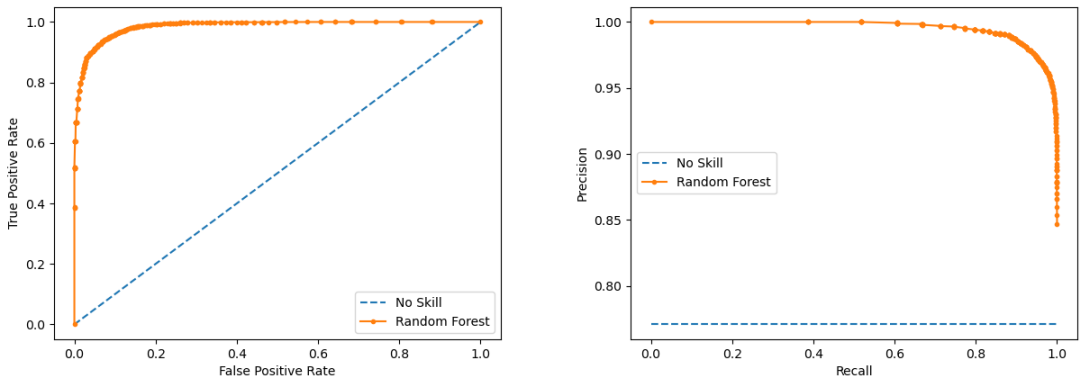

Figure 4. Random Forest ROC and PR Curves – TP53

The ROC curve (left) for TP53 shows the relationship between True Positive and False Positive rates for the pathogenic target class using the random forest algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for TP53 shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

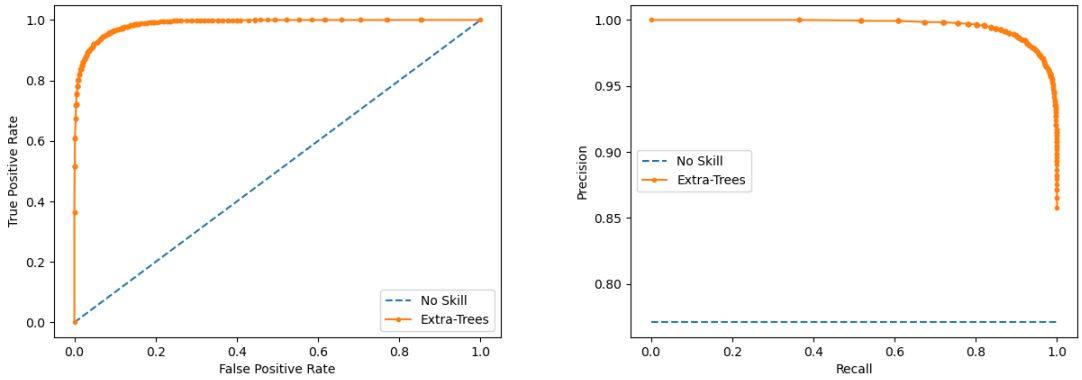

Figure 5. Extra-Trees ROC and PR Curves – TP53

The ROC curve (left) for TP53 shows the relationship between True Positive and False Positive rates for the pathogenic target class using the extremely randomized trees (extra-trees) algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for TP53 shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

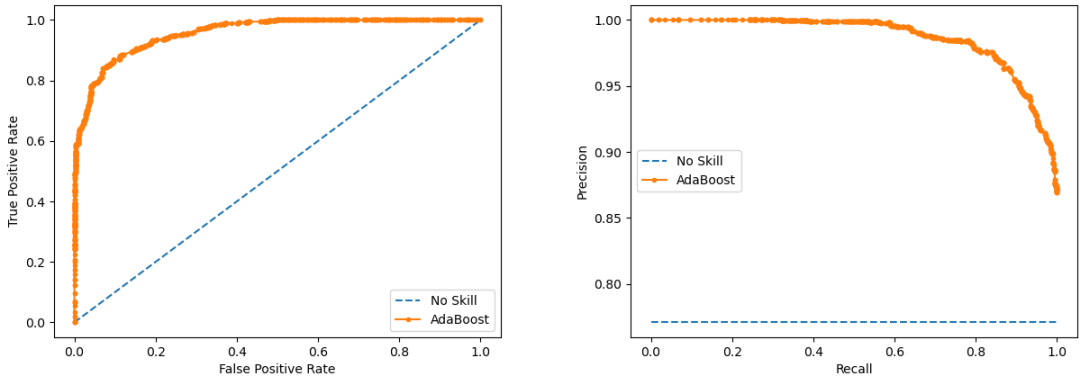

Figure 6. AdaBoost ROC and PR Curves – TP53

The ROC curve (left) for TP53 shows the relationship between True Positive and False Positive rates for the pathogenic target class using the adaboost algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for TP53 shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

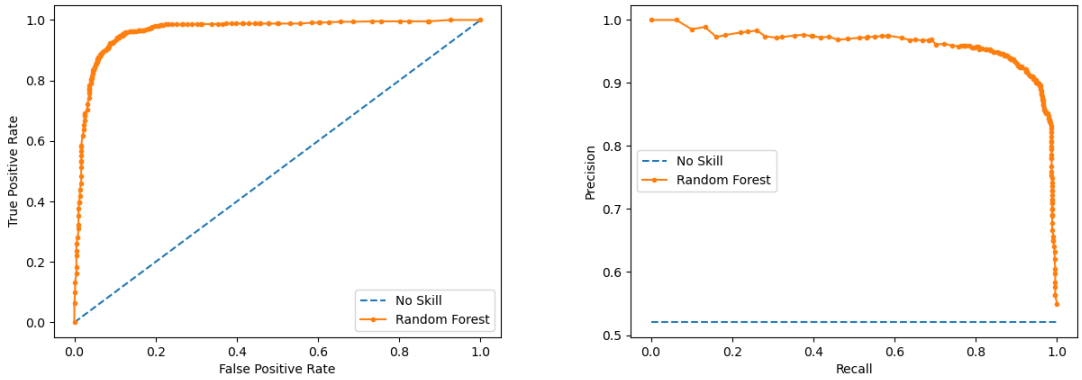

Figure 7. Random Forest ROC and PR Curves – RB1

The ROC curve (left) for RB1 shows the relationship between True Positive and False Positive rates for the pathogenic target class using the random forest algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for RB1 shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

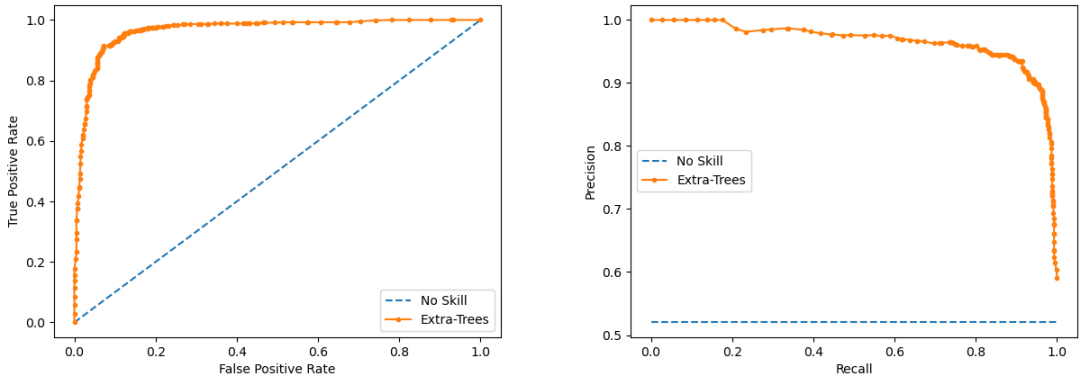

Figure 8. Extra-Trees ROC and PR Curves – RB1

The ROC curve (left) for RB1 shows the relationship between True Positive and False Positive rates for the pathogenic target class using the extremely randomized trees (extra-trees) algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for RB1 shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

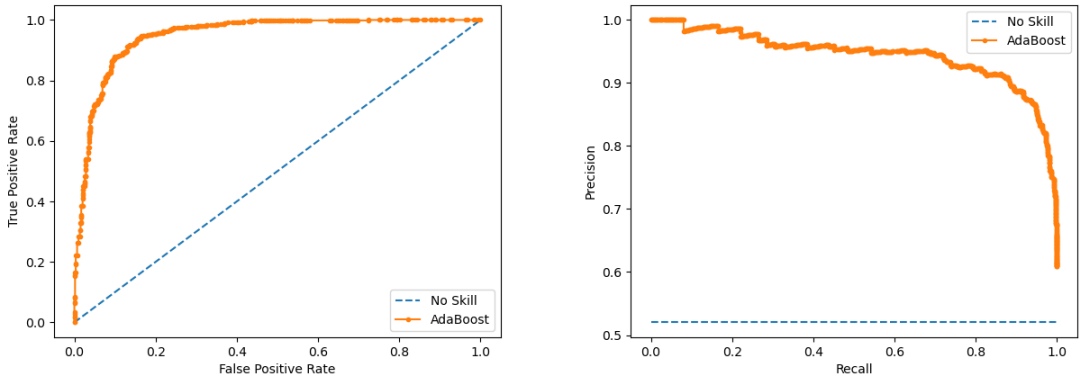

Figure 9. AdaBoost ROC and PR Curves – RB1

The ROC curve (left) for RB1 shows the relationship between True Positive and False Positive rates for the pathogenic target class using the adaboost algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for RB1 shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

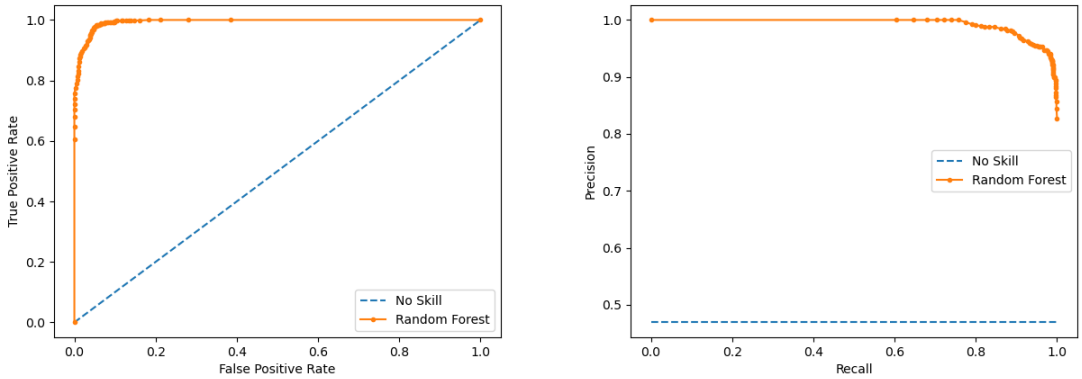

Figure 10. Random Forest ROC and PR Curves – EGFR

The ROC curve (left) for EGFR shows the relationship between True Positive and False Positive rates for the pathogenic target class using the random forest algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for EGFR shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

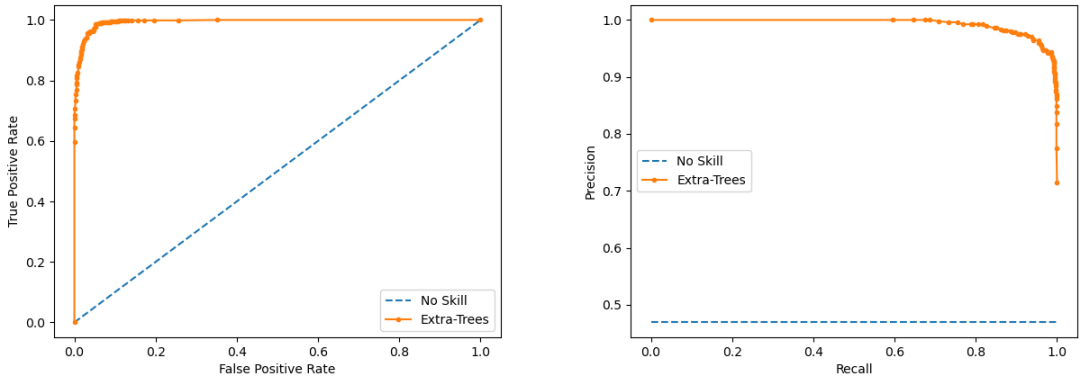

Figure 11. Extra-Trees ROC and PR Curves – EGFR

The ROC curve (left) for EGFR shows the relationship between True Positive and False Positive rates for the pathogenic target class using the extremely randomized trees (extra-trees) algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for EGFR shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

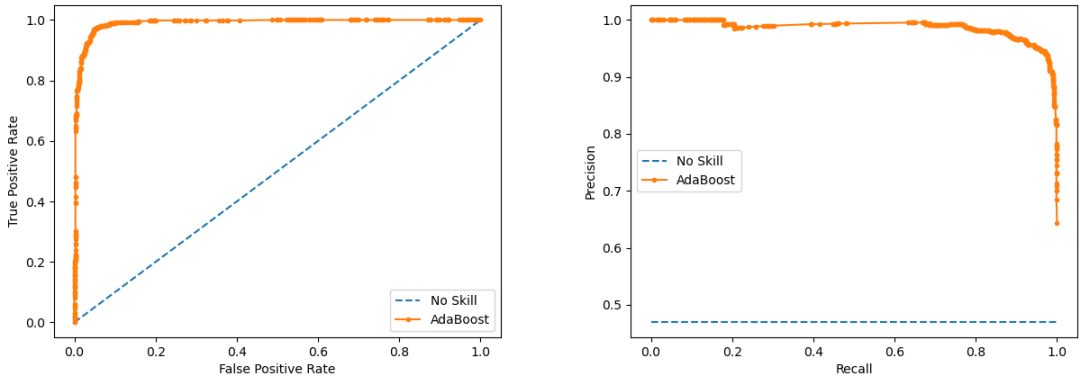

Figure 12. AdaBoost ROC and PR Curves – EGFR

The ROC curve (left) for EGFR shows the relationship between True Positive and False Positive rates for the pathogenic target class using the adaboost algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for EGFR shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

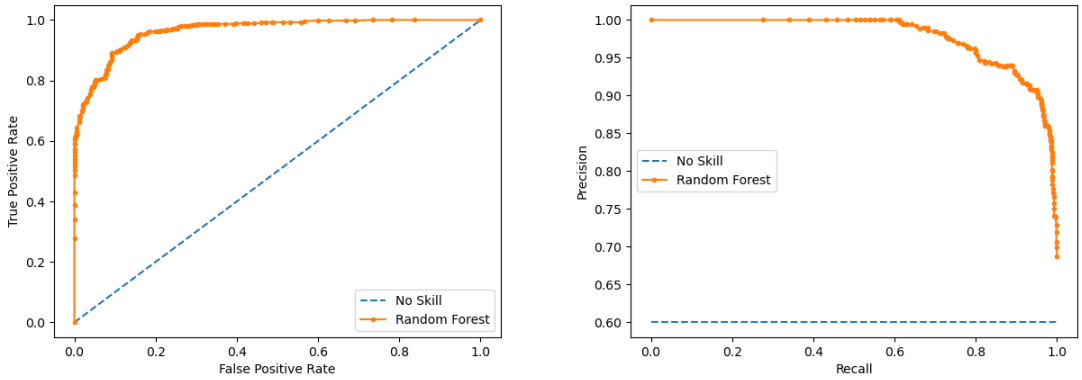

Figure 13. Random Forest ROC and PR Curves – ERBB2

The ROC curve (left) for ERBB2 shows the relationship between True Positive and False Positive rates for the pathogenic target class using the random forest algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for ERBB2 shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

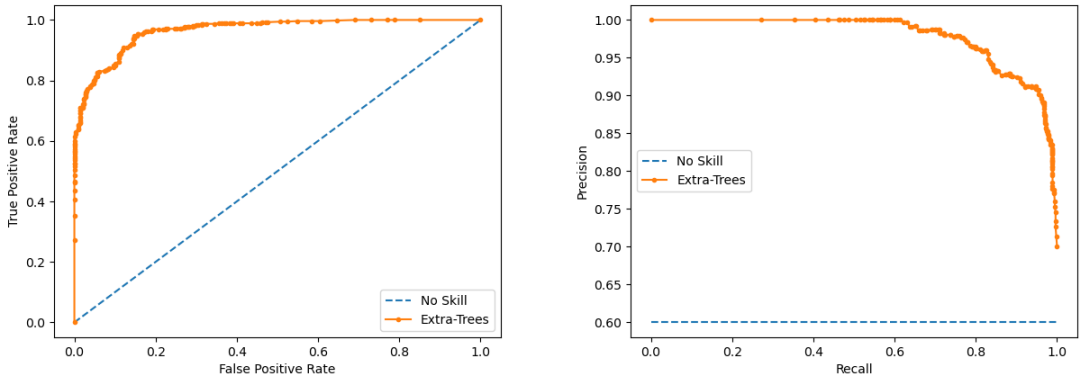

Figure 14. Extra-Trees ROC and PR Curves – ERBB2

The ROC curve (left) for ERBB2 shows the relationship between True Positive and False Positive rates for the pathogenic target class using the extremely randomized trees (extra-trees) algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for ERBB2 shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

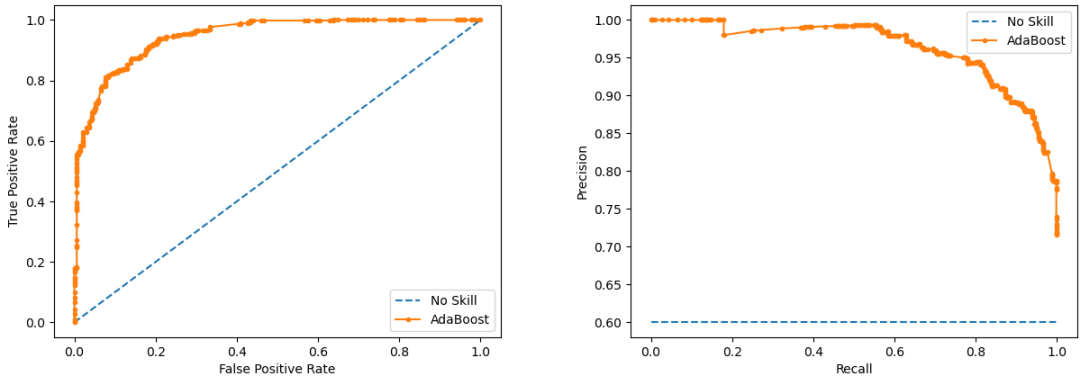

Figure 15. AdaBoost ROC and PR Curves – ERBB2

The ROC curve (left) for ERBB2 shows the relationship between True Positive and False Positive rates for the pathogenic target class using the adaboost algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for ERBB2 shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

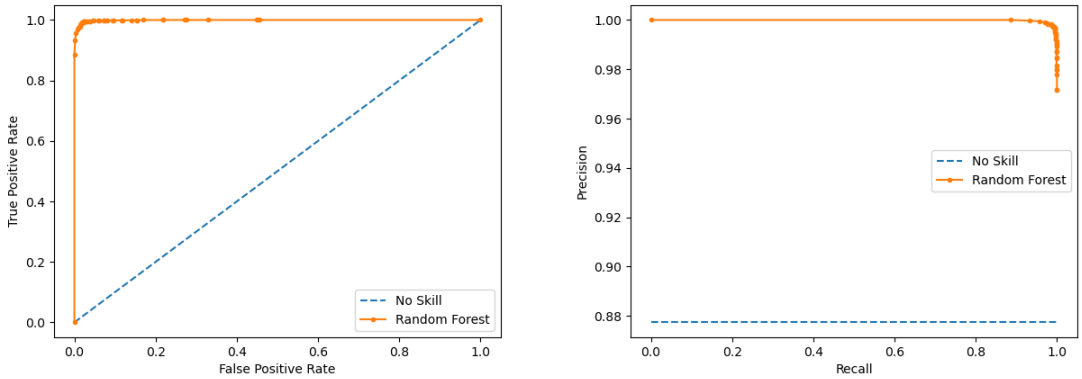

Figure 16. Random Forest ROC and PR Curves – PIK3CA

The ROC curve (left) for PIK3CA shows the relationship between True Positive and False Positive rates for the pathogenic target class using the random forest algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for PIK3CA shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

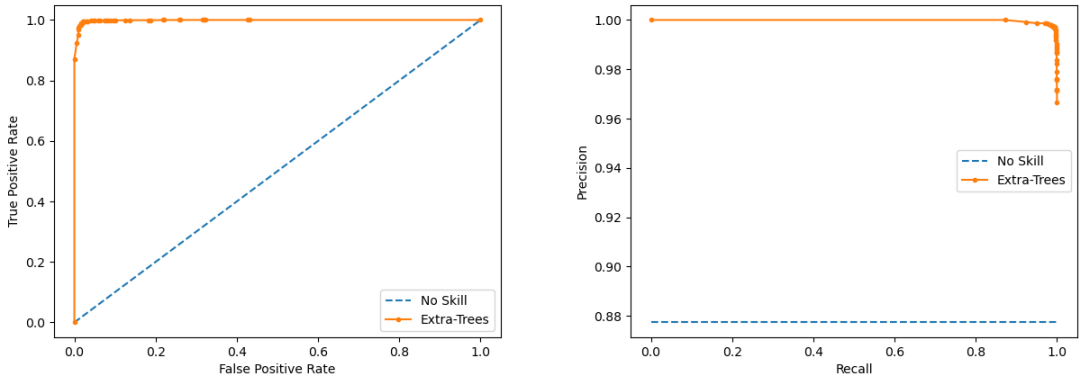

Figure 17. Extra-Trees ROC and PR Curves – PIK3CA

The ROC curve (left) for PIK3CA shows the relationship between True Positive and False Positive rates for the pathogenic target class using extremely randomized trees (extra-trees) algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for PIK3CA shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

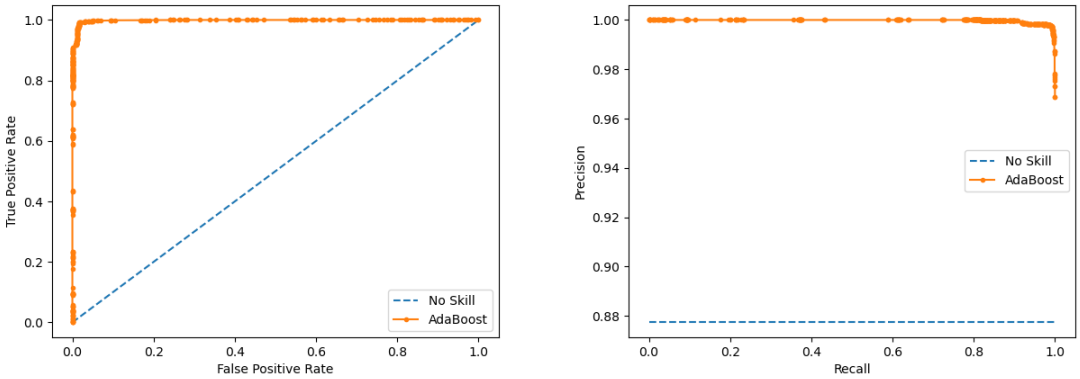

Figure 18. AdaBoost ROC and PR Curves – PIK3CA

The ROC curve (left) for PIK3CA shows the relationship between True Positive and False Positive rates for the pathogenic target class using adaboost algorithm. The plot was generated using output prediction scores and the target class entries for the test set. The PR Curve (right) for PIK3CA shows the relationship between precision (PPV) and recall (sensitivity) for the pathogenic target class. The PR curve was generated using output prediction scores and the target class entries for the test set. Both plots were generated using Matplotlib.

References

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., Getz, G., . . . Stratton, M. R. (2020). The repertoire of mutational signatures in human cancer. Nature, 578(7793), 94–101. https://doi.org/10.1038/s41586-020-1943-3

Barnes, J. L., Zubair, M., John, K., Poirier, M. C., & Martin, F. L. (2018). Carcinogens and DNA damage. *Biochemical Society transactions*, 46(5), 1213–1224. https://doi.org/10.1042/BST20180519

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., & Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behavioural brain research*, 125(1-2), 279–284. https://doi.org/10.1016/s0166-4328(01)00297-2

Bergstrom, E. N., Barnes, M., Martincorena, I., & Alexandrov, L. B. (2020). Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator. *BMC Bioinformatics*, 21(1). https://doi.org/10.1186/s12859-020-03772-3

Bertucci, F., Borie, N., Ginestier, C., Groulet, A., Charafe-Jauffret, E., Adélaïde, J., Geneix, J., Bachelart, L., Finetti, P., Koki, A., Hermitte, F., Hassoun, J., Debono, S., Viens, P., Fert, V., Jacquemier, J., & Birnbaum, D. (2004). Identification and validation of an ERBB2 gene expression signature in breast cancers. *Oncogene*, 23(14), 2564–2575. https://doi.org/10.1038/sj.onc.1207361

Berry, J. L., Polski, A., Cavenee, W. K., Dryja, T. P., Murphree, A. L., & Gallie, B. L. (2019). The RB1 Story: Characterization and Cloning of the First Tumor Suppressor Gene. *Genes*, 10(11), 879. https://doi.org/10.3390/genes10110879

Bister K. (2015). Discovery of oncogenes: The advent of molecular cancer research. *Proceedings of the National Academy of Sciences of the United States of America*, 112(50), 15259–15260. https://doi.org/10.1073/pnas.1521145112

Brady, C. A., & Attardi, L. D. (2010). p53 at a glance. *Journal of cell science*, 123(Pt 15), 2527–2532. https://doi.org/10.1242/jcs.064501

Brash, D. E., Rudolph, J. A., Simon, J. A., Lin, A., McKenna, G. J., Baden, H. P., Halperin, A. J., & Pontén, J. (1991). A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. *Proceedings of the National*

Academy of Sciences of the United States of America, 88(22), 10124–10128. https://doi.org/10.1073/pnas.88.22.10124

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

Buljan, M., & Bateman, A. (2009). The evolution of protein domain families. *Biochemical Society transactions*, 37(Pt 4), 751–755. https://doi.org/10.1042/BST0370751

Carbone A. (2020). Cancer Classification at the Crossroads. *Cancers*, 12(4), 980. https://doi.org/10.3390/cancers12040980

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., & Karchin, R. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer research*, 69(16), 6660–6667. https://doi.org/10.1158/0008-5472.CAN-09-1133

Chevalier, A., Silva, D. A., Rocklin, G. J., Hicks, D. R., Vergara, R., Murapa, P., Bernard, S. M., Zhang, L., Lam, K. H., Yao, G., Bahl, C. D., Miyashita, S. I., Goreshnik, I., Fuller, J. T., Koday, M. T., Jenkins, C. M., Colvin, T., Carter, L., Bohn, A., Bryan, C. M., … Baker, D. (2017). Massively parallel de novo protein design for targeted therapeutics. *Nature*, 550(7674), 74–79. https://doi.org/10.1038/nature23912

Chinnam, M., & Goodrich, D. W. (2011). RB1, development, and cancer. Current topics in developmental biology, 94, 129–169. https://doi.org/10.1016/B978-0-12-380916-2.00005-X

Edwards, G. (2020, January 28). *Machine Learning | An Introduction - Towards Data Science*. Medium. Retrieved February 26, 2021, from https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0

Dean, L., & Kane, M. (2015). Trastuzumab Therapy and ERBB2 Genotype. In V. M. Pratt (Eds.) et. al., *Medical Genetics Summaries*. National Center for Biotechnology Information (US).

Deng, N., Zhou, H., Fan, H., & Yuan, Y. (2017). Single nucleotide polymorphisms and cancer susceptibility. *Oncotarget*, 8(66), 110635–110649. https://doi.org/10.18632/oncotarget.22372

Di Lonardo, A., Nasi, S., & Pulciani, S. (2015). Cancer: we should not forget the past. Journal of Cancer, 6(1), 29–39. https://doi.org/10.7150/jca.10336

Dunnen, J. T., & Hong, W. (2019b, September 30). *Sequence Variant Nomenclature*. Human Genome Variation Society. Retrieved August 3, 2021, from https://varnomen.hgvs.org/bg-material/simple/

Fan, R., Zhong, M., Wang, S., Zhang, Y., Andrew, A., Karagas, M., Chen, H., Amos, C. I., Xiong, M., & Moore, J. H. (2011). Entropy-based information gain approaches to detect and to characterize gene-gene and gene-environment interactions/correlations of complex diseases. *Genetic epidemiology*, 35(7), 706–721. https://doi.org/10.1002/gepi.20621

Forbes, S., Clements, J., Dawson, E., Bamford, S., Webb, T., Dogan, A., Flanagan, A., Teague, J., Wooster, R., Futreal, P. A., & Stratton, M. R. (2006). COSMIC 2005. British journal of cancer, 94(2), 318–322. https://doi.org/10.1038/sj.bjc.6602928

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Gao, S., Ye, H., Gerrin, S., Wang, H., Sharma, A., Chen, S., Patnaik, A., Sowalsky, A. G., Voznesensky, O., Han, W., Yu, Z., Mostaghel, E. A., Nelson, P. S., Taplin, M. E., Balk, S. P., & Cai, C. (2016). ErbB2 Signaling Increases Androgen Receptor Expression in Abiraterone-Resistant Prostate Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 22(14), 3672–3682. https://doi.org/10.1158/1078-0432.CCR-15-2309

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R., Creixell, P., Karchin, R., Vazquez, M., Fink, J. L., Kassahn, K. S., Pearson, J. V., Bader, G. D., Boutros, P. C., Muthuswamy, L., Ouellette, B. F., Reimand, J., Linding, R., Shibata, T., Valencia, A., Butler, A., Dronov, S., … International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group (2013). Computational approaches to identify functional genetic variants in cancer genomes. *Nature methods*, 10(8), 723–729. https://doi.org/10.1038/nmeth.2562

Gottesdiener, L. S., O'Connor, S., Busam, K. J., Won, H., Solit, D. B., Hyman, D. M., & Shoushtari, A. N. (2018). Rates of ERBB2 Alterations across Melanoma Subtypes and a Complete Response to Trastuzumab Emtansine in an ERBB2-Amplified Acral Melanoma. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 24(23), 5815–5819. https://doi.org/10.1158/1078-0432.CCR-18-1397

Guha, T., & Malkin, D. (2017). Inherited TP53 Mutations and the Li-Fraumeni Syndrome. *Cold Spring Harbor perspectives in medicine*, 7(4), a026187. https://doi.org/10.1101/cshperspect.a026187

Hankey, W., Frankel, W. L., & Groden, J. (2018). Functions of the APC tumor suppressor protein dependent and independent of canonical WNT signaling: implications for therapeutic targeting. *Cancer metastasis reviews*, 37(1), 159–172. https://doi.org/10.1007/s10555-017-9725-6

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hanson, C. A., & Miller, J. R. (2005). Non-traditional roles for the Adenomatous Polyposis Coli (APC) tumor suppressor protein. *Gene*, 361, 1–12. https://doi.org/10.1016/j.gene.2005.07.024

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hastie, T., Rosset, S., Zhu, J., & Zou, H. (2009). Multi-class AdaBoost. *Statistics and Its Interface*, 2(3), 349–360. https://doi.org/10.4310/sii.2009.v2.n3.a8

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. doi:10.1109/MCSE.2007.55

Jemal, A., Devesa, S. S., Fears, T. R., & Fraumeni, J. F., Jr (2000). Retinoblastoma incidence and sunlight exposure. *British journal of cancer*, 82(11), 1875–1878. https://doi.org/10.1054/bjoc.2000.1215

Kaminker, J. S., Zhang, Y., Waugh, A., Haverty, P. M., Peters, B., Sebisanovic, D., Stinson, J., Forrest, W. F., Bazan, J. F., Seshagiri, S., & Zhang, Z. (2007). Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer research*, 67(2), 465–473. https://doi.org/10.1158/0008-5472.CAN-06-1736

Kamps, R., Brandão, R. D., Bosch, B. J., Paulussen, A. D., Xanthoulea, S., Blok, M. J., & Romano, A. (2017). Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification. *International journal of molecular sciences*, 18(2), 308. https://doi.org/10.3390/ijms18020308

Karakas, B., Bachman, K. E., & Park, B. H. (2006). Mutation of the PIK3CA oncogene in human cancers. *British Journal of Cancer*, 94(4), 455–459. https://doi.org/10.1038/sj.bjc.6602970

Kastenhuber, E. R., & Lowe, S. W. (2017). Putting p53 in Context. *Cell*, 170(6), 1062–1078. https://doi.org/10.1016/j.cell.2017.08.028

Kim, N., Eum, H. H., & Lee, H. O. (2021). Clinical Perspectives of Single-Cell RNA Sequencing. *Biomolecules*, 11(8), 1161. https://doi.org/10.3390/biom11081161

Kitts, A. (2011, February 2). The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation - The NCBI Handbook - NCBI Bookshelf. NCBI. Retrieved August 24, 2021, from https://www.ncbi.nlm.nih.gov/books/NBK21088/

Lee, E. Y., & Muller, W. J. (2010). Oncogenes and tumor suppressor genes. *Cold Spring Harbor perspectives in biology*, 2(10), a003236. https://doi.org/10.1101/cshperspect.a003236

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18(17), 1–5. Opgehaal van http://jmlr.org/papers/v18/16-365

Lesko, A. C., Goss, K. H., & Prosperi, J. R. (2014). Exploiting APC function as a novel cancer therapy. *Current drug targets*, 15(1), 90–102. https://doi.org/10.2174/1389450114666131108155418

Ligresti, G., Militello, L., Steelman, L. S., Cavallaro, A., Basile, F., Nicoletti, F., Stivala, F., McCubrey, J. A., & Libra, M. (2009). PIK3CA mutations in human solid tumors: role in sensitivity to various therapeutic approaches. *Cell cycle (Georgetown, Tex.)*, 8(9), 1352–1358. https://doi.org/10.4161/cc.8.9.8255

Liu, H., Zhang, B., & Sun, Z. (2020). Spectrum of EGFR aberrations and potential clinical implications: insights from integrative pan-cancer analysis. *Cancer communications*, 40(1), 43–59. https://doi.org/10.1002/cac2.12005

Macconaill, L. E., & Garraway, L. A. (2010). Clinical implications of the cancer genome. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28(35), 5219–5228. https://doi.org/10.1200/JCO.2009.27.4944

Mardis, E. R., & Wilson, R. K. (2009). Cancer genome sequencing: a review. *Human molecular genetics*, 18(R2), R163–R168. https://doi.org/10.1093/hmg/ddp396

Marnell, C. S., Bick, A., & Natarajan, P. (2021). Clonal hematopoiesis of indeterminate potential (CHIP): Linking somatic mutations, hematopoiesis, chronic inflammation and cardiovascular disease. *Journal of molecular and cellular cardiology*, 161, 98–105. https://doi.org/10.1016/j.yjmcc.2021.07.004

Martin, S., & Santaguida, S. (2020). Understanding Complexity of Cancer Genomes: Lessons from Errors. *Developmental Cell*, 53(5), 500–502. https://doi.org/10.1016/j.devcel.2020.05.004

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Reds), *Proceedings of the 9th Python in Science Conference* (bll 56–61). doi:10.25080/Majora-92bf1922-00a

Mei, L., & Nave, K. A. (2014). Neuregulin-ERBB signaling in the nervous system and neuropsychiatric diseases. *Neuron*, 83(1), 27–49. https://doi.org/10.1016/j.neuron.2014.06.007

Meyers, N., Gérard, C., Lemaigre, F. P., & Jacquemin, P. (2020). Differential impact of the ERBB receptors EGFR and ERBB2 on the initiation of precursor lesions of pancreatic ductal adenocarcinoma. *Scientific Reports*, 10(1). https://doi.org/10.1038/s41598-020-62106-8

Miller, M., Reznik, E., Gauthier, N., Aksoy, B., Korkut, A., Gao, J., Ciriello, G., Schultz, N., & Sander, C. (2015). Pan-Cancer Analysis of Mutation Hotspots in Protein Domains. *Cell Systems*, 1(3), 197–209. https://doi.org/10.1016/j.cels.2015.08.014

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2020). Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1), 297–304. https://doi.org/10.1093/genetics/156.1.297

NHGRI. (2019, March 13). *DNA Sequencing Costs: Data*. Genome.Gov. Retrieved February 20, 2021, from https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Negro, A., Brar, B. K., & Lee, K. F. (2004). Essential roles of Her2/erbB2 in cardiac development and function. *Recent progress in hormone research*, 59, 1–12. https://doi.org/10.1210/rp.59.1.1

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? *Bioinformatics*, 34(21), 3711–3718. https://doi.org/10.1093/bioinformatics/bty373

Olivier, M., Hollstein, M., & Hainaut, P. (2010). TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harbor perspectives in biology*, 2(1), a001008. https://doi.org/10.1101/cshperspect.a001008

Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. *Journal of global health*, 8(2), 020303. https://doi.org/10.7189/jogh.08.020303

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85), 2825–2830. Opgehaal van http://jmlr.org/papers/v12/pedregosa11a.html

Pfäffle, H. N., Wang, M., Gheorghiu, L., Ferraiolo, N., Greninger, P., Borgmann, K., Settleman, J., Benes, C. H., Sequist, L. V., Zou, L., & Willers, H. (2013). EGFR-

activating mutations correlate with a Fanconi anemia-like cellular phenotype that includes PARP inhibitor sensitivity. *Cancer research*, 73(20), 6254–6263. https://doi.org/10.1158/0008-5472.CAN-13-0044

Phillips, K. A., & Douglas, M. P. (2018). The Global Market for Next-Generation Sequencing Tests Continues Its Torrid Pace. *The Journal of precision medicine*, 4, https://www.thejournalofprecisionmedicine.com/wp-content/uploads/2018/11/Phillips-Online.pdf.

Ponting, C. P., & Hardison, R. C. (2011). What fraction of the human genome is functional?. *Genome research*, 21(11), 1769–1776. https://doi.org/10.1101/gr.116814.110

Rana, J. S., Khan, S. S., Lloyd-Jones, D. M., & Sidney, S. (2021). Changes in Mortality in Top 10 Causes of Death from 2011 to 2018. *Journal of general internal medicine*, 36(8), 2517–2518. https://doi.org/10.1007/s11606-020-06070-z

Rawla, P., Sunkara, T., & Barsouk, A. (2019). Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Przeglad gastroenterologiczny*, 14(2), 89–103. https://doi.org/10.5114/pg.2018.81072

Roberts, S. A., & Gordenin, D. A. (2014). Hypermutation in human cancer genomes: footprints and mechanisms. *Nature Reviews Cancer*, 14(12), 786–800. https://doi.org/10.1038/nrc3816

Saito, M., Momma, T., & Kono, K. (2018). Targeted therapy according to next generation sequencing-based panel sequencing. Fukushima journal of medical science, 64(1), 9–14. https://doi.org/10.5387/fms.2018-02

Samuels, Y., & Waldman, T. (2010). Oncogenic mutations of PIK3CA in human cancers. *Current topics in microbiology and immunology*, 347, 21–41. https://doi.org/10.1007/82_2010_68

Sboner, A., Mu, X., Greenbaum, D., Auerbach, R. K., & Gerstein, M. B. (2011). The real cost of sequencing: higher than you think! *Genome Biology*, 12(8), 125. https://doi.org/10.1186/gb-2011-12-8-125

Sha, D., Jin, Z., Budczies, J., Kluck, K., Stenzinger, A., & Sinicrope, F. A. (2020). Tumor Mutational Burden as a Predictive Biomarker in Solid Tumors. *Cancer discovery*, 10(12), 1808–1825. https://doi.org/10.1158/2159-8290.CD-20-0522

Shaikh, R. (2018, November 9). *Choosing the right Encoding method-Label vs OneHot Encoder*. Medium. Retrieved August 11, 2021, from https://towardsdatascience.com/choosing-the-right-encoding-method-label-vs-onehot-encoder-a4434493149b

Shastry B. S. (2009). SNPs: impact on gene function and phenotype. *Methods in molecular biology (Clifton, N.J.)*, 578, 3–22. https://doi.org/10.1007/978-1-60327-411-1_1

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., Day, I. N., & Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human mutation*, 34(1), 57–65. https://doi.org/10.1002/humu.22225

Shyr, D., & Liu, Q. (2013). Next generation sequencing in cancer research and clinical application. *Biological procedures online*, 15(1), 4. https://doi.org/10.1186/1480-9222-15-4

Sidey-Gibbons, J., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC medical research methodology*, 19(1), 64. https://doi.org/10.1186/s12874-019-0681-4

Sigismund, S., Avanzato, D., & Lanzetti, L. (2018). Emerging functions of the EGFR in cancer. *Molecular oncology*, 12(1), 3–20. https://doi.org/10.1002/1878-0261.12155

Soliman, S. E., D'Silva, C. N., Dimaras, H., Dzneladze, I., Chan, H., & Gallie, B. L. (2018). Clinical and genetic associations for carboplatin-related ototoxicity in children treated for retinoblastoma: A retrospective noncomparative single-institute experience. *Pediatric Blood & Cancer*, 65(5), e26931. https://doi.org/10.1002/pbc.26931

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, 458(7239), 719–724. https://doi.org/10.1038/nature07943

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2018). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

Temko, D., Tomlinson, I. P. M., Severini, S., Schuster-Böckler, B., & Graham, T. A. (2018). The effects of mutational processes and selection on driver mutations across cancer types. *Nature Communications*, 9(1), 1857. https://doi.org/10.1038/s41467-018-04208-6

Team, T.P.D. (2020). pandas-dev/pandas: Pandas (Version latest). doi:10.5281/zenodo.3509134

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

Voldborg, B. R., Damstrup, L., Spang-Thomsen, M., & Poulsen, H. S. (1997). Epidermal growth factor receptor (EGFR) and EGFR mutations, function and possible role in clinical trials. *Annals of Oncology*, 8(12), 1197–1206. https://doi.org/10.1023/a:1008209720526

Wood, D. E., White, J. R., Georgiadis, A., Van Emburgh, B., Parpart-Li, S., Mitchell, J., Anagnostou, V., Niknafs, N., Karchin, R., Papp, E., McCord, C., LoVerso, P., Riley, D., Diaz, L. A., Jr, Jones, S., Sausen, M., Velculescu, V. E., & Angiuoli, S. V. (2018). A machine learning approach for somatic mutation discovery. *Science translational medicine*, 10(457), eaar7939. https://doi.org/10.1126/scitranslmed.aar7939

Xiao, Y., Wu, J., Lin, Z., & Zhao, X. (2018). A deep learning-based multi-model ensemble method for cancer prediction. *Computer methods and programs in biomedicine*, 153, 1–9. https://doi.org/10.1016/j.cmpb.2017.09.005

Xu, M. J., Johnson, D. E., & Grandis, J. R. (2017). EGFR-targeted therapies in the post-genomic era. *Cancer metastasis reviews*, 36(3), 463–473. https://doi.org/10.1007/s10555-017-9687-8

Yanık, Ö., Gündüz, K., Yavuz, K., Taçyıldız, N., & Ünal, E. (2015). Chemotherapy in Retinoblastoma: Current Approaches. *Turkish Journal of Ophthalmology*, 45(6), 259–267. https://doi.org/10.4274/tjo.06888

Yao, Y., & Dai, W. (2014). Genomic Instability and Cancer. *Journal of carcinogenesis & mutagenesis*, 5, 1000165. https://doi.org/10.4172/2157-2518.1000165

Yiu, T. (2019, June 12). *Understanding Random Forest - Towards Data Science*. Medium. Retrieved September 15, 2021, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Young, L. C., Keuling, A. M., Lai, R., Nation, P. N., Tron, V. A., & Andrew, S. E. (2007). The associated contributions of p53 and the DNA mismatch repair protein Msh6 to spontaneous tumorigenesis. *Carcinogenesis*, 28(10), 2131–2138. https://doi.org/10.1093/carcin/bgm153

Yu, N., Li, Z., & Yu, Z. (2018). Survey on encoding schemes for genomic data representation and feature learning—from signal processing to machine learning. *Big Data Mining and Analytics*, 1(3), 191–210. https://doi.org/10.26599/bdma.2018.9020018

Zadeh, G., Karimi, S., & Aldape, K. D. (2016). PIK3CA mutations in meningioma. *Neuro-oncology*, 18(5), 603–604. https://doi.org/10.1093/neuonc/now029

Zhang, L., & Shay, J. W. (2017). Multiple Roles of APC and its Therapeutic Implications in Colorectal Cancer. *Journal of the National Cancer Institute*, 109(8), djw332. https://doi.org/10.1093/jnci/djw332