UNIVERSITY OF CALIFORNIA,
IRVINE

Cancer Risk Determination through Chromosomal Scale Length Variations of Germline DNA

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biomedical Engineering

by

Charmeine Shumeng Ko

Dissertation Committee:

Associate Professor James P. Brody, Chair
Professor Elliot Botvinick
Associate Professor Timothy Downing

2023

Dedication

To my mother

for her unfailing love and encouragement to strive for excellence,

my sister

for inspiration and unwavering companionship,

my friends

for their support,

to the Leu family

for their incredible strength and faith during Aunt Wendy’s battle against cancer

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my sincere thanks and deepest appreciation to my committee chair, Dr. James P. Brody, whose guidance and support have been profoundly fundamental to this thesis. He always patiently listens to and encourages me to explore my ideas. His scientific insight inspires me to tackle the multifaceted topic critically. This work would not be possible without him.

I would like to thank my committee member, Dr. Elliot Botvinick. His passion for research and innovation is truly inspirational, motivating me to consider a broader scope and application of engineering techniques.

I would like to thank my committee member, Dr. Timothy Downing. His expertise and dedication to scientific research urge me to examine my work analytically.

Vita

Charmeine Shumeng Ko

2017	B.S. in Biochemistry and Cell Biology,
	University of California, San Diego
2018-20	Graduate Researcher & Teaching Assistant, Department of Biomedical Engineering
2020	M.S. in Biomedical Engineering,
	University of California, Irvine
2020-2022	Graduate Researcher & Teaching Assistant, Department of Biomedical Engineering
2023	Ph.D. in Biomedical Engineering,
	University of California, Irvine

FIELD OF STUDY

Computational Genomics and Machine Learning in Complex Human Diseases

PUBLICATIONS

A genetic risk score for glioblastoma multiforme based on copy number variations – Ko, Brody. *Cancer Treatment and Research Communications.* 2021. https://doi.org/10.1016/j.ctarc.2021.100352

A genetic risk score using human chromosomal-scale length variation can predict breast cancer – Ko. Brody. *In Review*. 2022. https://10.21203/rs.3.rs-1999108/v1

Abstract of The Dissertation

Cancer Risk Determination through Chromosomal Scale Length Variation of Germline DNA

By

Charmeine Shumeng Ko

Doctor of Philosophy in Biomedical Engineering

University of California, Irvine, 2023

Professor James P. Brody, Chair

Cancer is a complex disease with significant genetic components. Previous efforts to uncover the genetic basis of carcinogenesis tend to focus on linear combinations of single genetic mutations, ignoring the complex non-linear network of interactions that are known to regulate cellular processes. The goal of this line of research is the ability to predict whether a person will develop a specific cancer later in their life.

This study evaluates how well machine learning classification algorithms trained with germline chromosomal scale length variation (CSLV) data from cancer patients can predict whether a person will develop cancer later in life. CSLVs were developed to condense pertinent copy number variation (CNV) information into a smaller number of parameters, allowing the usage of machine learning models.

We investigated cancer risk prediction and diagnosis classification from germline CSLV data alone. Our findings indicate that CSLVs contribute to inherited cancer likelihood through a complicated network interaction. We first tested 33 different types of cancer using the 11,000 patients from the Cancer Genome Atlas (TCGA). Lung squamous cell carcinoma (AUC = 0.69), Glioblastoma multiforme (AUC = 0.78), colon adenocarcinoma (AUC = 0.67), and many others could be differentiated from other cancer types better than random chance.

We also evaluated the method in a second dataset, the UK Biobank. Each cancer type dataset was paired with an age- and gender-matched randomized control set. 125 CSLVs were computed, 4 averages and 1 standard deviation from each of the 22 autosomes and 3 sex chromosomes (X, Y, and XY), to be used as features in the model. The AUC of lung cancer was found to be 0.597, the AUC of brain cancer was 0.567, and the AUC of colorectal cancer was 0.565. These results were comparable to current published risk scores and demonstrate the viability of CSLVs as genetic risk scores for certain cancer types.

Utilizing germline chromosomal scale length variation data from large public databases and machine learning models, we developed a novel and promising method to predict cancer diagnosis. This technique can be further improved and augmented for more clinical relevance, and it can be beneficial in personalized diagnostics and cancer preventive measures.

Introduction

Cancer is a genetic disease, with a significant hereditary component in its development, as demonstrated by past research[1]. There has been accumulating evidence indicating that genetic variation accounts for a considerable portion of susceptibility to cancer[2], the identifications of inherited genetic variations associated with the disease and understanding of how they contribute to cancer biology become a priority in elucidating etiology in cancers[3]. In recent years, the rapid progress in sequencing technologies allows for cheaper and more efficient comprehensive genome analysis, and thus many large genomic databases have been established. With an immense wealth of information available, it becomes increasingly important and complex to develop methods to analyze and utilize the data to draw meaningful conclusions.

Genome-wide association studies (GWAS) are a common approach for investigating the genetic basis of complex diseases such as cancer and mainly focus on single nucleotide polymorphisms (SNPs). GWAS have identified many cancer risk loci locating at non-coding regions of the genome, generally through studies performed on somatic tissues[4,5]. As cancer is a multifactorial disease, it results from an interaction between hereditary and somatic factors. The utilization of somatic samples may introduce environmental factors acquired during the individual's lifetime and complicate the homogeneity of inherited component in cancer pathogeny. Therefore, it is important to separate the two by focusing on the germline genetics.

We are interested in chromosomal scale length variation of germline DNA, which condenses multiple copy number variations (CNVs). CNVs are extensive structural variants in the human genome composed of repeats and deletions. CNVs have been shown to exhibit functional impact on gene expression and are a hallmark of cancer[6]. The discovery and mapping of these genetic variants owes in parts to the development of Next Generation Sequencing (NGS), which has greatly advanced the field of genomic research. The reduced cost and production needs allow for faster whole genome sequencing with greater accuracy and precision, for instance, the human genome can be sequenced within a day using NGS, while the same task performed with Sanger sequencing technology would require over a decade[7]. Many population genomic databases have therefore been built and become publicly available, proving to be tremendously valuable to researchers interested in uncovering the genetic basis of complex diseases, e.g. identification of the hereditary component in cancer development.

This endeavor is further aided by the advent of computational technologies in the past decade. High-throughput sequencing outputs large quantities of data to be processed for further analysis. Cancer samples are complex and heterogeneous, for the disease mechanism involves a multitude of processes that encompass genomic to cellular functions. As the size and intricacy of cancer genomic datasets continue to growth, storing and querying terabytes or even petabytes of data can be immensely challenging to researchers without sufficient computational resources[8]. Therefore, the availability of scalable computing resources, i.e. the “cloud”, is crucial for facilitating rapid and cost-effective data analysis[9].

An important goal of cancer genomic data analysis methods is to transform the wealth of sequencing results into understanding of the relationship between various molecular characteristics of cells. The development of machine learning methods contributes to this effort for its ability to apply complex mathematical calculations to large, complex datasets in an automated fashion to produce predictive models. Understanding the decision-making progress involved in model building would reveal potential biomarkers as therapeutic targets; translating and incorporating the results with clinical data would provide insight for physicians, patients, and researchers. Furthermore, the interaction between the hereditary component and other risk factors may be also studied in depth.

Objective and Specific Aims

Genetic variation has been associated with many complex diseases, i.e. cancer, and copy number variations (CNV) account for considerable amount of variability in human genome. In addition to genetic features, many factors contribute to the complexity of cancer development, with age as a major element[10]. In this study, we hypothesize that the epistatic interactions between germline genetic variants create a network effect that contributes to hereditary cancer risks and cancer onset age. Machine learning models can thus be utilized to predict risks and determine the influence of germline CNVs on cancer incidence age.

Objective 1

The primary objective of the study is to utilize germline CNV information from large, public databases to predict cancer diagnosis. We developed a method to transform CNV data into Chromosomal Scale Length Variation (CSLV) values for dimensionality reduction while preserving the epistatic interaction between CNVs across the genome.

Objective 2

The second objective is to study whether there is an inherited genetic risk in complex diseases such as cancer, and we aim to investigate if the CNVs across an individual's genetic landscape contribute to such hereditary component in disease development.

Objective 3

The third objective is to develop machine learning techniques utilizing CSLV data to predict different cancer diagnoses. These models would predict whether an individual possess higher inherited risk for specific types of cancer. Our models would be built from genomic data from large, public databases and address the non-linear and high-dimensional relationships between multitude of genetic variants. We would employ different machine learning algorithms to compare their performances and predictive powers and determine which option achieves the best results.

Objective 4

The fourth objective is to explore how CSLV features affect cancer prediction and optimize the predictive performance. We would compute different CSLV configurations and investigate the optimal dataset.

Objective 5

The next objective focuses on model interpretation. We would examine how the predictions are made, which variables carry substantial weights in the models, and the importance of different components. This would provide biological insight into the CNV regions that play the most significant role to a specific cancer type.

Objective 6

The final objective is to evaluate the performance of our CSLV predictive models through comparison to published risk scores for the specific cancer types that are computed in UK Biobank.

Background

Next-Generation Sequencing

The rapid development of different next-generation sequencing (NGS) platforms has revolutionized the biological sciences. In the field of genomics research, NGS reduces the high production needs and cost for the comprehensive analysis of genomes, transcriptomes, and interactomes, standardizing inexpensive and robust studies[11].

The advent of NGS allows for systematic study of cancer genomes, prompting various ongoing large-scale cancer genome projects around the world[7]. Some of the common goals that these projects aim to address are more precise cancer diagnosis and classification, increased prognosis accuracy, and identification of mutations for potential drug targets[7], forming the basis for development of personalized cancer treatment.

In comparison to previous sequencing technologies, NGS enables simultaneous identification of multiple modalities of genome alteration, i.e. copy number, mutation, due to deep coverage[12]. However, there exist several limitations of cancer genomics. Cancer genomes are characterized by high degrees of heterogeneity between cancer types and even individuals. Due to the incomplete penetrance of mutations, the identification of a pathogenic mutation in an individual does not necessarily implicate that other members in the subject's family carrying the same mutation will develop cancer. The complication is amplified with NGS because of the larger number of genes and variants being identified[13]. In addition to modification to screening protocol, researchers would benefit from proper statistical tools to distinguish clinically significant features and draw correct conclusions on mutations that are indicative of certain cancer types.

Copy Number Variations

There are many forms of genomic variability, including single nucleotide polymorphisms (SNPs), tandem repeats, transposable elements, structural alterations such as insertions, deletions, and inversions[14]. Copy number variants (CNVs) represents a copy number change involving a DNA fragment that is or greater than 1 kilobases in size[15]. CNVs account for a considerable amount of genetic variation in the human genome.

Population-based surveys have identified thousands of CNVs, and their functional impact has been demonstrated to have dramatic phenotypic consequences from alterations of gene dosage, disruptions in coding sequences and long-range gene regulation[16]. Increased CNVs can be positively or negatively correlated with gene expression levels[17], for instance, deletion of a transcriptional enhancer will repress gene expression. CNVs have also been shown to affect all classes of human disease with genetic basis, such as sporadic, Mendelian, complex and infectious[18]. Studies on CNVs and cancer risks typically focus on identification of single genes and their corresponding CNVs or rare single region CNVs, i.e. those with low population frequency[19,20].

It is important to distinguish somatic CNVs from germline CNVs (gCNVs) in human disease studies. Although both types contain inherited information pertaining to the pathogeny of interest, past research has shown the possibility of CNVs acquisition in somatic tissues through environmental factors[21], which may compromise the homogeneity in somatic CNVs, leading to conclusions drawn from genetic features acquired later in life. Therefore, we focus on gCNVs to properly study the effect of hereditary genetic variations alone on cancer risks and incidence age.

The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) was a cancer genomic program that molecularly characterized over 20000 primary cancer and matched normal samples from 11000 patients, spanning 33 cancer types. It started as a joint effort between the National Cancer Institute and the National Human Genome Research Institute in 2006, culminating in the creation of the Pan-Cancer Atlas[22]. Many accompanying studies provide insights into cancer classification through molecular similarities and genetic differences[23,24], and they further reveal the role of germline genetic variants and somatic mutations in cancer progression[25]. Though the project has ended after 12 years, the data remains publicly available, and the wealth of information has immense potential for new discovery in cancer research.

The 2.5 petabytes of data contain information such as mRNA expression, somatic mutations, DNA methylation, and our target of interest: copy number variation. CNVs are typically measured as segment means, and TCGA defines segment mean as the value of $log_2(\frac{CN}{2})$, where CN is the copy number of a specific segment of the genome. Each segment mean has a unique genomic address consisting of the chromosome number, start base pair position, and end base pair position. To anonymize the data, the TCGA employs a masking process of omitting the Y chromosomes and calculating segment means of large portions of a genome, sometimes spanning an entire chromosome, resulting in two datasets: masked and unmasked. The challenge of navigating and managing the extensive datasets in TCGA is offset by its compatibility with Cloud Service providers such as Cancer Genomic Cloud and Google Cloud Platform.

The Cancer Genome Atlas Pipeline

TCGA was primarily funded by the National Cancer Institute (NCI), and its genomic and clinical datasets were coordinated through NCI's Center for Cancer Genomics (CCG). This project was implemented through a standardized workflow called the Genome Characterization Pipeline, consisting of four major steps: Tissue Collection, Genome Characterization, and Genomic Data Analysis[26].

CCG collected tumor tissues and matched normal blood samples from patient who voluntarily participated in clinical trials and community oncology groups. The majority of the samples were formalin-fixed and paraffin-embedded, and the rest frozen, then sent to CCG's Biospecimen Core Resource (BCR). The Biospecimen Processing Center at Nationwide Children's Hospital, the first component of BCR, curated the samples to meet the rigorous quality standards. The second component of BCR, the Clinical Data Center at Information Management Services, Inc. oversaw informed consents and anonymization of clinical data to safeguard patients' privacy.

The tissues were then sent to the Genome Characterization Centers (GCCs): The Broad Institute that specializes in DNA and performs whole genome and whole exome sequencing, the University of North Carolina that specializes in RNA and performs total RNA sequencing, MD Anderson Cancer Center that specializes in proteins and performs reverse phase protein arrays. CNV data was generated by the Broad Institute[27]. The GCCs generated and sent the outputs, including raw sequencing data, associated metadata, and other characterization data, to the Genomic Data Commons (GDC), which shared the data with the Genomic Data Analysis Network (GDAN).

The GDAN consisted of scientists from 13 institutions across North America. They examined the raw data and utilized genomic characterization techniques to produce novel analyses and publish results in scientific journals. The data generated by the CCG pipeline are publicly available in the GDC for researchers all around to the world.

The CNV pipeline used Affymetrix Genome-Wide Human SNP Array 6.0 data to identify genomic repeats and infer the copy number of these regions based off of GRCh38[28]. It was built onto the TCGA data generated by Birdsuite, an open-source tool set created by the Broad Institute[29]. The data was processed through a circular binary segmentation analysis, which translated noisy intensity measurements into chromosomal regions of equal copy number, resulting in final output files that are segmented into genomic regions with the estimated copy number for each region[30]. These copy number values were further transformed into segment mean values ($log_2(\frac{CN}{2})$)[31] .

UK Biobank

The UK Biobank is one of the most well-established genetic projects and serves a major internal health resource. It was founded by the Wellcome Trust medical charity, the UK Medical Research Council, the UK Department of Health & Social Care, the Scottish Government, the Northwest Regional Development Agency, with additional funding from the Welsh Government, British Heart Foundation, Cancer Research UK, and Diabetes UK[32]. The project is primarily supported by the UK National Health Services.

The UK Biobank was started in 2006 by tracking longitudinally the health outcomes of 500,000 volunteers between the ages of 40 to 69 years old over their lifetimes[32]. It aims to provide important biological samples and environmental exposure data, further constituting a resource on the effects of genetic, environmental, and lifestyle factors on human morbidity, mortality, and health[33]. Genome-wide genotyping data are available for all 500,000 participants in the UK Biobank. The database went live in 2017, making 12 Petabytes of information such as genetic data, imaging and exercise data available to researchers worldwide. The health data of the participants is regularly updated. To date, approximately 89,000 cancer occurrences and 20,000 deaths have been recorded[34].

UK Biobank Pipeline

Blood samples were collected from participants at UK Biobank assessment center and were stored at the UK Biobank facility in Stockport, UK. After DNA extraction, the samples in 96-well plates of 94×50-μl aliquots were sent to Affymetrix Research Services Laboratory for genotyping[35]. During the automated sample retrieval process, special attention was paid to ensure that there was no systematic correlation between experimental units and baseline phenotypes such as age, sex, and ethnicity[36]. The UK Biobank Axiom array was used to genotype around 90% of the 500,000 UK Biobank participants, and the UK BiLEVE array was used to genotype the remaining 10%; the two arrays were very similar with over 95% common marker content[35]. Genomic assays of 820,967 SNPs were conducted, and genome-wide genotyping and imputation was performed by the Big Data Institute of Oxford University[37]. The resulting data are around 2 terabtyes in size and include information such as normal SNP genotyping data, calls, confidences, and intensities.

A wide range of phenotypic information has been collected along with the biological samples. Participants were asked to provide their socio-demographic background, lifestyle, medical history, and physical measures such as blood pressure and arterial stiffness[36]. Physical activity tracking data, monitored from 2013-2014, of 100,000 participants was also recorded. All participants were consented to provide health-related records that indicate death, cancer diagnoses, and hospitalizations.

Machine Learning

Machine learning is one of the major branches of artificial intelligence, for the ability to learn is a basic requirement for any intelligent being. Its history can be traced back as far as 1950s, during which the experimental and theoretical works were inspired by neurophysiological, biological, and psychological research[38]. The development of practical algorithms started to take roots in 1970s, and many were designed to analyze medical datasets[39].

From the early history of machine learning, three major branches emerged from the early development, as outlined in classical works in symbolic learning by Hunt et al[40]., in statistical methods by Nilsson[41], and in neural networks by Rosenblatt[42]. They later gave rise to advanced methods, respectively[43]: inductive learning of symbolic rules such as top down induction of decision trees, pattern recognition methods such as *k*-nearest neighbors, and artificial neural networks such as multilayered feedforward neural network.

Machine learning methods are split into two types: supervised and unsupervised learning. The main difference is that we have prior knowledge of the outcomes for the samples used for supervised learning to learn a function that best approximates the correct output values based on the inputs[44]. Unsupervised learning interprets the natural structure present in the data, without labeled outputs. Therefore, supervised learning is often employed for classification and regression tasks. A standard protocol involves splitting the data into training set and testing set; the former is used for model building, and the latter is used for model evaluation. Since the “correct” output is determined solely from the training data, incorrect or noisy data labels will impact model effectiveness.

Machine Learning Algorithms

Generalized Linear Model

Generalized Linear Models (GLMs) estimate regression models whose outcomes are assumed to follow exponential distributions, which include the Gaussian, Poisson, binomial, and gamma distributions. GLM can be used for regression or classification, depending on the distribution and link function. The H2O suite includes Gaussian regression, Poisson regression, binomial regression/classification, fractional binomial regression, quasibinomial regression, multinomial classification, Gamma regression, ordinal regression, negative binomial regression, and Tweedie distribution[45]. GLM does not require data to be sorted or special handling with imbalanced data.

Regularization is employed in GLM by introducing penalties to prevent overfitting, to reduce variance of the prediction error, and to handle collinearity. Some common techniques are ridge regression and least absolute shrinkage and selection operator (LASSO)[46]. The regularization process involves finding the optimal regularization parameters α and λ, and this is achieved by performing a grid search over α and “lambda search”, a specific type of grid search, over λ. The α parameter handles the distribution between the LASSO and ridge regression penalties, with a value of 1.0 representing LASSO whereas a value of 0.0 representing ridge regression. The λ parameter controls the amount of regularization employed in the model, with a λ of 0.0 denoting that no regularization is applied at all.

Distributed Random Forest

Random forests are an ensemble of independently trained decision trees, and the results of the individual trees are averaged to obtain a more optimized prediction. The training of random forests follows the general techniques of bootstrap aggregation: each tree is built with random sample with replacement. At each terminal node of the tree, a random subset of features is selected to prevent each learner from fixating on the apparently predictive features of the training set and becoming tuned too much to the noises. The resulting trees will be as uncorrelated from each other as possible, increasing generalization of the model. Random forests can contain hundreds or even thousands of trees, and they work well on noisy data. The fundamental principle of random forests is that a large number of uncorrelated trees operating as an ensemble will outperform any of the individual constituent model.

Gradient Boosting Machines

In the field of supervised learning, Gradient Boosted Decision Trees (GBDTs), or Gradient Boosting Machines (GBMs), have been shown to perform exceptionally and have rapidly gained popularity in the data science community[47,48]. The general paradigm of gradient boosting is aggregating weak classifiers to form a strong learner, as such some parameter tuning might be necessary to achieve good results. A training set of known inputs and corresponding outputs is used to find an approximator, built from the sum of weak learners, that minimizes some loss function to gradually step towards best fit[49]. In comparison to Random Forest, another popular tree-based algorithm, GBMs build the decision trees sequentially instead of in a parallel fashion, which results in a large number of trees that may be slow in real-time prediction. The algorithm utilizes the error of prior trees in the creation of subsequent tree; in mathematical terms, the residuals of a given model act as negative gradients to optimize the loss function. The iterative nature of GBMs proves advantageous over other methods, such as Artificial Neural Networks, in handling imbalanced datasets by amplifying the impact of the positive class.

The major limitation of GBM is its tendency for overfitting, that is, the model is tuned too much to the noise instead of the signal and thus performs significantly better in the training set than in the testing set. Therefore, k-fold cross validation is often implemented for the models. The sample is divided into k parts, one of which will be used for testing while the rest for training. The procedure is repeated k times, rotating the testing set. An expected performance metric will be selected to evaluate the results across iterations.

XGBoost

XGBoost is an implementation of gradient boosted decision trees designed to heighten performance and efficiency. It has become one of the most popular machine learning algorithms in recent years. Some features include penalization of trees, a proportional shrinking of leaf nodes, extra randomization parameter, and Newton Boosting, which uses curvature information, i.e. the second derivative, to take a more direct route than gradient descent to minimize a function.

Deep Learning

Deep learning is a subfield of machine learning algorithms based on artificial neural networks that are inspired by the function and structure of brain. When larger neural networks are constructed and trained with more data, their performance continuously increases, as opposed to reaching a plateau like some other machine learning algorithms. In addition to the scalability, another benefit of deep learning is feature learning, the ability of extracting features automatically from raw data. It detects the unknown structure in the input distribution to discover good representations at multiple levels, with high-level learned features defined in terms of low-level features.

The most common implementation of deep learning is a feedforward artificial neural network. It is trained with stochastic gradient descent using back-propagation. The network typically consists of many perceptrons organized into many hidden layers. Each perceptron has a rectifier, *tanh*, or some max-out activation function. It is typically important to shuffle training data when implementing deep learning because the rows are processed sequentially during training. The input layer is scaled to the number of columns, and this is typically an indication of the complexity of the model. After sample training, backpropagation and loss function assessment are performed. The algorithm will go through the complete training set a number of times as defined by the user, and this hyperparameter is called epoch. The epoch value is key to finding the model that represents that sample with less error.

Deep learning has become one of the more widely used machine learning algorithms in recent years, however, several shortcomings are present when applying deep learning to disease data. First, it is difficult to interpret how the model arrives at its predictions and infer biological insight behind the disease development. Next, deep learning does not perform much better than other machine learning algorithms unless the dataset is complete, with little to no sparsity, unlabeled, and contains hundreds of thousands, if not more, observations. Since the dataset we use consist of prelabeled diagnoses, deep learning might not achieve top predictive performance in our study.

Stacked Ensembles

Ensemble machine learning methods combine multiple machine learning algorithms to obtain better predictive performance than the result from any of the constituent models. Many popular machine learning algorithms such as Random Forest and Gradient Boosting Machines utilize the ensemble method. Stacked Ensemble employs Stacking, also called Super Learning and Stacked Regression, which trains a second-level meta-learner to find the optimal combinations of the base learners.

To train a stacked ensemble, a list of base algorithms is chosen with a specific set of model parameters, followed by a second-tier learner, which uses the predictions of the base models as features. The second-tier algorithm may be the same as one of the base learner. Next, each base algorithm is trained on the training set, and a k-fold cross validation is performed on each of these learners. The cross-validated prediction values from the base algorithms are combined and used to train the meta-learning algorithm. The resulted model, combined with the base learners, forms the ensemble model to generate predictions on the new data.

Research Design and Methods

Data Acquisition

The Cancer Genome Atlas

The CNV data and corresponding clinical information are stored in Google BigQuery^{TM}, for which the data is accessible through Standard Query Language (SQL), and these TCGA Big Query tables are publicly available. Cloud computing allows more efficient data storage and bulk data manipulation without straining the computing power of local machines. Data analysis was performed in the statistical programming language R. The "bigrquery" package was used to download the required data subsets from the cloud storage for further manipulation to be trained in GBM models.

UK Biobank

Researchers need to apply for approval to access UK Biobank data through the data showcase platform for a $500 application fee. Once the application is approved, another £2000 was required in user fees. The initial process took several months until the data became available for access.

We specified the exact data categories needed in our research when applying for access; for instance, we wanted to look at cancer diagnosis, and patient clinical data such as gender and age, we then chose the corresponding categories listed on the data showcase: https://biobank.ndph.ox.ac.uk/showcase/browse.cgi.

Once access was granted, we were sent a key through email to download and decrypt the data. The key was then stored as a file called "*integers*.key", where the integers correspond to our application number, in the working directory and made readable through the command “chmod 755 *integers*.key”. We were also given another encrypted file named ukb*integers*.enc (same numbers as the key). This file contains approximately 500k lines, where each line consists of individual patient IDs and their respective parameters chosen in the application, e.g. sex, cancer type, etc.

The first step was to retrieve all the necessary tools to download the UK Biobank data. The instructions are available here:

https://biobank.ndph.ox.ac.uk/showcase/download.cgi

To decrypt the downloaded file, we ran the following command:

```
$ ./ukb_unpack ukbintegers.enc kintegers.key
```

which produced the file ukb*integers*.enc_ukb.

The following commands were used to extract the decrypted data into useful formats:

```
$ ./ukb_conv ukbintegers.enc_ukb docs

$ ./ukb_conv ukbintegers.enc_ukb r
```

docs produced an html file that contains the documentation of the variables in the dataset, and r produced a tab delimited file and an R script for labeling the variables.

We have previously downloaded the l2r genetic data, but a new version was made available in early 2021, so we re-downloaded the data following the instruction listed here:

https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/ukbgene_instruct.html ukbgene is a Linux executable, and we had to first use the following command to download a small file that contained a list of patient IDs:

```
$ ./ukbgene l2r –c1 –m
```

This file was renamed as “patientIDSall.fam”. The -c1 argument specifies chromosome 1, but other chromosomes would also suffice since we only needed the patient ID information, which was the same across all chromosomes. The patientIDSall.fam file contains one patient ID per row, and the row order is important because it matches the column headers for the l2r genetic data.

To download the l2r genetic data, we wrote a shell script that executed the following:

```
$ ukbgene l2r –cN
```

where N is an integer ranged from 1 to 22, representing the chromosome number, and each file was saved as “ukb22431_l2r_cN_b0_v2” (22431 was our application number). We needed to download each chromosome individually, and it was a timely process to download 2.3 terabytes of data while performing error checking. It took around a week or two, and we needed to check on it every day to make sure that ukbgene didn’t fail and had to be restarted. The largest file was chromosome 1, 195GB, and the smallest file was chromosome 21, 34GB.

All the l2r data were plain text files, with numbers separated by spaces and no headers. They were formatted such that every column consisted of one sample, i.e. a single patient, with their ID given by the patientIDSall.fam file. Every row was the SNP location that the log2 ratio value was measured in the array.

Data Processing

The Cancer Genome Atlas

The dataset was formatted such that every row contained one observation, i.e. a single patient, which was denoted by case-barcode, a unique identifier. Every column was defined by the genomic address of a gene segment, consisting of chromosomal location, start and end base pair positions. Each cell then contained the segment mean for the gene segment defined by its column for the patient recorded in that particular row. The value could be blank if the CNV was normal and no information was available. For example, the unmasked data with 50 top CNVs had 50 columns, one for each of the CNVs. Lastly, the cancer types that the patients had been diagnosed with were recorded in another column. Additional information such as age, gender, and ethnicity may be incorporated, though the diagnosis classification models only included CNVs data.

UK Biobank

The l2r data was converted into the Chromosomal Scale Length Variation data through shell script to directly handle the large files, since the file size that could be processed in IDEs such as R Studio is limited by RAM size. Chromosomal Scale Length Variation is the average l2r value of large CNV segments across a chromosome, evenly split into desired numbers of pieces. The mathematical formula is as follows:

$$\frac{\sum_{i=0}^{n} log_2 \frac{CN_i}{2}}{n}$$

Where CN is the copy number value, normalized first by division of two for each allele then taken the base 2 log, resulting in a single l2r value. For instance, to calculate the average values of 4 splits of chromosome 22, which has 11342 lines, two splits would contain 2835 sequential lines, and the other two would contain 2836 sequential lines. Each column in all four splits would be averaged.

Server Specifications

The hardware specifications of the server are as follows: AMD Ryzen 9 5950X 3.4 GHz 16-core AM4 processor (upgraded from Intel Xeon E5-2960 2.90 GHz CPU in 2021), 32GB of DDR4 3600 MHz RAM (upgraded from DDR3 2133 MHz RAM in 2021), GeForce GT 710 GPU (2GB GDDR3), and 10 TB HDD. In addition, We created a 64 GB swap for additional memory on the hard disk. We set up our computer server to run Linux Ubuntu 20.04 (64-bit) LTS for its operating system.

R Statistical Programming Language Specifications

The initial work with TCGA data was conducted in R v3.6.3, and the UK Biobank portion was done in R v4.0.3. Detailed instructions to install different versions of R in the Linux environment are available here: https://cran.r-project.org/bin/linux/ubuntu/

H2O Machine Learning

We trained, tested, and validated our Gradient Boosting Machine (GBM) models using H2O, the leading, open-source machine learning platform. The distributed systems and in-memory computing of H2O accelerated machine learning with massive datasets, and its accessibility for many programming languages, e.g. R, python, allowed us to seamlessly deploy models while maintaining reproducibility of the data analysis in R. The software also utilizes many popular machine learning algorithms, both supervised and unsupervised, such as GBM, XGBoost, Random Forest, Deep Learning, etc., and this facilitates the process of algorithm comparison to determine the most suitable one. Each algorithm is equipped with extensive parameters for fine-tuning to improve model performance and handle issues such as overfitting.

H2O Generalized Linear Models

H2O follows the authoritative text by P. McCullagh and J.A. Nelder[50] on the generalization of linear models to non-linear distributions of the response variable Y by fitting GBLM models based on the maximum likelihood estimation via iteratively reweighed least squares[45,46].

Let $y_1, \dots, y_n$ be n observations of the independent, random response variable Y_i. Assume that the observations are distributed according to a function from the exponential family and have a probability density function of the form:

$$f(y_i) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\alpha_i(\phi)}\right] + c(y_i;\phi)$$

where θ and ϕ are location and scale parameters, and$\alpha_i(\phi)$, $b(\theta_i)$, and $c(y_i;\phi)$ are known functions.

α_i is of the form $\alpha_i = \frac{\phi}{p_i}$ where p_i is a known prior weight.

When Y has a probability distribution function from the exponential family:

$$E(Y_i) = \mu_i = b'var(Y_i) = \sigma_i^2 = b''(\theta_i)\alpha_i(\phi)$$

Let $g(\mu_i) = \eta_i$ be a monotonic, differentiable transformation of the expected value of y_i. The function η_i is the link function and follows a linear model:

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i'\beta$$

When inverted: $\mu = g^{-1}(\boldsymbol{x}_i'\beta)$

Maximum Likelihood Estimation

For an initial rough estimate of the parameters $\hat{\beta}$, use the estimate to generate fitted values:

$$\mu = g^{-1}(\hat{\eta}_i)$$

Let z be a working dependent variable such that $z_i = \hat{\eta}_i + (y - \hat{\eta}_i)\frac{d\eta_i}{d\mu_i}$, where $\frac{d\eta_i}{d\mu_i}$ is the derivative of the link function evaluated at the trial estimate.

Calculate the iterative weights: $w_i = \frac{p_i}{[b''(\theta_i)(\frac{d\eta_i}{d\mu_i})^2]}$, where b'' is the second derivative of $b(\theta_i)$ evaluated at the trial estimate.

Assume $\alpha_i(\phi)$ is of the form $\alpha_i = \frac{\phi}{p_i}$. The weight w_i is inversely proportional to the variance of the working dependent variable z_i for current parameter estimates and proportionality factor ϕ. Regress z_i on the predictors x_i using the weights w_i to obtain new estimates of β:

$$\hat{\beta} = (X'WX)^{-1}X'Wz$$

where X is the model matrix, W is a diagonal matrix of w_i, and z is a vector of the working response variable z_i.

The process is repeated until the estimates $\hat{\beta}$ change by less than the specified amount.

H2O Distributed Random Forest

Distributed random forest (DRF) is one of the powerful classification and regression tools available in H2O. The algorithm generates a set of classification or regression trees instead of a single tree, and each tree is a weak learner built from a subset of rows and columns from the given dataset, with the addition of more trees reducing the variance[51]. The final prediction is computed from the average prediction of all the trees in the model. Tree building and growth is stopped randomly by several stopping metrics, such as tree depth, number of leaves or nodes[54]. Random forest shares some similarities with gradient boosting machines, however, the former builds the weak learners and trees independently without input from the other trees in the model.

H2O Gradient Boosting Machines

GBM involves three major elements: a loss function to be optimized, a weak learner to make predictions, and an additive model to add weak learners to minimize the loss function. The loss function depends on the problem type; for instance, it may be squared error for regression and logarithmic for classification. Decision trees are used as the weak learner, as they output real values for splits. The trees are constructed in a greedy manner to minimize the loss function. Lastly, the trees are added sequentially, and a gradient descent procedure is employed to minimize the loss function. After the loss is calculated, a decision tree is added to the model that follows the gradient and thus reduces the loss. The tree is parameterized, and its parameters modified to move in the right direction by reducing the residual loss[52]. The output of the new tree is added to that of the existing sequence of trees to correct the final output of the model.

H2O's GBM algorithms follow the algorithm specified by Hastie et al[49]. The goal is to minimize the residuals r_{ikm}, which are the gradient values for each of the K bins. The iterative construction of regression trees, denoted as γ_{jkm}, allows for the results and errors of the previous tree to be incorporated into the creation of subsequent trees. The specific algorithm used by H2O is as followed[53]:

```
Initialize $f_{k0} = 0, k = 1,2, \dots, K$
For $m = 1$ to $M$:
    1. Set $p_k(x) = \frac{e^{f_k(x)}}{\sum_{l=1}^{k} e^{f_k(x)}}, k = 1,2, \dots, K$
    2. For $k = 1$ to $K$:
        a. Compute $r_{ikm} = y_{ik} - p_k(x_i), i = 1, 2, \dots, N$
        b. Fit a regression tree to the targets $r_{ikm}, i = 1, 2, \dots, N$, giving the terminal regions $R_{jim}, j = 1, 2, \dots, J_m$
        c. Compute $\gamma_{jkm} = \frac{K-1}{K} \frac{\sum_{x_i \in R_{jkm}} (r_{ikm})}{\sum_{x_i \in R_{jkm}} |r_{ikm}|(1-|r_{ikm}|)}, j = 1, 2, \dots, J_m$
        d. Update $f_{km}(x) = f_{k,m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jkm} I(x_i \in R_{jkm})$
  Output $\hat{f}_k(x) = f_{kM}(x), k = 1,2, \dots, K$
```

H2O Deep Learning Neural Networks

H2O's deep learning implementation is based on a multi-layer feedforward artificial neural network (ANN) trained with stochastic gradient descent using back propagation. The ANN model is also known as deep neural network (DNN), and it is the most common type of deep learning that works well with tabular data. The network contains a large number of layers consisting of neurons with tanh, rectifier, maxout activation functions. While the number and size of hidden layers can be customized by the user, the minimum is at least one hidden layer. High predictive accuracy is achieved through advanced features such as adaptive learning rate, rate annealing, momentum training, dropout, L1 or L2 regularization, checkpointing, and grid search[54]. H2O asynchronous trains multiple copies of the global model parameters on the local data at each compute node, and individual performance is periodically fed to the global model through model averaging across the network.

The default setting of deep learning in H2O sets two hidden layers of size 200 each and a stopping metric of log loss for classification. It is recommended to shuffle the training set because the training is done in order. The input layer automatically scales to the number of input features or columns for the given dataset; therefore, any complexity reduction would need to be done prior to feeding the training data into the neural network.

H2O Stacked Ensembles

Ensemble machine learning methods utilize multiple learning algorithms to achieve better predictive performance than the result obtained from the individual algorithms. H2O's Stacked Ensemble method employs a specific process called stacking to find the optimal combination of a collection of prediction algorithms. Stacking involves training a second-level metalearner to find the optimal combination of the base learners, its goal being to ensemble strong, diverse sets of learners together.

The steps below outline the procedure of training and testing a super learner ensemble. H2O automates most of the process for efficient model building on the platform[55].

1. Set up the ensemble.
    a. Specify a list of L base algorithms, with a specific set of model parameters.
    b. Specify a metalearning algorithms.
2. Train the ensemble.
    a. Train each of the L base algorithms on the training set.
    b. Perform k-fold cross-validation on each of these learners and collect the cross-validated predicted values from each of the L algorithms.

    c. The N cross-validated predicted values from each of the L algorithms can be combined to form a new N x L matrix. This matrix, along with the original response vector, is called the “level-one” data.
    d. Train the metalearning algorithm on the level-one data. The “ensemble model” consists of the L base learning models and the metalearning model, which can then be used to generate predictions on a test set.

3. Predict on new data.
    a. To generate ensemble predicts, first generate predictions from the base learners.
    b. Feed those predictions into the metalearner to generate the ensemble prediction

For cross validation, all base models need to have the same number of folds, and our experiment used 10-fold cross validation for all the models. The cross-validated prediction results have to be saved to train the metalearner. In addition, base models are trained on the same training data, with a minimum of two base learners required.

Additional R Packages

All the R packages can be found on CRAN: https://cran.r-project.org/web/packages/available_packages_by_name.html

ukbtools

This R toolset is used to visualize primary dataset from UK Biobank data files and query ICD Diagnoses, retrieve genetic metadata, read and write standard formats of genetic analyses[56].

tidyverse

The ‘tidyverse’ is a collection of open source R packages that are very useful in the field of data science. It is designed to simplify the process of loading multiple ‘tidyverse’ packages in a single step, since they all share the common data representations and API design[57].

dplyr

This is one of the core packages of the ‘tidyverse’, and it contains many functions that enable dataframe manipulation in an intuitive and user-friendly fashion[58].

tidyr

As the name implies, this R toolset helps transform messy data into tidy data, which follows these principles:

1. Every column is a variable.
2. Every row is an observation.
3. Every cell is a single value.

It pairs nicely with ‘dplyr’ in data wrangling and manipulation tasks[59].

ggplot2

One of the most popular R packages, ‘ggplot2’ was developed based on the “Grammar of Graphics”, which employs a data visualization scheme that breaks up graphs into semantic components such as layers and scales. It is a system for ‘declaratively’ creating graphics, allowing for versatile manipulation of data visualization that could replace the base graphics in R[60]. It is also a part of the ‘tidyverse’ collection.

ggthemes

This package provides additional themes, geoms, and scales for ‘ggplot2’ and allows for more customization and aesthetics for data visualization. Some examples are Stata graph schemes, range frame, Tufte’s box plot, ‘The Economist’ color scheme, ‘Wall Street Journal’ theme, etc[61].

Training GBMs on TCGA Unmasked Germline CNV Data for Diagnosis Classification

Germline CNVs from unmasked TCGA studies were used for GBM model training. Only the genomic information from normal blood sample was included; therefore, cancers derived from hematopoietic cells, i.e. Acute Myeloid Leukemia (LAML) and Chronic Myelogenous Leukemia (LCML), were excluded to prevent their results from skewing other models. When building model for each cancer, samples of the target type retained their cancer labels, e.g. OV, BRCA, PCPG, while the rest of the samples were labeled as “Normal”. The models were trained with 100 trees and balanced classes, with no max depth specified, and the training was followed by ten-fold cross validation to avoid overfitting and obtain the AUC values for each model. The procedure was repeated five times for each cancer type, and the results were averaged for model performance evaluation.

Training GBMs on TCGA Germline CNV Data for Incidence Age Prediction

Similarly, only germline CNV data from normal blood sample was used in this analysis, and LAML and LCML were thus excluded. However, samples from both masked and unmasked TCGA studies were combined. In addition to gCNVs, factors such as gender, race, ethnicity, and their combinations were taken into consideration. During model training, the data was divided into subsets by cancer type, such that only samples of a single cancer type were included for each model. To better evaluate whether the prediction was better than random guess, control sets were constructed with a simple random sampling without replacement method. To generate randomized dataset for each cancer type, the ages of the patients were scrambled, while the CNVs and other parameters were kept the same. Models were then built for the controls, simulating results from random chances.

Training Machine Learning Algorithms on UK Biobank Germline L2R Data for Diagnosis Classification

To verify whether our results were database-specific, i.e. highly influenced by artifacts found in the TCGA, we validated our methods and models on the data from UK Biobank, which has the additional advantage of containing data from healthy individuals. After acquisition of the germline l2r data and associating clinical data and transformation into CSLV values, the UK Biobank data were first separated into normal-patient group and cancer-patient groups.

Results

The Cancer Genome Atlas (TCGA)

Prediction and Classification of Cancer Diagnosis

The initial results show that it is possible to predict cancer types with unmasked germline copy number variation (CNV) data from The Cancer Genome Atlas (TCGA). There exist inherent genetic differences between patients diagnosed with different cancer types to make germline chromosomal scale length variation-based classification feasible.

For each of the 32 cancer types from TCGA database, 10 gradient boosting tree models were built. However, the two blood-related cancers, Chronic Myelogenous Leukemia (LCML) and Acute Myloid Leukemia (LAML), were omitted because the data was collected from normal blood samples.

To select the number of CNVs for the model building, the segments were first sorted by their corresponding measurement counts in descending order. The analysis utilized top 30 CNV segments, and each CNV carries an individual segment mean. The value is normalized average of copy number of specific regions of the genome. In comparison to the masked data, the unmasked data includes Y chromosomes and calculates segment means for small genomic segments, instead of large portions that account for almost the entire chromosome in some cases, providing a more precise measurement.

	Chromosome	Start	End	Length	Count	Mean	SD
1	3	68697108	68698262	1154	3508	-0.52	1.32
2	7	54318812	54318824	12	2956	-0.46	1.86
3	8	39378084	39529446	151362	2830	-0.68	1.49
4	13	71903417	71906418	3001	2815	-1.40	1.83
5	5	58030200	58037706	7506	2707	1.36	0.86
6	Y	2782397	56872112	54089715	2664	-1.00	0.28
7	6	103290101	103314186	24085	2601	-0.37	1.00
8	7	70958264	70961155	2891	2523	-1.87	1.55
9	4	172067997	172068382	385	2471	-0.70	1.24
10	5	104524672	104524903	231	2361	-0.13	1.62
11	3	193160173	193165114	4941	2275	-0.25	1.31
12	2	146106836	146109366	2530	2096	-0.62	1.32
13	1	109690352	109697556	7204	2085	0.04	1.39
14	4	114254167	114261019	6852	2083	-0.57	1.09
15	9	23363117	23373486	10369	2030	-0.84	1.65
16	7	154601461	154607903	6442	2020	-1.20	1.43
17	20	80664	1580353	1499689	1982	0.01	0.02
18	5	46271828	46273400	1572	1950	-1.46	1.74
19	6	149661	254283	104622	1938	0.02	0.06
20	13	37497899	37510620	12721	1918	-0.58	1.40
21	16	55764310	55764867	557	1900	2.54	0.92
22	4	9459504	9477699	18195	1810	0.33	1.33
23	5	177800550	177805604	5054	1784	2.20	0.97
24	5	181006225	181363319	357094	1776	0.02	0.06
25	8	40920333	40920952	619	1735	-1.90	0.85
26	12	9481274	9575696	94422	1694	-0.32	1.34
27	1	152789447	152796224	6777	1682	-0.46	1.01
28	8	645892	645908	16	1645	2.77	2.55
29	19	51639099	51644944	5845	1643	-1.02	1.22
30	8	111282050	111283031	981	1628	0.23	1.32

Table 1: Top 30 CNVs Ranked by Count in TCGA

We selected CNVs based on the number of patients in which they appeared. These CNVs were identified as part of the TCGA bioinformatics pipeline. This table shows the top 30 CNVs ranked by count. The location of the CNV is characterized by its chromosome number, start, and end points in HG38 coordinates. The length of each CNV, its respective count, i.e. the number of patients out of 8859 who had this CNV, the mean and standard deviation of different listed values for that CNV.

A common error of model building is overfitting. This phenomenon occurs when the model is tuned too preferentially to the noise, instead of the targeted signal, so it performs exceptionally well with the training set but poorly with the testing set. Therefore, a 10-fold cross validation was implemented for each cancer type. Receiver operating characteristic (ROC) curves were plotted for the cross-validation results of individual models, in respect to ROC curve with an area-under-curve (AUC) of 0.5 that represents a model formed by chance. The corresponding AUCs were averaged to evaluate individual model performance, as shown in Table 1.

Cancer Type	Average AUC	Standard Deviation
Uterine Corpus Endometrial Carcinoma	0.632	0.008
Bladder Urothelial Carcinoma	0.682	0.004
Prostate Adenocarcinoma	0.639	0.008
Breast Invasive Carcinoma	0.696	0.002
Ovarian Serous Cystadenocarcinoma	0.819	0.004
Sarcoma	0.607	0.013
Glioblastoma Multiforme	0.782	0.002
Skin Cutaneous Melanoma	0.659	0.004
Head and Neck Squamous Cell Carcinoma	0.690	0.010
Pancreatic Adenocarcinoma	0.556	0.017
Lung Squamous Cell Carcinoma	0.698	0.003
Kidney Renal Papillary Cell Carcinoma	0.676	0.005
Brain Lower Grade Glioma	0.626	0.004
Lung Adenocarcinoma	0.618	0.005
Stomach Adenocarcinoma	0.688	0.006
Thyroid Carcinoma	0.685	0.006
Liver Hepatocellular Carcinoma	0.750	0.007
Colon Adenocarcinoma	0.667	0.009
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	0.644	0.008
Pheochromocytoma and Paraganglioma	0.836	0.009
Mesothelioma	0.790	0.013
Esophageal Carcinoma	0.779	0.013
Rectum Adenocarcinoma	0.621	0.008
Testicular Germ Cell Tumors	0.728	0.012
Kidney Renal Clear Cell Carcinoma	0.649	0.008
Thymoma	0.784	0.022
Uveal Melanoma	0.780	0.015
Adrenocortical Carcinoma	0.745	0.018
Uterine Carcinosarcoma	0.709	0.020
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	0.598	0.017
Cholangiocarcinoma	0.679	0.044
Kidney Chromophobe	0.406	0.136

Table 2: Average Performance of GBM Models

The average Area-Under-Curves (AUCs) of GBM models trained on top 30 germline CNVs indicate performance better than chance for most cancer types.

Cancer types such as Ovarian Cancer (OV), Pheochromocytoma and Paraganglioma (PCPG) and Glioblastoma Multiforme (GBM) performed notably well, with AUC values over 0.75 in analysis using only top 30 gCNVs (Fig. 1). Kidney Chromophobe (KICH) has poor performance likely due to the insufficient sample size (n = 9), as it is a rare, genetic disorder. The results indicate that the modeling technique was able to classify cancer diagnosis using gCNVs from unmasked data, and its performance was better than random guessing.

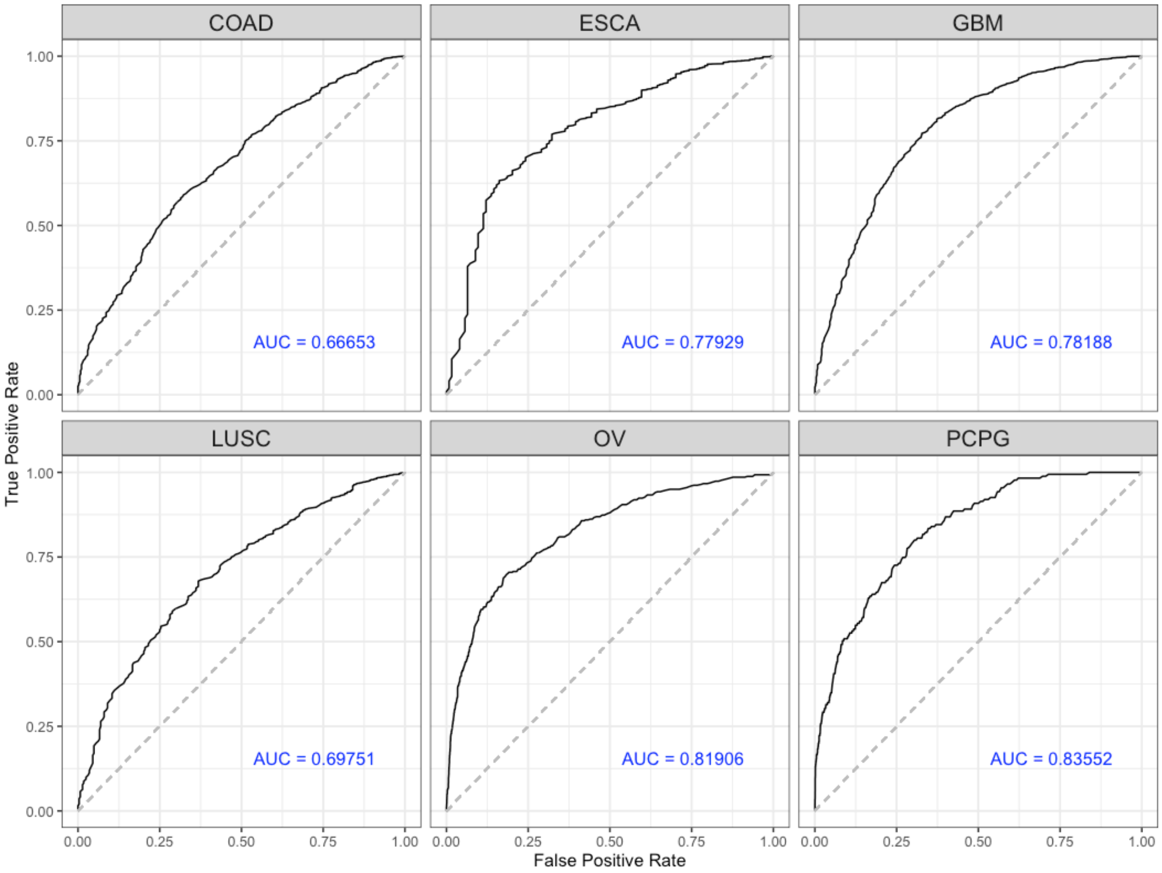

Figure 1: Receiver Operating Characteristic (ROC) Curves of Germline CNV Cross-Validation Models
Selection of ROC curves for cross-validation metrics of gradient-boosted machine classification models on germline CNVs. Six models are shown: Colon adenocarcinoma (COAD), Esophageal Carcinoma (ESCA), Glioblastoma Multiforme (GBM), Lung Squamous Cell Carcinoma (LUSC), Ovarian Serous Cystadenocarcinoma (OV), and Pheochromocytoma and Paraganglioma (PCPG), with PCPG in lead (AUC = 0.84).

AUC Comparison of Models Utilizing Different Numbers of Top CNV

As noted earlier, the unmasked data contains fewer samples, raising the question whether the same number of top gCNVs used in analysis for unmasked data should be used for masked data. Therefore, we conducted an analysis to investigate the effect of increasing gCNVs on AUCs and the point where the incremental change in performance tapers off by building cancer diagnosis classification model with different numbers of top gCNVs: 30 (used with masked data), 50, 75, 100, 150, and 200.

The following results indicate that the respective AUC value of the model built for a cancer type increases with the number of top CNV used (Fig. 2). Although using every CNV in the unmasked models could potentially optimize the predictive powers, it would be too computationally exhaustive and might even result in diminishing return, as the great number of parameters provide too much interference. The model performance plateaued at top 75 CNVs for many cancer types (Fig. 3), suggesting top 75 to be a suitable cutoff for efficient and accurate cancer prediction from unmasked germline CNVs.

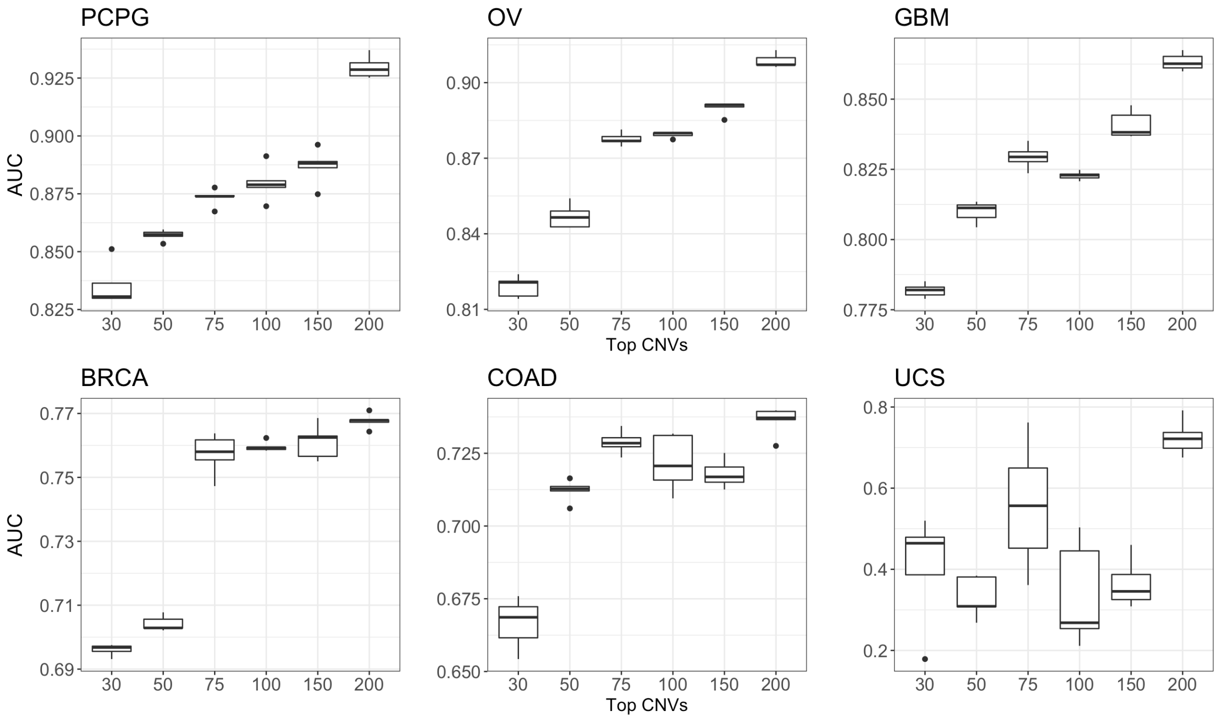

Figure 2: Area-Under-Curve Values of Germline CNV Models utilizing different numbers of top CNVs
Selection of Area-Under-Curve (AUC) values vs. numbers of top CNVs for gradient-boosted machine classification models on germline CNVs. The six models show discernable trend of proportionality between top CNVs and AUC.

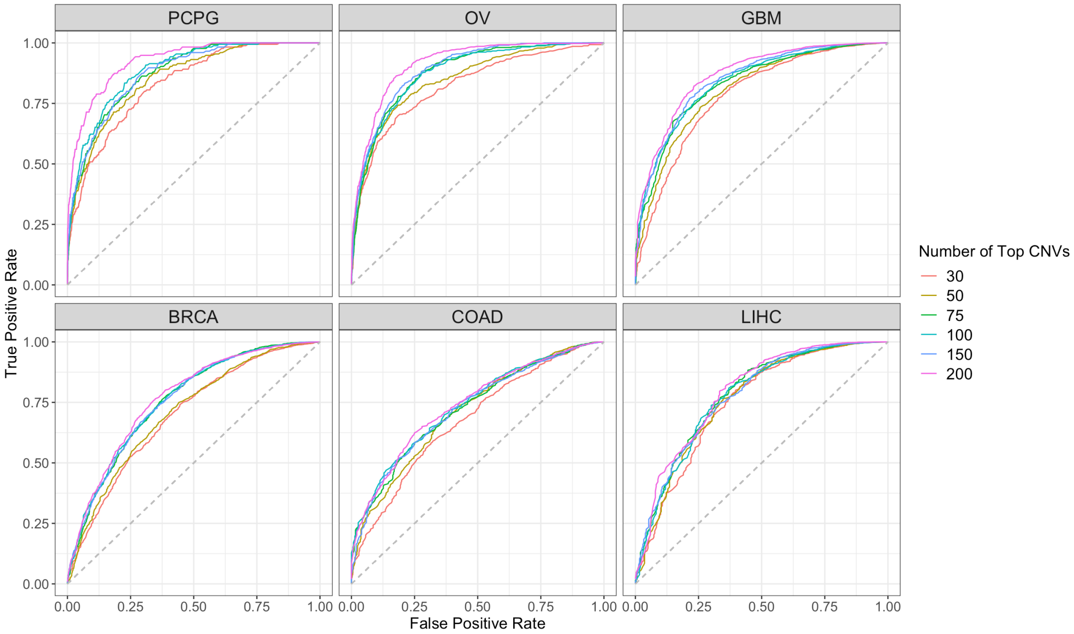

Figure 3: ROC Curves of Germline CNV Models utilizing different numbers of top CNVs
Selection of ROC curves of top CNVs for gradient-boosted machine classification models on germline CNVs. The six models show discernable trend of proportionality between top CNVs and AUC.

Cancer Incidence Age Prediction

In addition to cancer classification, we were interested in whether chromosomal scale length variation of DNA play a role in cancer incidence age. Though the accuracy of cancer onset age prediction using chromosomal scale length variation alone from combined unmasked and masked TCGA data is dubious, the results suggest that the prediction models performed better than chance for some cancers.

To determine whether the age prediction was more accurate than chance, we first constructed control datasets with a simple random sampling without replacement method. For each cancer type, the ages of the patients were scrambled, while the CNVs were kept the same, to generate a randomized dataset. Gradient Boosting regression model was then built from the control, simulating results from random guessing.

This analysis used top 50 CNV segments. For each type of cancer, 10 gradient boosting regression models with gCNVs as predictors and Age as predicted variable were built per set: the original and control datasets. The models were subsequently cross validated 10 times, and correlation coefficients between the predicted and observed ages and the root mean square errors (RMSEs) of each model were obtained as evaluation metrics. To determine whether the differences in correlation coefficients and RMSEs of between the original and control datasets were statistically significant, one-sided Welch's t-tests were conducted, as it was hypothesized that the correlation coefficients of the original datasets would be higher than control, while the RMSEs of the original datasets would be lower.

The overall low correlation coefficients, i.e. less than 0.1, suggested that age prediction based on gCNVs alone was not entirely feasible. Head and Neck Squamous Cell Carcinoma (HNSC) resulted in the highest average correlation coefficient of 0.1021 (Fig. 4). The t-test outputs showed that the prediction models were better than random guessing for slightly less than half of the cancer types (Fig. 5), as 13 out of 28 cancers indicated statistically significant results. For RMSE, the t-test outputs likewise demonstrated that age prediction based on gCNVs was better than chance for some cancer types: 10 out of 28 cancers had statistically significance differences (Table 3).

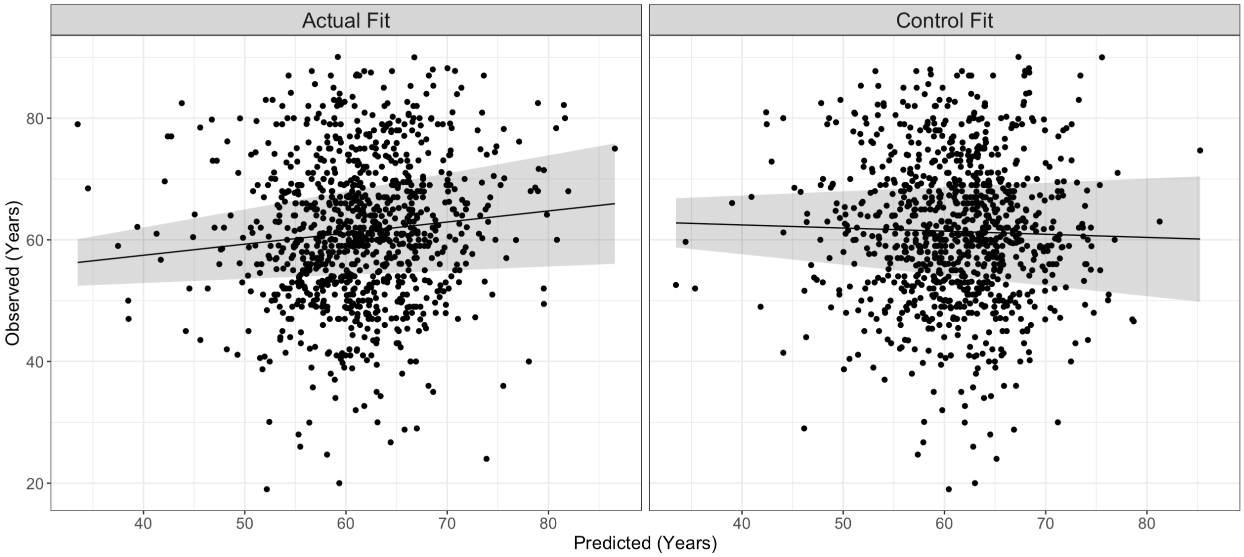

Figure 4: Performance of Multifactorial Age Prediction Models
Head and Neck Squamous Cell Carcinoma results in the best performance, with a correlation coefficient of 0.1021. The actual fit showed a stronger correlation between observed and predicted age than control. The 95% confidence intervals of the linear fit are indicated by the grey-shaded areas

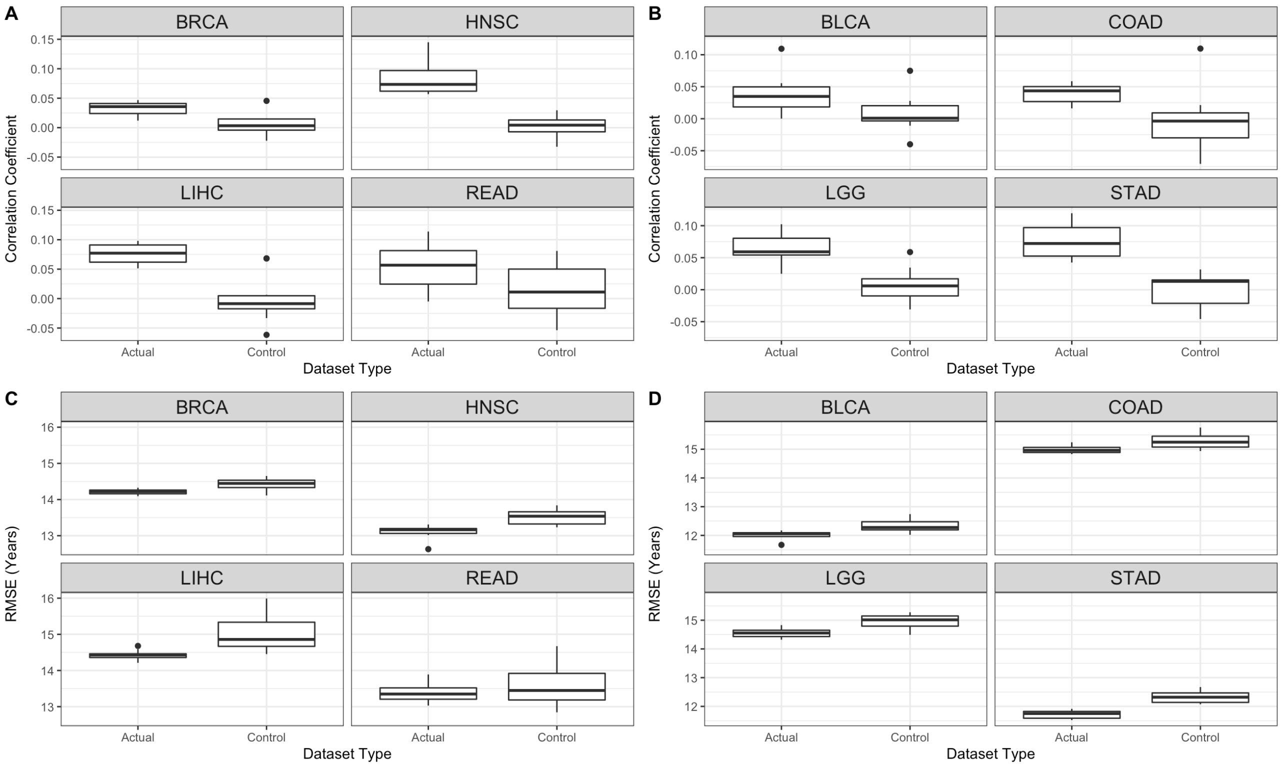

Figure 5: Performance of CNV Age Prediction Models
Selection of age prediction models with statistically significant differences in performance. A) and B) compare correlation coefficients between predicted and observed age of actual and control datasets. C) and D) compare cross-validation RMSEs of actual and control datasets.

Next, we were interested in whether the addition of other phenotypic factors, i.e. gender, race, and ethnicity, would increase the predictive power. These elements were incorporated into the previously constructed CNV models stepwise, i.e. in different combinations such as CNV+gender, CNV+race+ethnicity. For the models utilizing all parameters, 15 out of the 28 cancer types exhibited significant differences in correlation coefficients and RMSEs between original and control datasets (Table 3). In addition, the overall correlation coefficients increased, with decrease in RMSE observed in many cancer types (Fig. 6, 7). The results suggest that the inclusion of gender, race, and ethnicity in the CNV models improves the predictive power, and the difference in performance in respect to chance is more evident.

	Number of Significant Cancers (out of 28)	
Model	**Correlation Coefficient**	**RMSE**
CNV Only	13	10
CNV+Ethnicity	15	12
CNV+Race	16	10
CNV+Gender	14	10
CNV+Race+Ethnicity	16	11
CNV+Gender+Ethnicity	15	12
CNV+Gender+Race	15	12
All	15	15

Table 3: Number of Significant Cancers of Different Multifactorial Age Prediction Models

The eight multifactorial age prediction models were evaluated using Correlation Coefficient and RMSE as criteria. t-tests were performed between actual and control datasets to determine whether the differences in correlation coefficients or RMSE were significant. The addition of more phenotypic factors results in greater contrasts in age prediction performance.

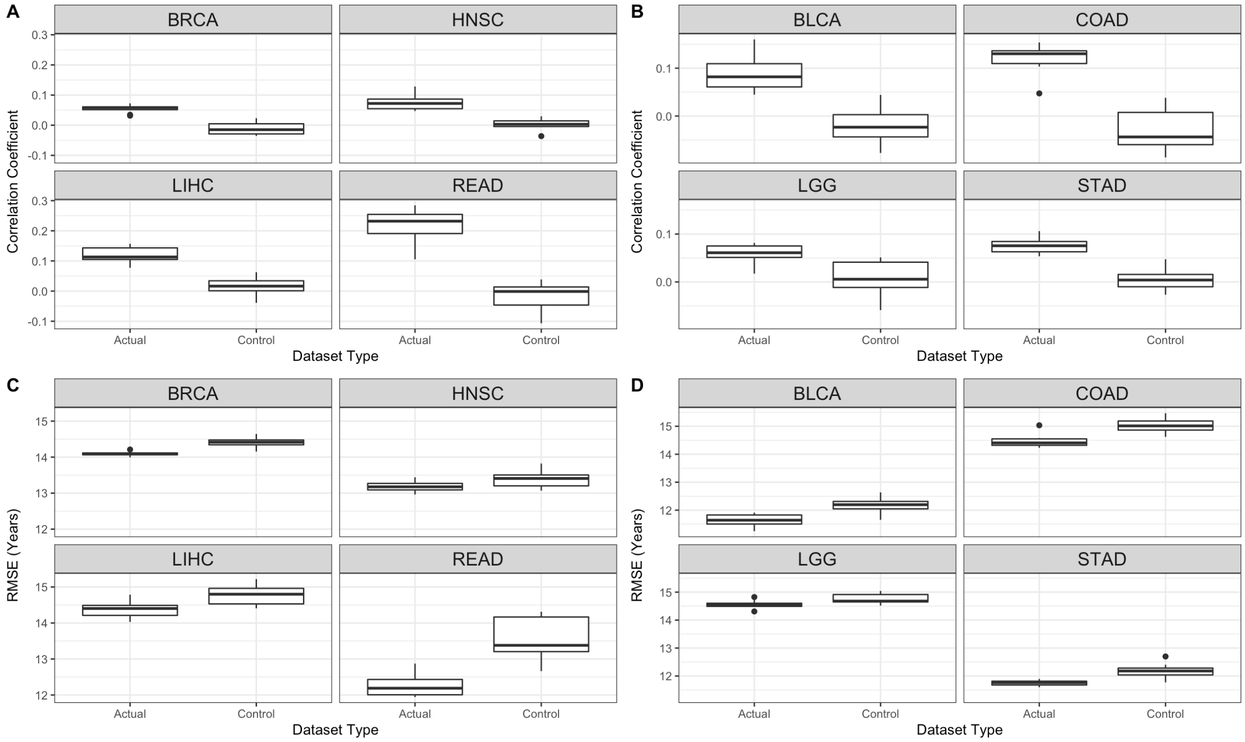

Figure 6: Performance of Multifactorial Age Prediction Models

Selection of age prediction models with statistically significant differences in performance. Models were built from CNVs, gender, race, and ethnicity. A) and B) compare correlation coefficients between predicted and observed age of actual and control datasets. C) and D) compare cross-validation RMSEs of actual and control datasets.

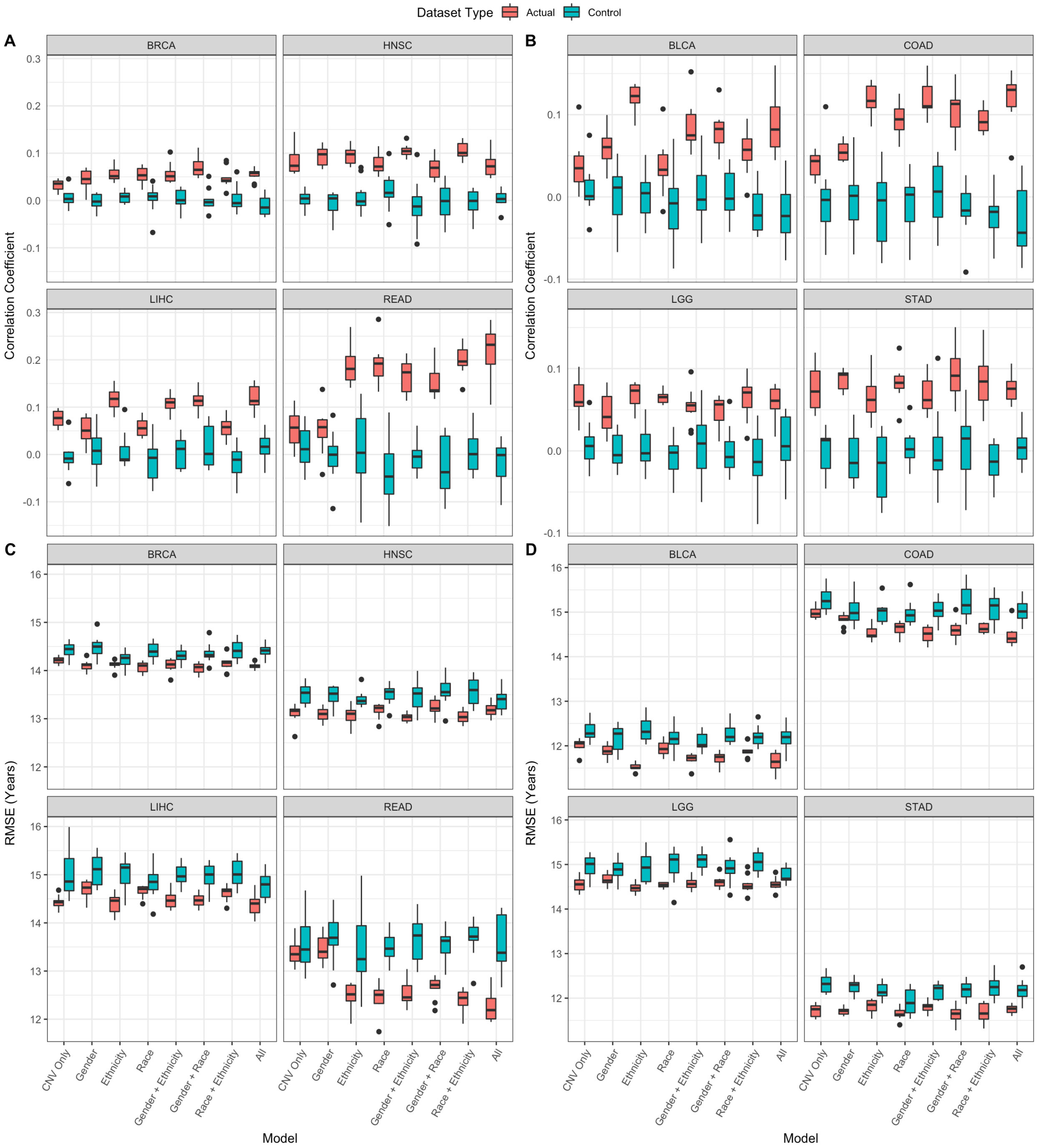

Figure 7: Performance of Different Multifactorial Age Prediction Models

Selection of age prediction models with statistically significant differences in performance. Models were built from eight combinations of CNVs, gender, race, and ethnicity: CNV only, CNV+Gender, CNV+Ethnicity, CNV+Race, CNV+Gender+Ethnicity, CNV+Gender+Race, CNV+Race+Ethnicity, and all. A) and B) compare correlation coefficients between predicted and observed age of actual and control datasets. C) and D) compare cross-validation RMSEs of actual and control datasets.

GBM Algorithm Library Comparisons

There are different implementations of the gradient boosting method available in H2O: the aforementioned GBM, XGBoost, and LightGBM, the latter two building and improving upon the traditional GBM. XGBoost employs a regularization function to control overfitting[62], and LightGBM incorporates two techniques to improve performance: one is gradient-based one-side sampling that emphasizes the most informative samples, and the other is exclusive feature bundling that groups similar features to reduce complexity[63].

Cancer diagnosis classification was implemented with the three algorithm libraries, with comparison performed across all the top CNVs. All three libraries followed the trend of increasing AUCs with the number of CNVs used in the models, and their performances were generally comparable. However, for some cancer types, XGBoost and LightGBM resulted in greater AUCs than the base GBM library, especially at higher numbers of CNVs (Fig. 8). The most exceptional case would be KICH, which was the only cancer scoring an AUC below 0.5 with top 30 CNVs in the GBM model, likely due to its small sample size. Its XGBoost- and LightGBM-implemented models were able to consistently achieve AUCs greater than 0.5 across all the top CNVs. In addition to their faster running time, the improved GBM library implementations should be taken into consideration for future studies.

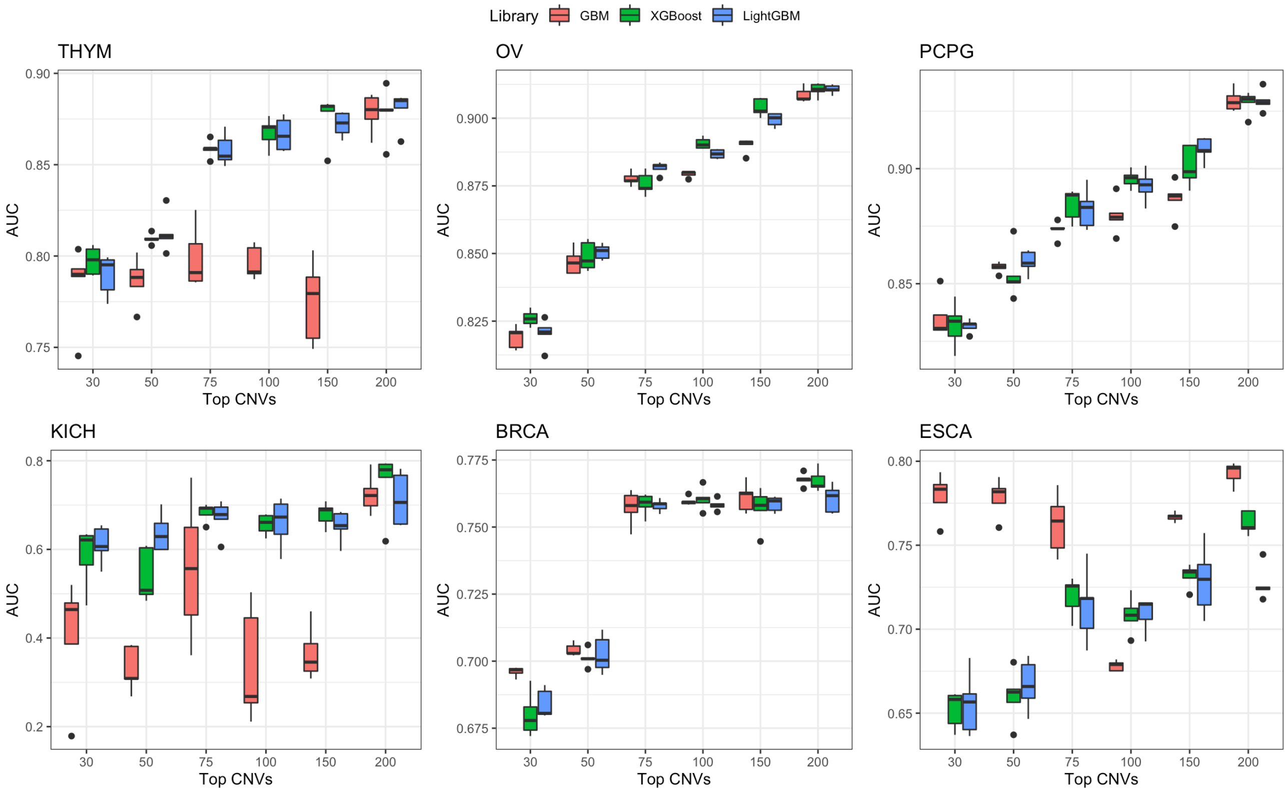

Figure 8: Cancer Classification Performance Comparison of Different GBM Algorithm Libraries

Selection of AUC values vs. numbers of top CNVs for three implementations of gradient-boosted machine classification models on germline CNVs. THYM, OV, and PCPG show significant differences in performance between base GBM library and the improved GBM libraries: XGBoost and LightGBM. KICH, the only cancer for which the prediction was worse than random chance, was able to consistently achieve AUC values greater than 0.5 across all top CNVs in XGBoost- and LightGBM- implemented models.

Genetic Risk Score for Glioblastoma Multiforme based on Copy Number Variation

Glioblastoma multiforme (GBM) is the most common form of brain cancer[64]. It is an aggressive form of cancer, with the median survival time measured in months. In 2008, patients diagnosed in the US had a median survival time ranging from 31.9 months to 5.6 months, depending on their age[65]. Several lines of evidence suggest that glioblastoma multiforme has a genetic basis. First, multiple cases of this rare disease have been reported to occur within single families[66]. Second, the only environmental factor associated with glioblastoma multiforme, high doses of ionizing radiation, is rare and not present for the vast majority of people diagnosed with this disease[67]. Most importantly, GWASs have identified several SNP alleles that are present significantly more in glioblastoma multiforme patients than expected[68].

The maximum accuracy of a genetic test is a function of the heritability and prevalence of a disease[69]. The heritability of glioblastoma multiforme in a Northern European population is about 26% (95% confidence interval: 17%-35%[70]. Based on this number and the prevalence of glioblastoma multiforme in a similar population (about 2-3 per 100,000 persons), the maximum accuracy of a genetic test measured by the area under the receiver operating characteristic curve (AUC) should exceed 0.95[69]. Tests based on SNPs do not come close to this AUC value. We set out to determine how well a glioblastoma multiform predictive DNA test based on copy number variations could perform.

We have previously found that the gradient boosting algorithm performs the best with this particular dataset and thus employed this algorithm to compute genetic risk score for glioblastoma multiforme. We wanted to investigate how the performance of classification changed with the number of distinct CNVs included in the model by examining how well these overlapping sets of CNVs could predict whether an individual would develop glioblastoma multiforme. Figure 9 shows that the predictive ability, quantified by AUC, of the gradient boosting classification models varies with the number of different top ranked CNVs included in the model. We also discovered that the classification performance improved with more features. The respective receiver operating characteristic curves are displayed in Figure 10.

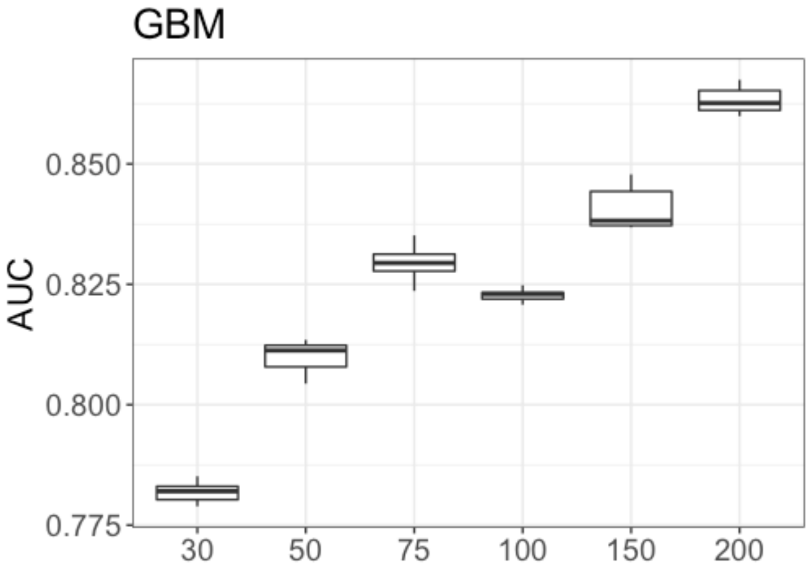

Figure 9: Area-Under-Curve Values of Glioblastoma Multiforme Classification Models utilizing different numbers of top CNVs
Area-Under-Curve (AUC) values vs. numbers of top CNVs for glioblastoma multiforme classification models on germline CNVs. The six models show discernable trend of proportionality between top CNVs and AUC.

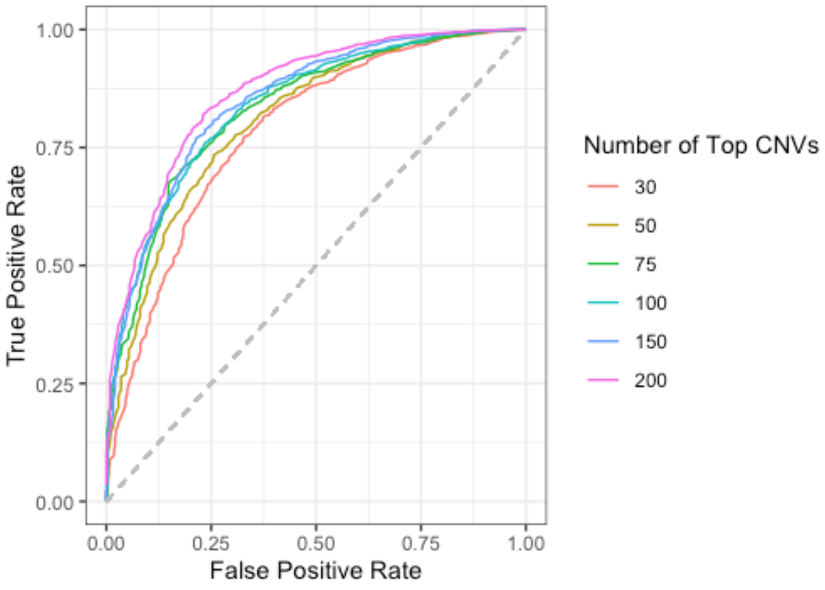

Figure 10: ROC Curves of Glioblastoma Multiforme Classification Models utilizing different numbers of top germline CNVs

ROC curves of top CNVs for glioblastoma multiforme gradient-boosted machine classification models on germline CNVs. The six models show discernable trend of proportionality between top CNVs and AUC.

Next, we characterized how well the classification model would work as a genetic risk score. Five-fold cross validation was used on the gradient boosting model with 200 top CNVs to obtain genetic risk scores for each of the 8726 patients in the dataset, and these individuals were ranked by their respective scores and assigned a percentile. Table 4 shows the results in tabular form, in which the samples were grouped into quintiles, each consisting of 20 percentile points.

Quintile	Normal	GBM	Total	Odds Ratio	95% Confidence Interval
1	1738	8	1746	0.07770379	0.04-0.15
2	1718	27	1745	0.26530325	0.18-0.39
3	1699	46	1745	0.45705285	0.34-0.62
4	1651	94	1745	0.96113136	0.77-1.20
5	1432	313	1745	3.68980390	3.17-4.30

Table 4: Odds Ratio Quintile of Glioblastoma Multiforme Prediction

From the five-fold cross validation, each individual in the dataset was assigned a genetic risk score from the gradient boosting classification model. The samples were ranked from lowest to highest then separated into quintiles. The table presents the number of patients with and without glioblastoma multiforme in each quintile along with the odds ratio (relative to the entire group) and the 95% confidence interval for the odds ratio.

Figure 11 presents a graph of odds ratio, relative to the entire dataset, of the patients in the given percentile having glioblastoma multiforme. The 8726 individuals are binned into 50 equal groups, each consisting of two percentile points.

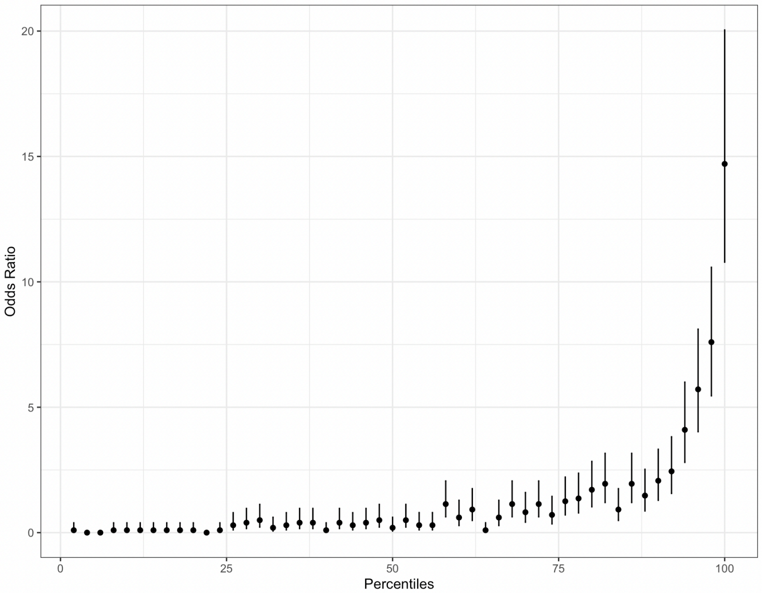

Figure 11: Odds Ratio Graph of Glioblastoma Multiforme Prediction

Patients ranked higher by the gradient boosting classification model are significantly more likely to have glioblastoma multiforme. The predictive model ranked all the people in the dataset based on their likelihood of having glioblastoma multiforme. This ranking was then grouped into 50 equal partitions. The plot shows the odds ratio of each of the 50 equal partitions along with the 95% confidence intervals.

To examine the feature contribution, we split the dataset of 8726 patients into training set (80%) and testing set (20%) then trained a gradient boosting machine to predict whether an individual had glioblastoma multiforme. Figure 12 presents the SHAP contribution plot that shows which CNVs play the most significant role in gradient boosting predictive model.

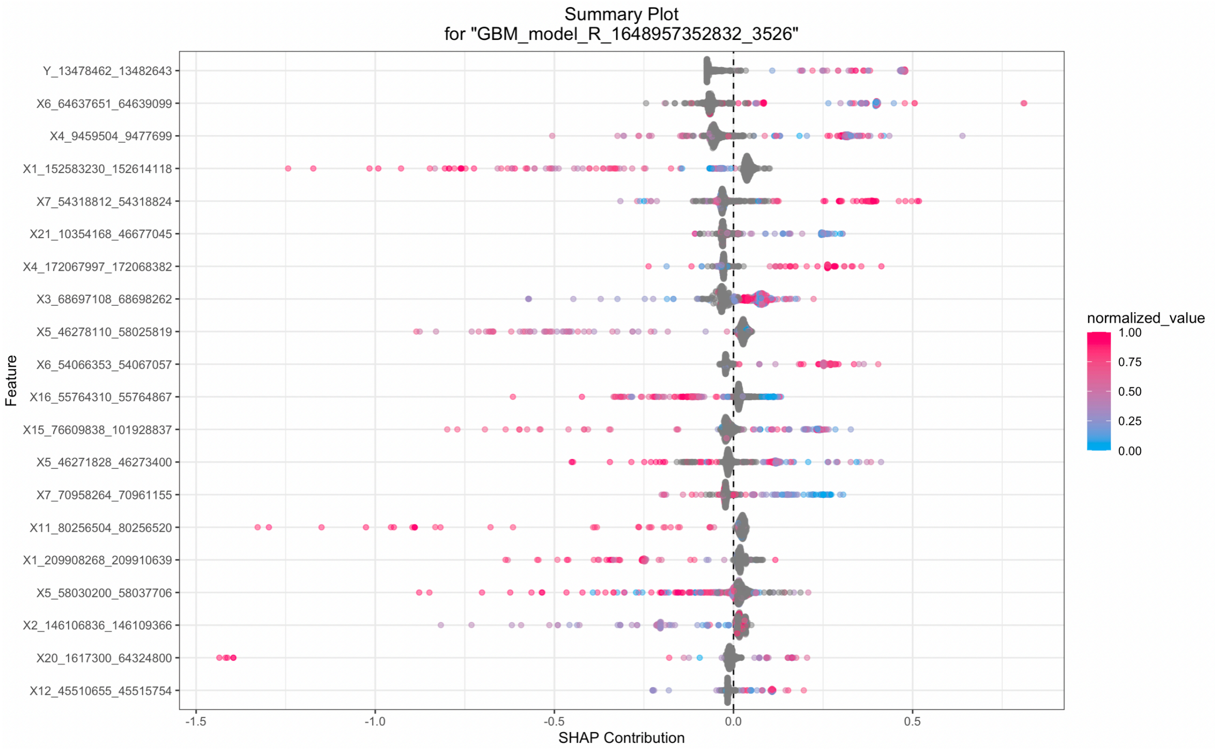

Figure 12: SHAP Contribution Plot of Predictive Model of Glioblastoma Multiforme Prediction
This plot ranks the importance to the predictive model of each CNV. Each individual is represented by a dot. The color of the dot represents the normalized chromosome length, and the position of the dot on the x-axis represents the impact of that CNV on the model prediction result for that respective patient. The plot indicates that Y_12378462_13482643 is more important than X12_45510655_45515754 in predicting Glioblastoma Multiforme.

Recent genome wide association studies have shown that glioblastoma multiforme is distinct from other forms of glioma, and these studies have identified a few regions of the germ line genome that are significantly different in people who develop glioblastoma multiforme[68]. The three SNPs with the highest levels of significance are: rs10069690 at 5p15.33, rs634537 at 9p21.3 and rs2297440 at 20q13.33. We checked whether our copy number variation data overlaps with these three SNPs. None of the three overlap with our data. The closest is the SNP rs634537, which still lies about 1.3 megabases away from the #15 copy number variation in our dataset shown in Table 1. This distance indicates the two are not related. This finding suggests that the copy number variation data is complementary to the SNP data provided by GWAS studies. The overall predictive accuracy of a germline test should increase by combining copy number variation data with SNP data.

Initial work by the TCGA project on glioblastoma multiforme focused on genome differences between normal germ line DNA and the somatic DNA found in glioblastoma tumors[71,72]. That work identified specific mutations and complex rearrangements that most glioblastoma tumors shared. Other work on the TCGA dataset related to glioblastoma multiforme identified prognostic indicators, somatic alterations in the tumor's DNA that could predict survival[73,74]. In contrast, we examined only germ line DNA copy number variation to measure how well these germ line DNA alterations can predict whether a person will develop glioblastoma multiforme.

Other work on using germ line DNA copy number variation to predict development of a disease also exists. Our group used chromosomal-scale length variations, a large scale version of copy number variation, to predict whether a person will develop ovarian cancer[75] and other forms of cancer[76] using TCGA data. Another group used a GWAS-type analysis employing logistic regression with copy number variation data collected from about 1800 ovarian cancer cases and 1800 controls to demonstrate that some germ line DNA copy number variations occur more frequently in women who develop epithelial ovarian cancer than in those who don't develop that form of cancer[77].

Genetic risk scores have been developed for several other forms of cancer. A large prostate cancer study of 1370 cases and 1239 controls found that a polygenic risk score built from 65 SNPs could predict prostate cancer with an AUC of 0.67[78]. Breast cancer can also be predicted by genetic risk scores. A recent study used a genetic risk score based on 287 SNPs in a European population and found an AUC of about 0.63. They also found that this same genetic risk score is applicable to a Chinese population, where it had an AUC of about 0.61[79]. One genetic risk score to predict breast cancer risk is commercially available and has been validated in several large cohorts with over 100,000 women. Women scoring in the top 1% of this commercially available genetic risk score have an odds ratio of about 2.0 of developing breast cancer compared to women scoring in the 40-60 percentile[80].

Our study has several limitations. We performed a reanalysis of existing data that was not collected for this purpose. It would be better to design a prospective study where samples could be collected ahead of time from a diverse, but well defined, group of people. Since we used a non-linear machine learning algorithm rather than logistic regression, we could not correct for population substructure, as is typically done in GWAS studies[81].

Our analysis is based on TCGA, a single dataset. Although TCGA was designed to be inclusive and it included a wide selection of people with different racial and ethnic backgrounds, it is not clear how well these results will generalize to any specific population. Future studies should be done to validate these findings by applying these predictions to different populations and testing how well they perform in a new population.

Finally, our control population consists of people who have been diagnosed with many different types of cancer patients, but not glioblastoma multiforme. This is an unfortunate aspect of using the TCGA dataset. It would be better to draw the control population from the general population instead of limiting it to only those who have been diagnosed with other forms of cancer.

Genetic Risk Score for Colorectal Cancer based on Copy Number Variation

Colorectal cancer (CRC) is the second leading cause of cancer-related death and the third most common cancer worldwide[82]. In 2018, it accounts for approximately 1.8 million new cancer diagnoses and more than 880,000 deaths[83]. The age-standardized incidence rates of CRC vary greatly across continents[84,85], with the highest present in Australia and New Zealand, and the lowest in Africa and South-Central Asia. These differences could be attributed to hereditary susceptibility, socioeconomic status, diet, lifestyle, and screening practices[86], as CRC encompasses a heterogenous cancer group, influenced by both exogenous and endogenous factors.

CRC is most often diagnosed in elderly individuals; however, in recent years there has been a rise of incidence rate of early on-set CRC, which is generally defined as CRC diagnosed in individuals under 50 years of age, worldwide, and the reason behind such phenomenon is poorly understood[87,88]. Patients with early-onset CRC are more likely to be diagnosed with advanced-stage disease than individuals with late-onset CRC. A lack of awareness, recognition of symptoms, and screening of the disease might contribute to delayed diagnosis and prevalence of advanced-stage disease at the time of diagnosis.

There have been a number of studies on the CRC predisposition. One study examined 53 SNPs for CRC and constructed genetic risk score. It found that those in the highest decile of genetic risk score were 3-fold more likely to have colorectal cancer compared to the lowest decile[89]. Other studies have also found that colorectal cancers can be predicted from genetics with similar effectiveness[72,90]. A better understanding of the underlying genetic basis of the disease could help guide prevention, early detection, and treatment strategies.

We aimed to develop a strategy based on structural variation rather than SNPs to compute genetic risk scores for CRC. We employed machine learning algorithms to account for the non-linear effects between CNVs, instead of linear combinations.

From the TCGA dataset, we constructed a case-control study to test the genetic risk score built from CNV data. 504 patients had been diagnosed CRC, and 8222 individuals had not been diagnosed with any form of CRC. The TCGA group statistics is shown in Table 5.

	Diagnosed with CRC	**Not Diagnosed with CRC**
% Women	239/504 = 47.4%	4397/8222 = 53.5%
% Men	265/504 = 52.6%	3825/8222= 46.5%
% Black	62/504 = 12.3%	708/8222 = 8.6%
% White	273/504 = 54.2%	6104/8222= 74.2%
% Asian	11/504 = 2.2%	573/8222= 7.0%
Mean Age	66.4	58.7
Total	504	8222

Table 5: TCGA Group Statistics of CRC Patients

From the TCGA dataset, we constructed two groups. One consisted of all individuals who had been diagnosed with CRC. The other contained patients that had not been diagnosed with CRC. This table compares characteristics of the two groups.

We carried out a preliminary investigation on the classification performance and number of different top ranked CNVs. We wanted to examine how well these overlapping sets of CNVs could predict whether an individual would develop CRC. Figure 13 shows that the predictive ability, quantified by AUC, of the gradient boosting classification models varies with the number of different top ranked CNVs included in the model. We discovered that the classification performance improved with more features, but it appeared to reach a plateau at 150 top CNVs. The respective receiver operating characteristic curves are displayed in Figure 14.

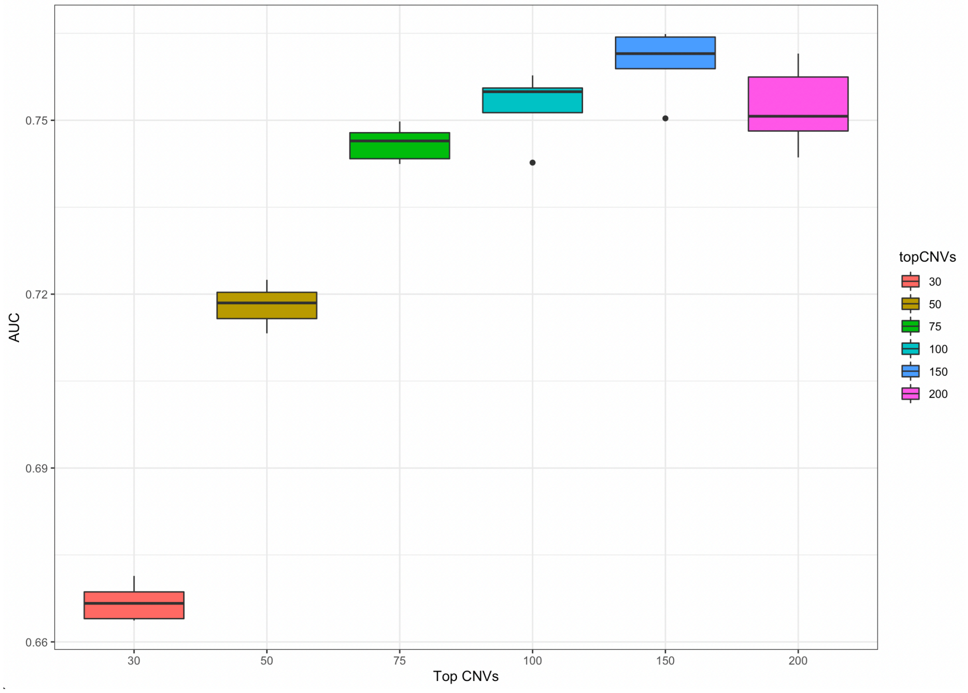

***Figure 13**: Area-Under-Curve Values of CRC Classification Models utilizing different numbers of top CNVs*
AUC values vs. numbers of top CNVs for six different classification models, each utilizing different number of CNVs. The performance generally increases with the number of CNVs but appears to plateau at 150 top CNVs.

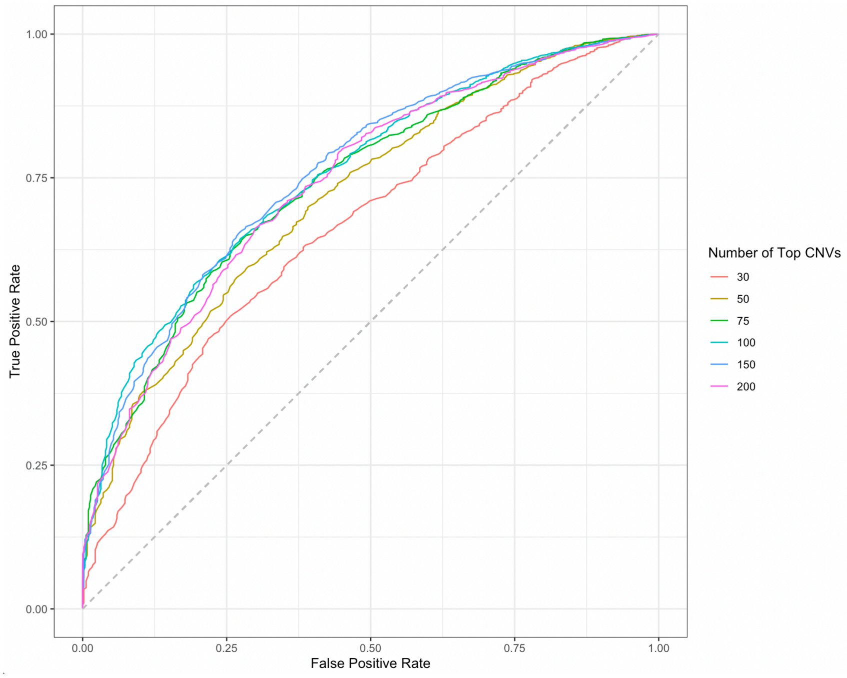

Figure 14: ROC Curves of CRC Classification Models utilizing different numbers of top germline CNVs
ROC curves of top CNVs CRC gradient-boosted machine classification models on germline CNVs. The six models show discernable trend of proportionality between top CNVs and AUC.

Next, we evaluated the predictive performance of different machine learning algorithms. Using 150 top CNVs, we measured how well these algorithms could identify whether a patient had been diagnosed with CRC or not. Each model building and classification was repeated five times. The performance metric we employed was the area-under-curve (AUC) value of the receiver operating characteristic (ROC) curve. The outcomes of these models are shown in tabular form in Table 6, and the graphical results, including AUC comparison and corresponding ROC curves are presented in Figure 15 and 16. The Gradient Boosting Machine is shown to achieve the highest AUC; however, this is not a conclusive result, since it might be possible to fine tune a deep learning network to attain superior performance. The gradient boosting machine has the advantages of faster running time and easier model tuning and manipulation. Others have found that gradient boosting machine does perform well on many different types of datasets.

Algorithm	**Average AUC**
Gradient Boosting Machine	0.76
XGBoost	0.75
Extremely Randomized Trees	0.71
Distributed Random Forest	0.69
Deep Learning	0.68
Generalized Linear Model	0.68

Table 6: Comparison of Machine Learning Algorithms for CRC Classification
We evaluated six popular and powerful machine learning algorithms from the H2O package in R for predicting CRC classification from CNV data. The algorithms are ranked by the best average AUC from five-fold cross validation, repeated five times.

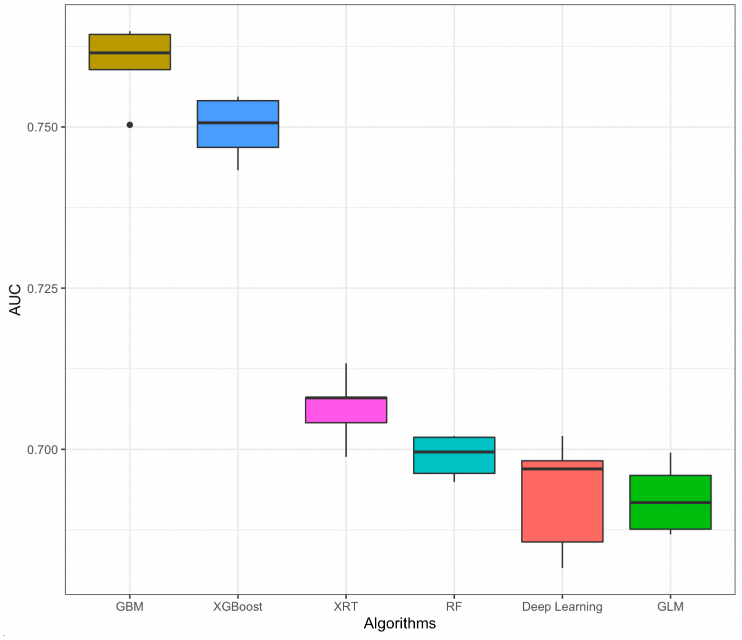

Figure 15: Area-Under-Curve Values of CRC Classification Models utilizing different algorithms
AUC values of six different CRC classification models, each employing different machine learning algorithm. Tree-based algorithms, specifically GBM and XGBoost, achieved the best performance.

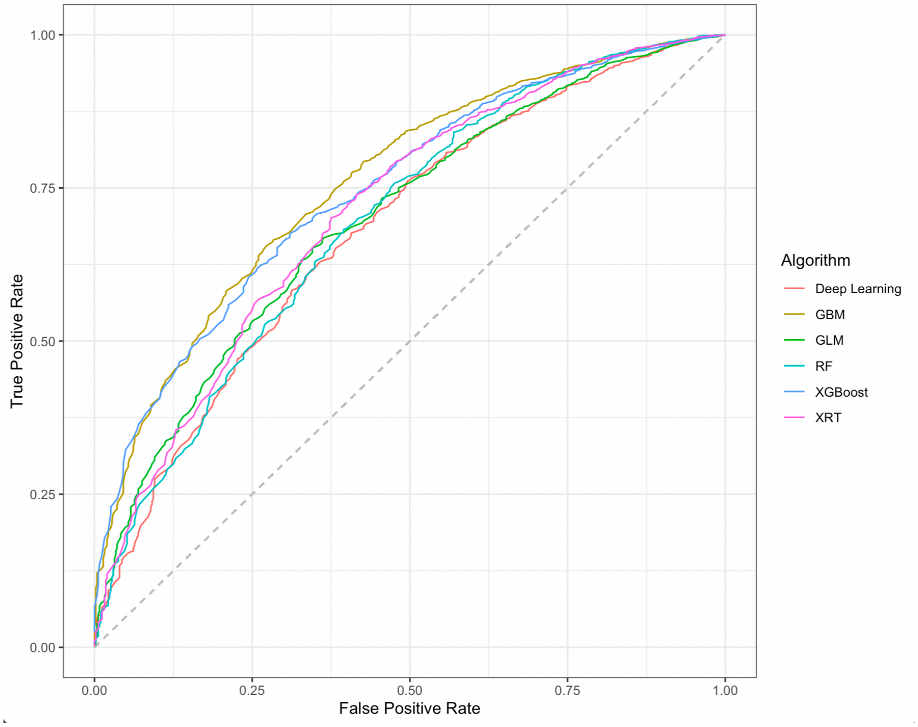

Figure 16: ROC Curves of CRC Classification Models utilizing different algorithms
ROC curves of CRC classification models employing different machine learning algorithms. The AUC for the best model, i.e. the gradient boosting machine model, was 0.76.

Table 6 indicates the effectiveness of Gradient boosting machine, thus we used this algorithm for the rest of the analysis in this study. For the following step, we aimed to classify the 8726 patients in the dataset. A ten-fold cross validation was carried out, randomly partitioning the dataset into ten equal groups. The first nine groups were used to train the model to assign individuals as CRC or non-CRC, and the last group was held out to be used as the test set. The model gave each patient in the test set a numerical score to quantify the likelihood of that particular individual belong to the CRC class. The process was repeated ten times, with different hold-out group each round, resulting in a numerical score for every single individual of the 8726 patients.

Decile	Number of Patients without CRC	Number of Patients with CRC	Total Number of patients	Odds Ratio	95% Confidence Interval
1	871	2	873	0.04	0.01 - 0.12
2	864	9	873	0.17	0.09 - 0.32
3	859	14	873	0.27	0.16 - 0.45
4	846	27	873	0.52	0.35 - 0.77
5	837	36	873	0.70	0.50 - 0.99
6	825	48	873	0.95	0.70 - 1.28
7	823	49	872	0.97	0.72 - 1.31
8	813	59	872	1.18	0.90 - 1.56
9	776	96	872	2.02	1.60 - 2.54
10	708	164	872	3.78	3.12 - 4.58

Table 7: Odds Ratio Deciles of CRC

Using 10-fold cross validation, each individual in the dataset was assigned a score from the CRC classification model. The patients were ranked from lowest to highest, then segmented into ten deciles. This table shows the number of individuals with and without CRC in each decile, the respective odds ratio relative to the entire group, and the 95% confidence interval for the odds ratio.

The predicted results were compared to the known CRC status of the patients, who were first ranked by their scores, from the least likely to have CRC to the most likely to be from the CRC class. We could evaluate the classification performance of the model through the comparison of the ranking with the “correct” CRC status of the patients. The relative risk of the samples split into ten different groups is shown in tabular form in Table 7. Similar information is presented in Figure 17, except the samples were split into 50 groups.

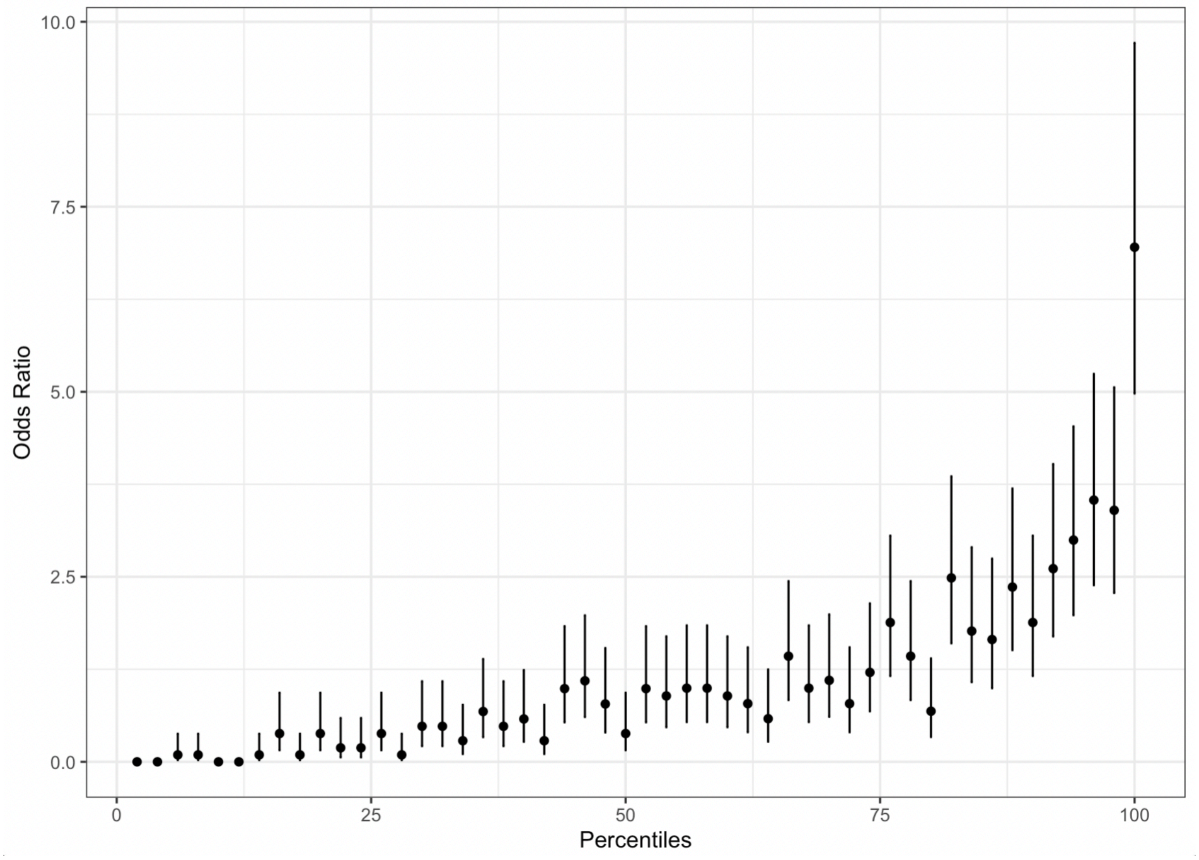

Figure 17: Odds Ratio Graph of CRC Prediction
Patients ranked higher by the gradient boosting classification model are significantly more likely to have CRC. The predictive model ranked all the people in the dataset based on their likelihood of having glioblastoma multiforme. This ranking was then grouped into 50 equal partitions. The plot shows the odds ratio of each of the 50 equal partitions along with the 95% confidence intervals.

Figure 18 shows the SHAP contribution plot of the gradient boosting machine model utilizing 150 top CNVs. It helps explain how the model arrives at its predictive results and the importance and weight of the features.

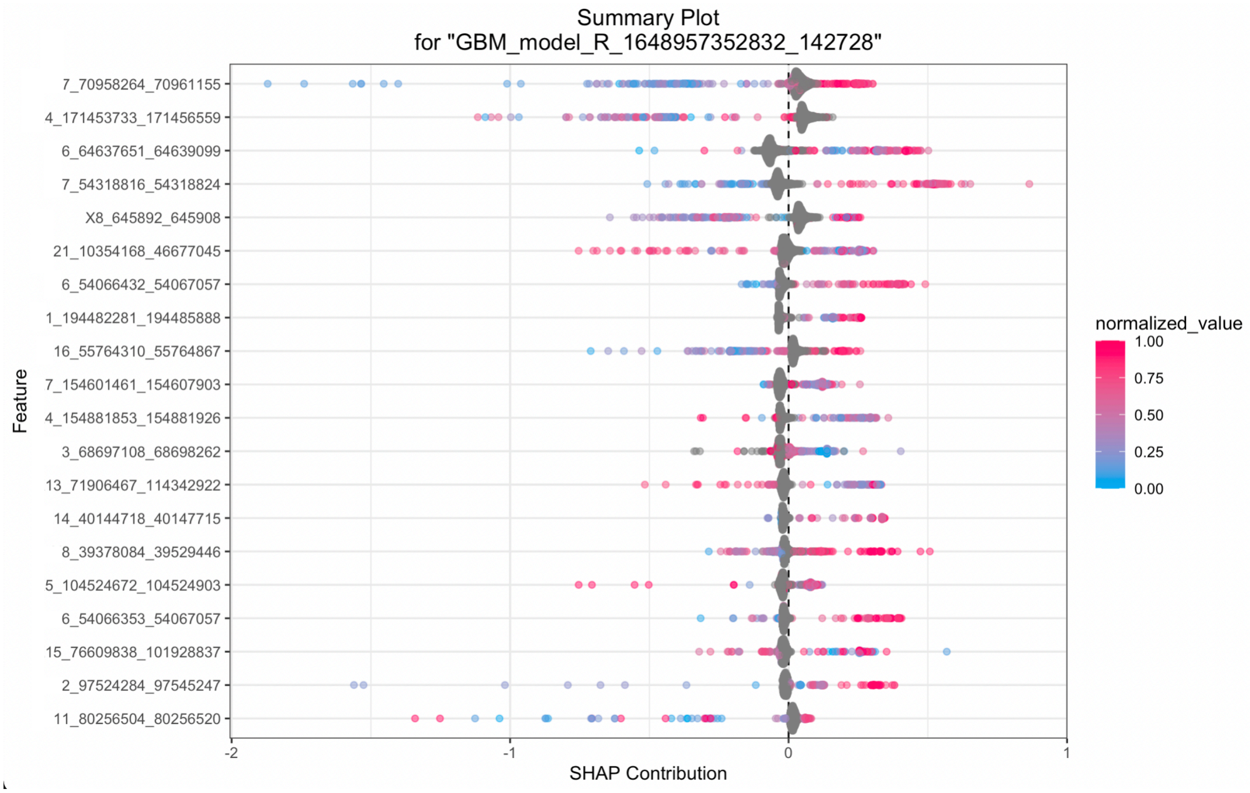

Figure 18: SHAP Contribution Plot of Predictive Model of CRC Prediction

This plot ranks the importance to the predictive model of each CNV. Each individual is represented by a dot. The color of the dot represents the normalized chromosome length, and the position of the dot on the x-axis represents the impact of that CNV on the model prediction result for that respective patient. The plot indicates that 7_70958264_70961155 is more important than 11_80256504_80256520 in predicting CRC.

The study yields promising results. A previous study found that the patients in the top 10% of genetic risk score had a 3-fold increase than those in the 10%. As shown in Table 7, the top 10% in our results were 90 times more likely to have colorectal cancer than the individuals scores in the bottom 10%.

One disadvantage of this approach is the difficulty in understanding and extracting biological meaning, in comparison to the more traditional SNP-based genetic risk scores. There is a fundamental difference between statistical methods for prediction and those attribution. The method we presented here focuses on prediction, while SNP-based risk scores are developed from GWAS studies, which were designed to identify specific genes responsible for diseases.

UK Biobank

Prediction and Classification of Cancer Diagnosis

We aimed to verify whether our cancer diagnosis prediction results were database-specific, so we validated our methods and models on UK Biobank, which has an additional advantage of containing data from healthy individuals. UK Biobank has the copy number variation data in the raw log2 ratio format, which needs to be transformed for dimensionality reduction to be usable in machine learning models. Therefore, the l2r data was converted into Chromosomal Scale Length Variation values by splitting each chromosome into four segments and computing the average of each segment. We followed Data-Field 41270, which contains summary ICD 10 Diagnoses for 440,019 participants, for the cancer diagnosis information. This study examines the more prevalent cancer types, listed in Table 8, along with their respective incidence counts.

Malignant Neoplasm Type	**Incidence Count**
Lung	5353
Brain	1051
Colorectal	8917
Kidney	2138
Esophagus	1627
Pancreas	1651
Bladder	4129
Stomach	1257
Prostate	13090
Ovary	2138
Breast	16496
Uterine	1795

Table 8: Incidence Count of Common Types of Malignant Neoplasm In UK Biobank

The table shows the malignant neoplasm types with the highest incidence counts, i.e. at least over 1000 individuals with l2r data available.

For each cancer type, the dataset consisted of diseased samples paired with age- and gender-matched cancer-free individuals. This control served to limit the heterogeneity originating from non-genetic sources, for some cancers exhibit higher incidence rates in one gender over the other, or certain cancer risks scale with age.

Figures 19-30 present the performance of different machine learning algorithms for each type of cancer. We found that the stacked ensemble models consistently performed best, and there are slight differences between algorithms and their performances. Gradient Boosting Machine and XGBoost was often among the top performing algorithms. For most cancers, except stomach cancer, all algorithms could predict cancer diagnosis significantly better than chance, represented by an AUC of 0.50, as shown in Figures 19-30. However, This indicates that a patient’s germline genetics, specifically the chromosomal-scale length variation values, demonstrate inherited predisposition to many cancer types.

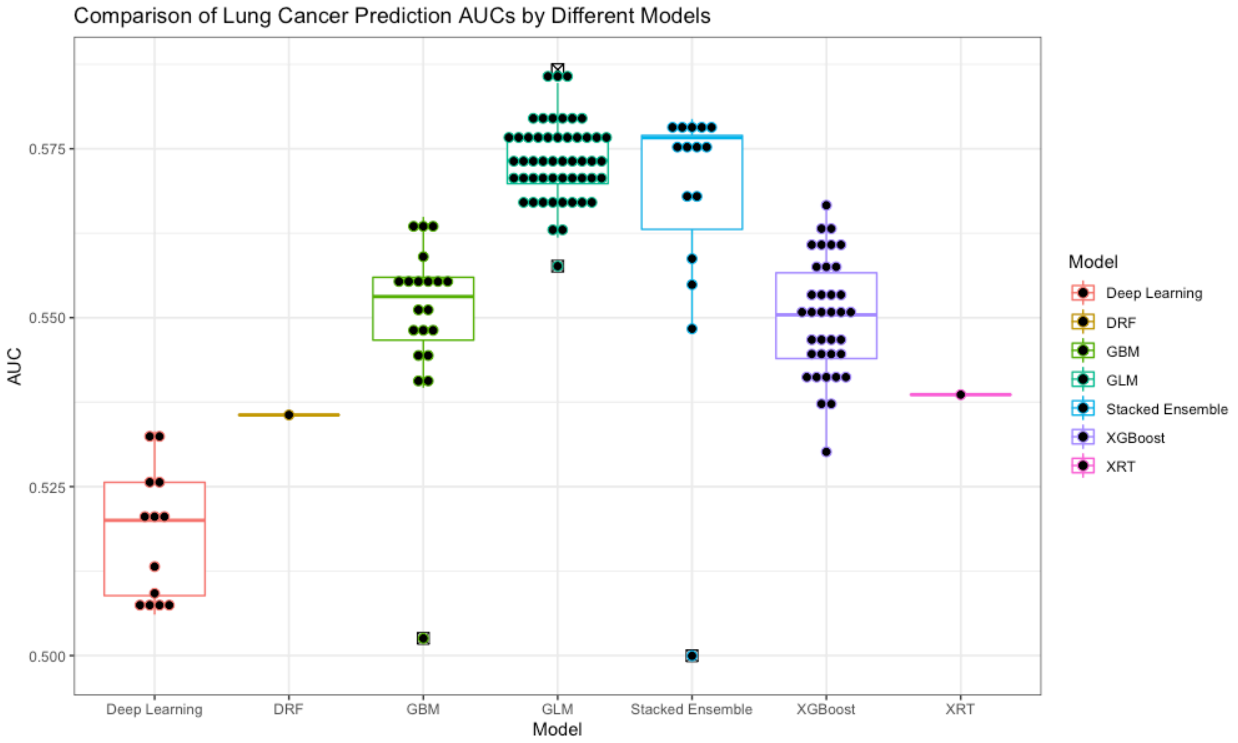

Figure 19: Comparison of Lung Cancer Prediction AUCs by Different Models
The predictive performance, represented by the metric AUC, of different machine learning algorithms. The dataset consists of lung-cancer patients with age- and gender-matched non-cancer individuals from the UK Biobank population as control. H2O was used to carry out a grid search for the best algorithms, and the top-performing models were selected and evaluated with AUCs. The algorithms tested were Deep Learning, Distributed Random Forests (DRF), Gradient Boosting Machine (GBM), General Linear Model (GLM), XGBoost, Extreme Randomized Tree (XRT), and Stacked Ensemble (combination of all the different models).

The average area-under-curve (AUC) value of the receiver operating characteristic curves (ROCs) for all machine learning models was 0.557, with a standard deviation of 0.020 and 95% confidence interval of (0.554, 0.560). The AUC differs from 0.5, which is equivalent to an AUC of random chance. Excluding the less optimal models, i.e. the deep learning models, the average AUC is 0.561, with a standard deviation of 0.015 and 95% confidence interval of (0.559, 0.564). Both AUCs are significantly different from 0.5, with $p<0.00001$.

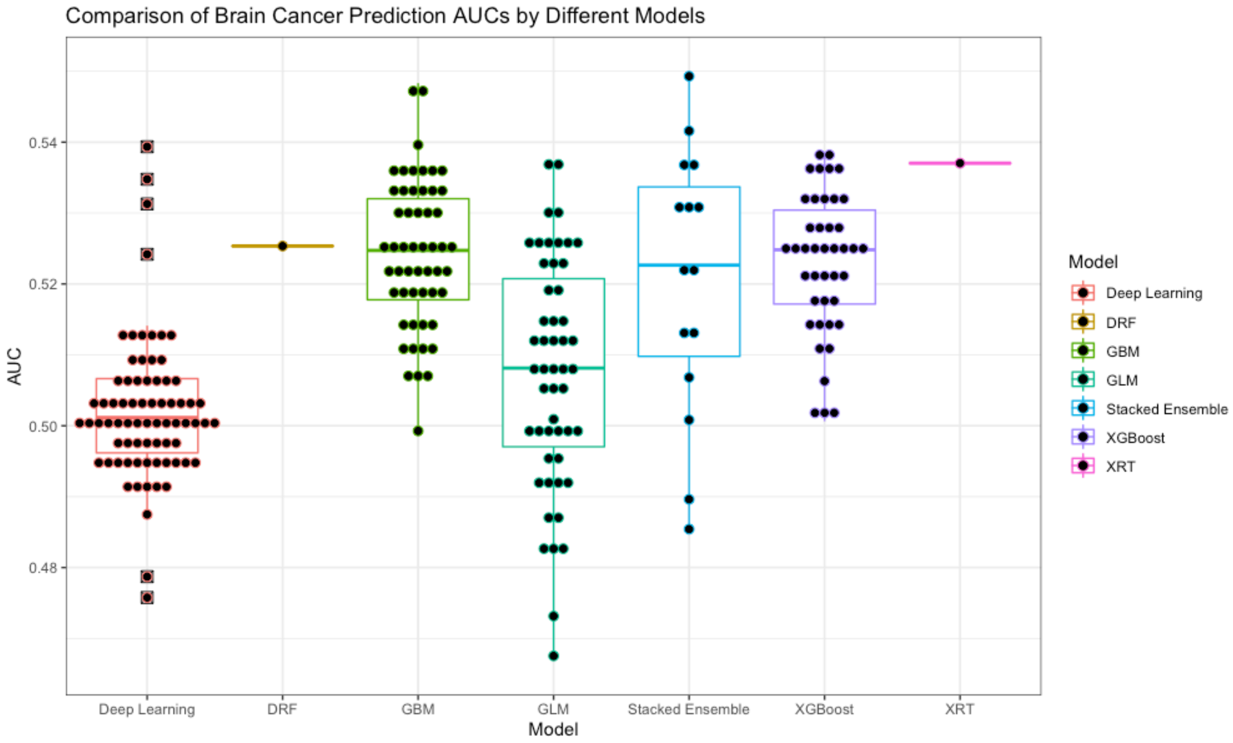

Figure 20: Comparison of Brain Cancer Prediction AUCs by Different Models
The predictive performance, represented by the metric AUC, of different machine learning algorithms. The dataset consists of brain-cancer patients with age- and gender-matched non-cancer individuals from the UK Biobank population as control. H2O was used to carry out a grid search for the best algorithms, and the top-performing models were selected and evaluated with AUCs. The algorithms tested were Deep Learning, Distributed Random Forests (DRF), Gradient Boosting Machine (GBM), General Linear Model (GLM), XGBoost, Extreme Randomized Tree (XRT), and Stacked Ensemble (combination of all the different models).

The AUC value of the ROCs for all machine learning models was 0.513, with a standard deviation of 0.016 and 95% confidence interval of (0.511, 0.515). The AUC differs from 0.5, which is equivalent to an AUC of random chance. Excluding the less optimal models, i.e. the deep learning and GLM models, the average AUC is 0.523, with a standard deviation of 0.012 and 95% confidence interval of (0.521, 0.526). Both AUCs are significantly different from 0.5, with $p<0.00001$.

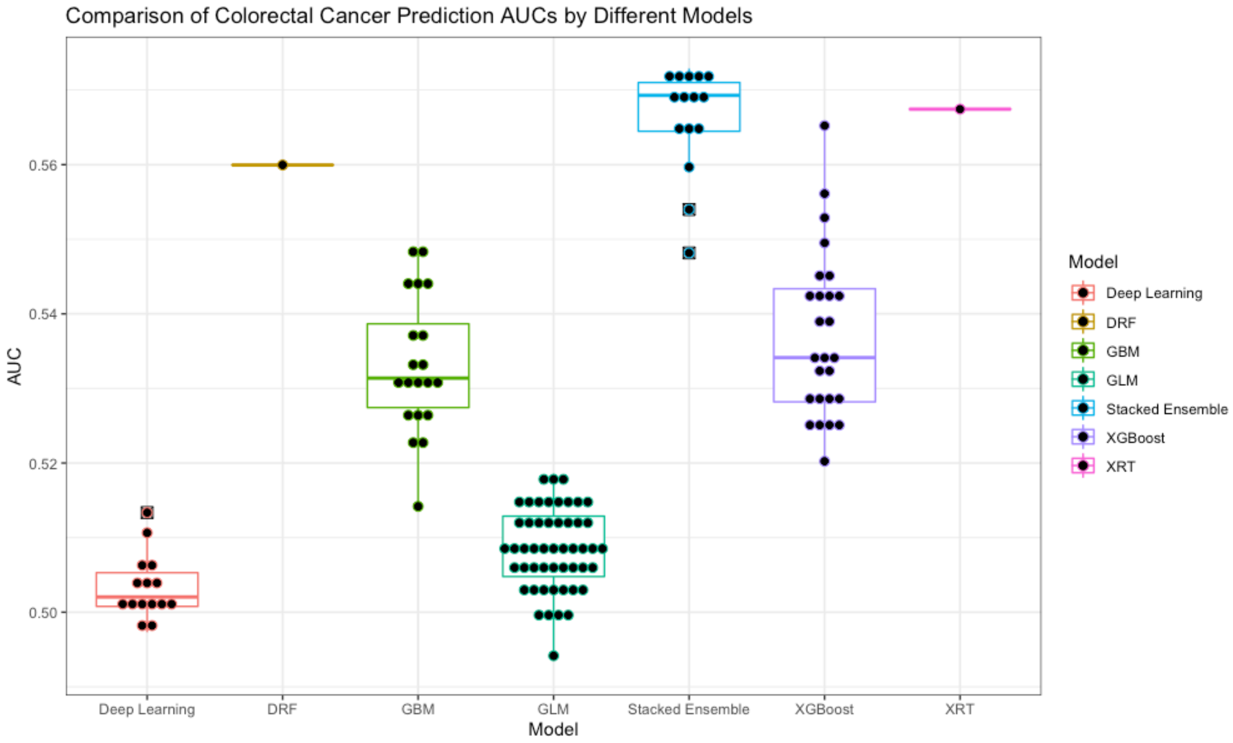

Figure 21: Comparison of Colorectal Cancer Prediction AUCs by Different Models
The predictive performance, represented by the metric AUC, of different machine learning algorithms. The dataset consists of colorectal-cancer patients with age- and gender-matched non-cancer individuals from the UK Biobank population as control. H2O was used to carry out a grid search for the best algorithms, and the top-performing models were selected and evaluated with AUCs. The algorithms tested were Deep Learning, Distributed Random Forests (DRF), Gradient Boosting Machine (GBM), General Linear Model (GLM), XGBoost, Extreme Randomized Tree (XRT), and Stacked Ensemble (combination of all the different models).

The AUC value of the ROCs for all machine learning models was 0.525, with a standard deviation of 0.022 and 95% confidence interval of (0.521, 0.529). The AUC differs from 0.5, which is equivalent to an AUC of random chance. Excluding the less optimal models, i.e. the deep learning models, the average AUC is 0.544 with a standard deviation of 0.017 and 95% confidence interval of (0.539, 0.548). Both AUCs are significantly different from 0.5, with p<0.00001.

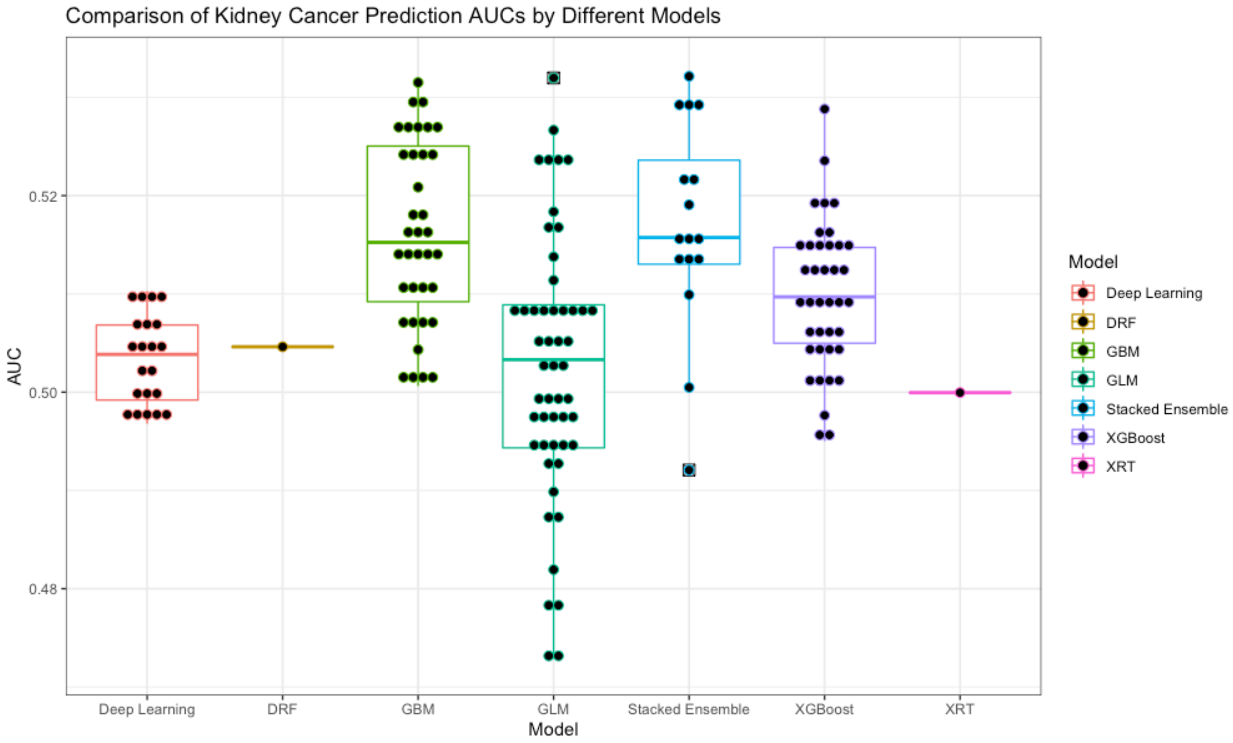

Figure 22: Comparison of Kidney Cancer Prediction AUCs by Different Models
The predictive performance, represented by the metric AUC, of different machine learning algorithms. The dataset consists of kidney-cancer patients with age- and gender-matched non-cancer individuals from the UK Biobank population as control. H2O was used to carry out a grid search for the best algorithms, and the top-performing models were selected and evaluated with AUCs. The algorithms tested were Deep Learning, Distributed Random Forests (DRF), Gradient Boosting Machine (GBM), General Linear Model (GLM), XGBoost, Extreme Randomized Tree (XRT), and Stacked Ensemble (combination of all the different models).

The AUC value of the ROCs for all machine learning models was 0.509, with a standard deviation of 0.012 and 95% confidence interval of (0.507, 0.510). The AUC differs from 0.5, which is equivalent to an AUC of random chance. Excluding the less optimal models, i.e. the deep learning, distributed random forest, and extreme randomized tree models, the average AUC is 0.510, with a standard deviation of 0.009 and 95% confidence interval of (0.508, 0.512). Both AUCs are significantly different from 0.5, with $p<0.00001$.

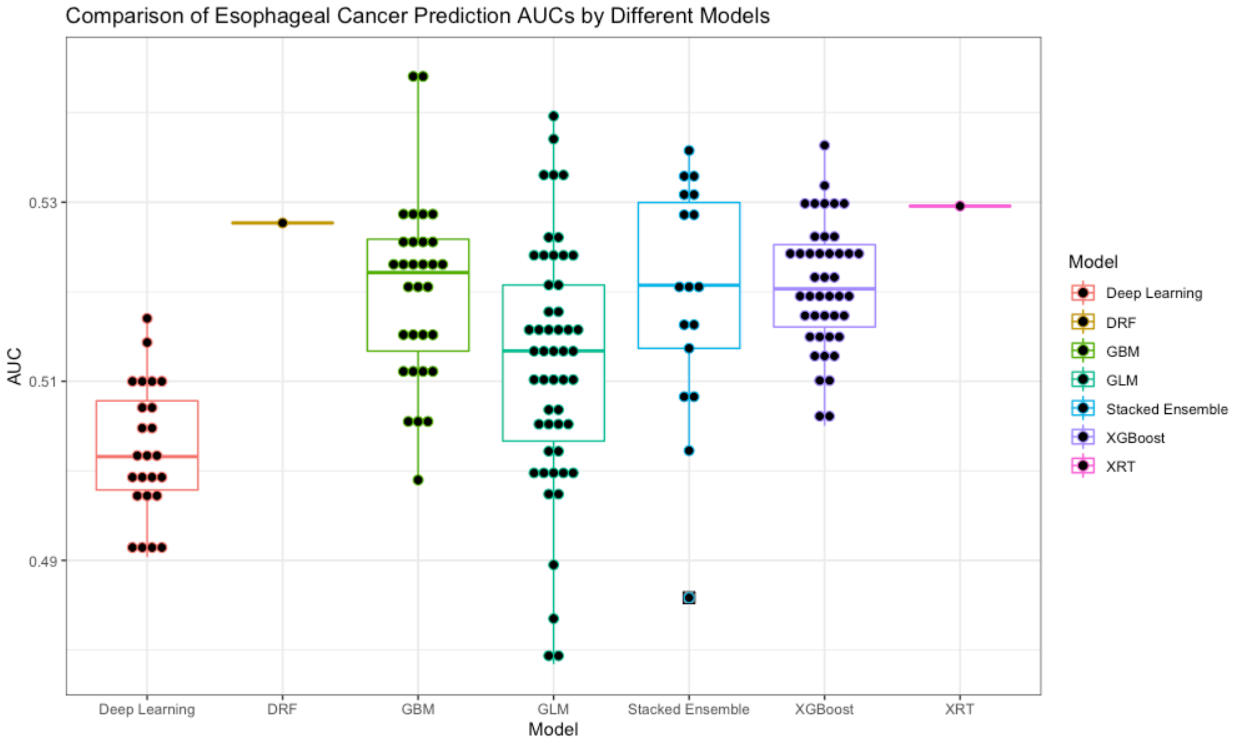

Figure 23: Comparison of Esophageal Cancer Prediction AUCs by Different Models
The predictive performance, represented by the metric AUC, of different machine learning algorithms. The dataset consists of esophageal-cancer patients with age- and gender-matched non-cancer individuals from the UK Biobank population as control. H2O was used to carry out a grid search for the best algorithms, and the top-performing models were selected and evaluated with AUCs. The algorithms tested were Deep Learning, Distributed Random Forests (DRF), Gradient Boosting Machine (GBM), General Linear Model (GLM), XGBoost, Extreme Randomized Tree (XRT), and Stacked Ensemble (combination of all the different models).

The AUC value of the ROCs for all machine learning models was 0.516, with a standard deviation of 0.012 and 95% confidence interval of (0.514, 0.518). The AUC differs from 0.5, which is equivalent to an AUC of random chance, with $p<0.00001$.

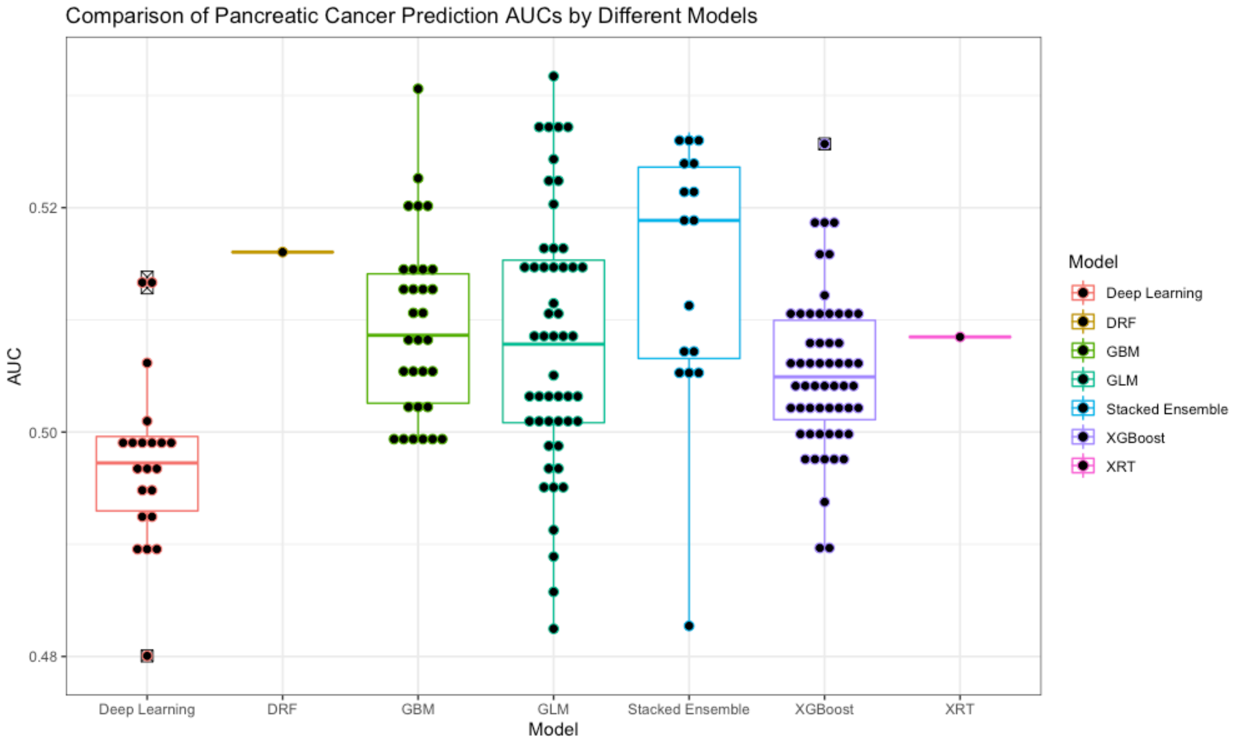

Figure 24: Comparison of Pancreatic Cancer Prediction AUCs by Different Models

The predictive performance, represented by the metric AUC, of different machine learning algorithms. The dataset consists of pancreatic-cancer patients with age- and gender-matched non-cancer individuals from the UK Biobank population as control. H2O was used to carry out a grid search for the best algorithms, and the top-performing models were selected and evaluated with AUCs. The algorithms tested were Deep Learning, Distributed Random Forests (DRF), Gradient Boosting Machine (GBM), General Linear Model (GLM), XGBoost, Extreme Randomized Tree (XRT), and Stacked Ensemble (combination of all the different models).

The AUC value of the ROCs for all machine learning models was 0.507, with a standard deviation of 0.010 and 95% confidence interval of (0.505, 0.508). The AUC differs from 0.5, which is equivalent to an AUC of random chance. Excluding the less optimal models, i.e. the deep learning, models, the average AUC is 0.508, with a standard deviation of 0.010 and 95% confidence interval of (0.506, 0.510). Both AUCs are significantly different from 0.5, with p<0.00001.

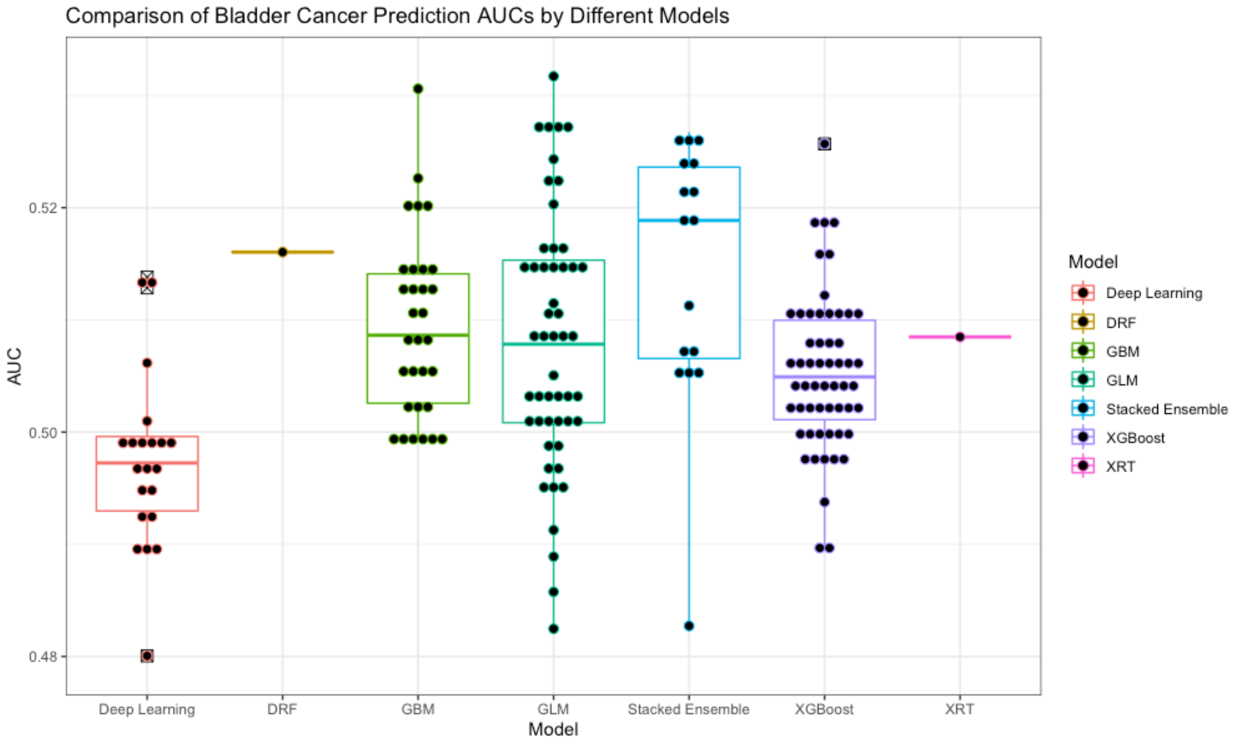

Figure 25: Comparison of Bladder Cancer Prediction AUCs by Different Models
The predictive performance, represented by the metric AUC, of different machine learning algorithms. The dataset consists of bladder-cancer patients with age- and gender-matched non-cancer individuals from the UK Biobank population as control. H2O was used to carry out a grid search for the best algorithms, and the top-performing models were selected and evaluated with AUCs. The algorithms tested were Deep Learning, Distributed Random Forests (DRF), Gradient Boosting Machine (GBM), General Linear Model (GLM), XGBoost, Extreme Randomized Tree (XRT), and Stacked Ensemble (combination of all the different models).

The AUC value of the ROCs for all machine learning models was 0.507, with a standard deviation of 0.010 and 95% confidence interval of (0.505, 0.508). The AUC differs from 0.5, which is equivalent to an AUC of random chance, with p<0.00001.

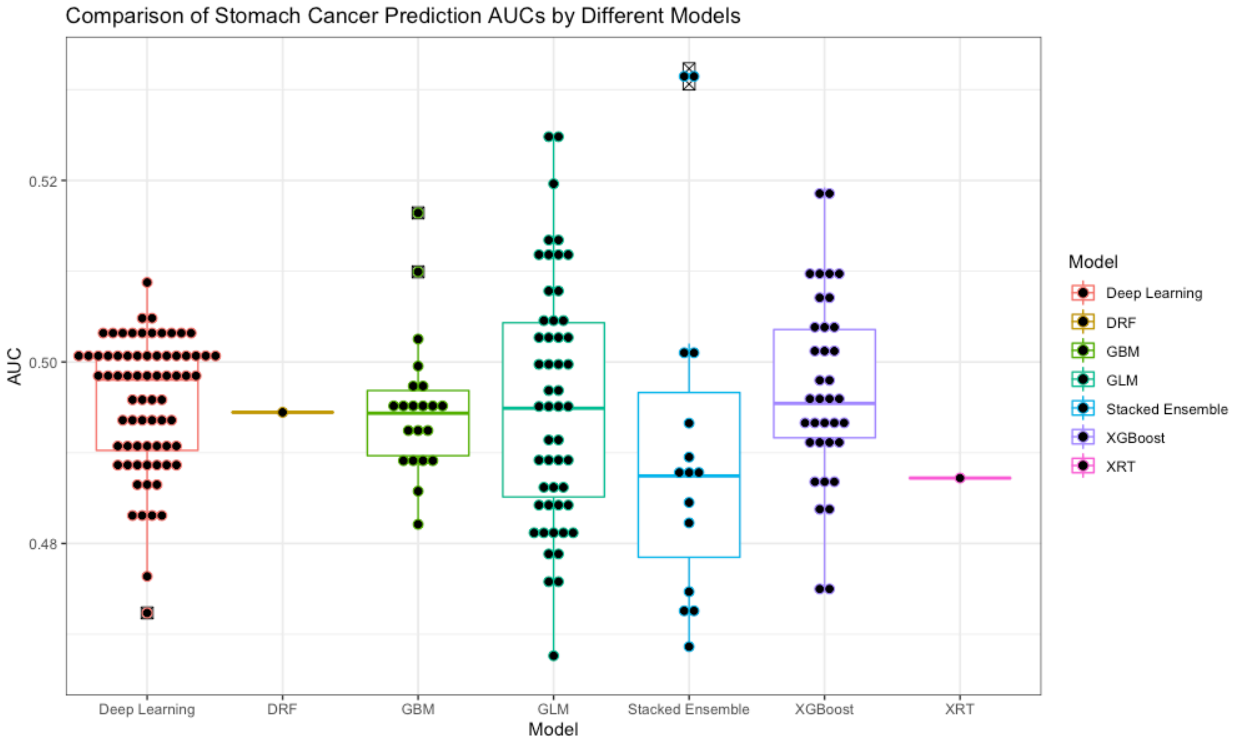

Figure 26: Comparison of Stomach Cancer Prediction AUCs by Different Models
The predictive performance, represented by the metric AUC, of different machine learning algorithms. The dataset consists of prostate-cancer patients with age- and gender-matched non-cancer individuals from the UK Biobank population as control. H2O was used to carry out a grid search for the best algorithms, and the top-performing models were selected and evaluated with AUCs. The algorithms tested were Deep Learning, Distributed Random Forests (DRF), Gradient Boosting Machine (GBM), General Linear Model (GLM), XGBoost, Extreme Randomized Tree (XRT), and Stacked Ensemble (combination of all the different models).

The AUC value of the ROCs for all machine learning models was 0.495, with a standard deviation of 0.011 and 95% confidence interval of (0.490, 0.497). The AUC does not differ from 0.5 significantly.

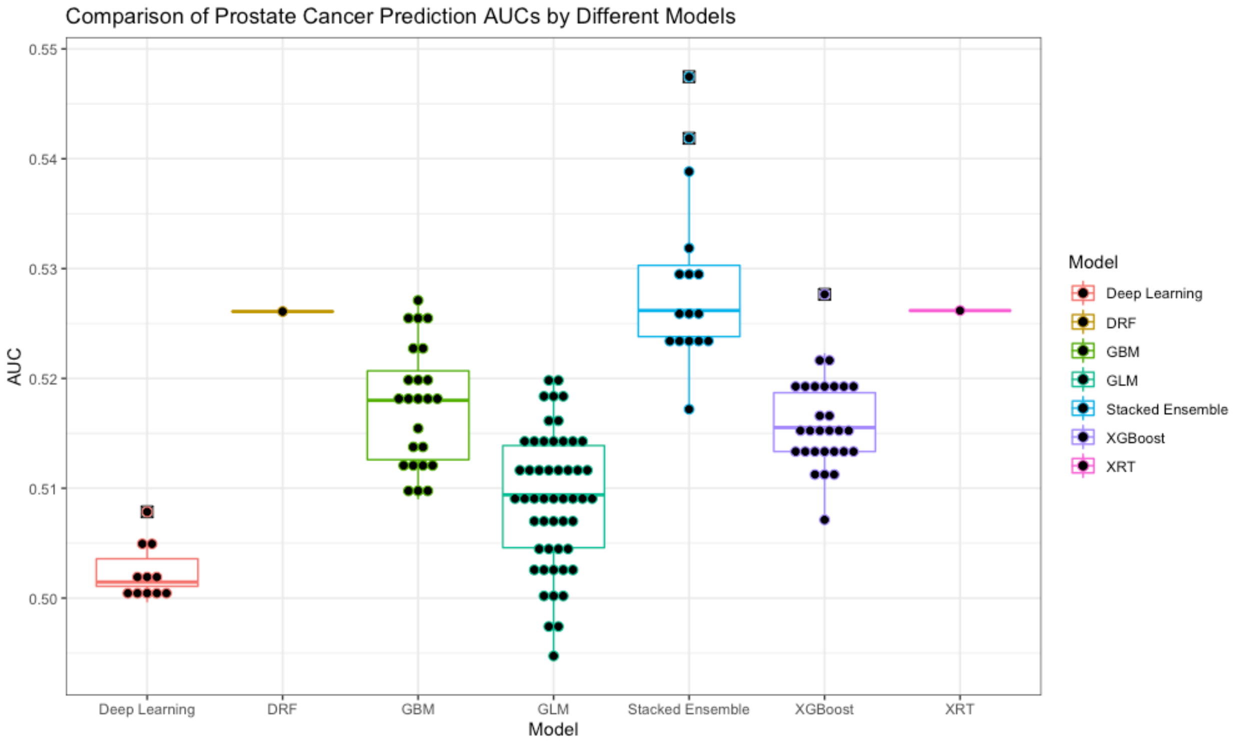

Figure 27: Comparison of Prostate Cancer Prediction AUCs by Different Models
The predictive performance, represented by the metric AUC, of different machine learning algorithms. The dataset consists of prostate-cancer patients with age- and gender-matched non-cancer individuals from the UK Biobank population as control. H2O was used to carry out a grid search for the best algorithms, and the top-performing models were selected and evaluated with AUCs. The algorithms tested were Deep Learning, Distributed Random Forests (DRF), Gradient Boosting Machine (GBM), General Linear Model (GLM), XGBoost, Extreme Randomized Tree (XRT), and Stacked Ensemble (combination of all the different models).

The AUC value of the ROCs for all machine learning models was 0.514, with a standard deviation of 0.009 and 95% confidence interval of (0.512, 0.515). The AUC differs from 0.5, which is equivalent to an AUC of random chance. Excluding the less optimal models, i.e. the deep learning and generalized linear models, the average AUC is 0.520, with a standard deviation of 0.008 and 95% confidence interval of (0.518, 0.521). Both AUCs are significantly different from 0.5, with $p<0.00001$.

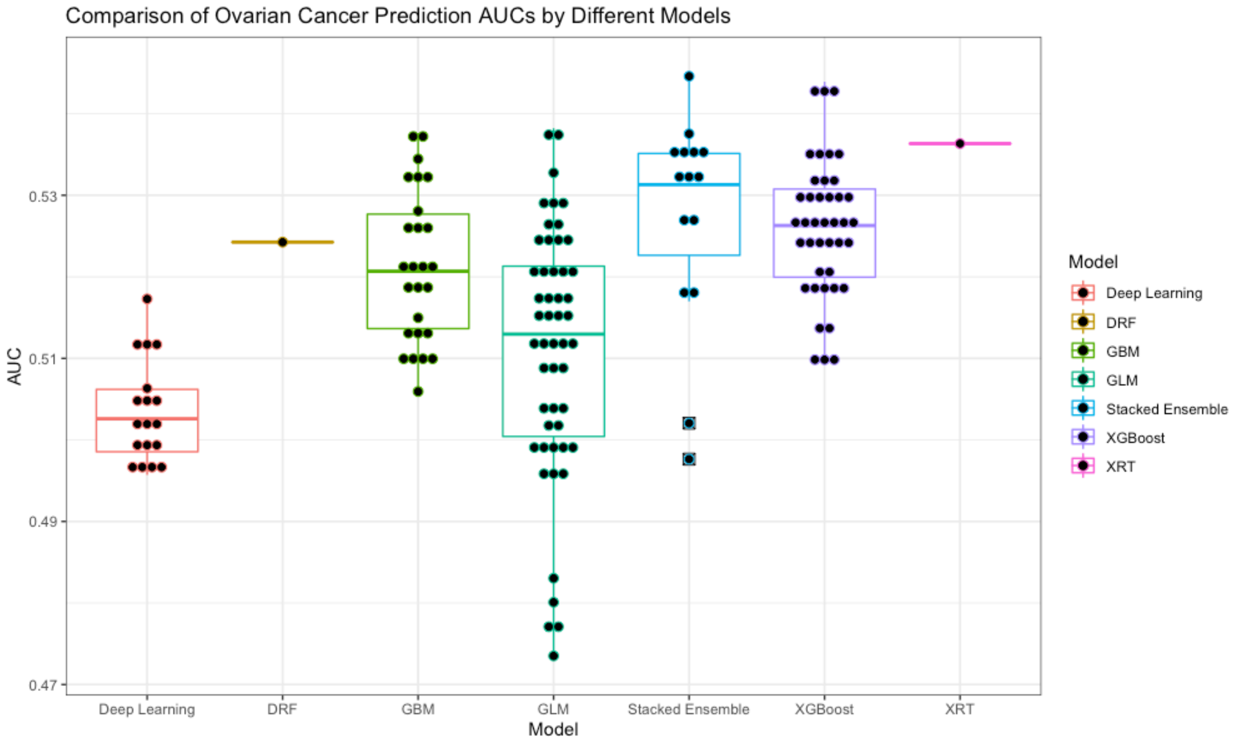

Figure 28: Comparison of Ovarian Cancer Prediction AUC by Different Models
The predictive performance, represented by the metric AUC, of different machine learning algorithms. The dataset consists of ovarian-cancer patients with age- and gender-matched non-cancer individuals from the UK Biobank population as control. H2O was used to carry out a grid search for the best algorithms, and the top-performing models were selected and evaluated with AUCs. The algorithms tested were Deep Learning, Distributed Random Forests (DRF), Gradient Boosting Machine (GBM), General Linear Model (GLM), XGBoost, Extreme Randomized Tree (XRT), and Stacked Ensemble (combination of all the different models).

The AUC value of the ROCs for all machine learning models was 0.518, with a standard deviation of 0.014 and 95% confidence interval of (0.515, 0.520). The AUC is significantly different from 0.5, with $p<0.00001$.

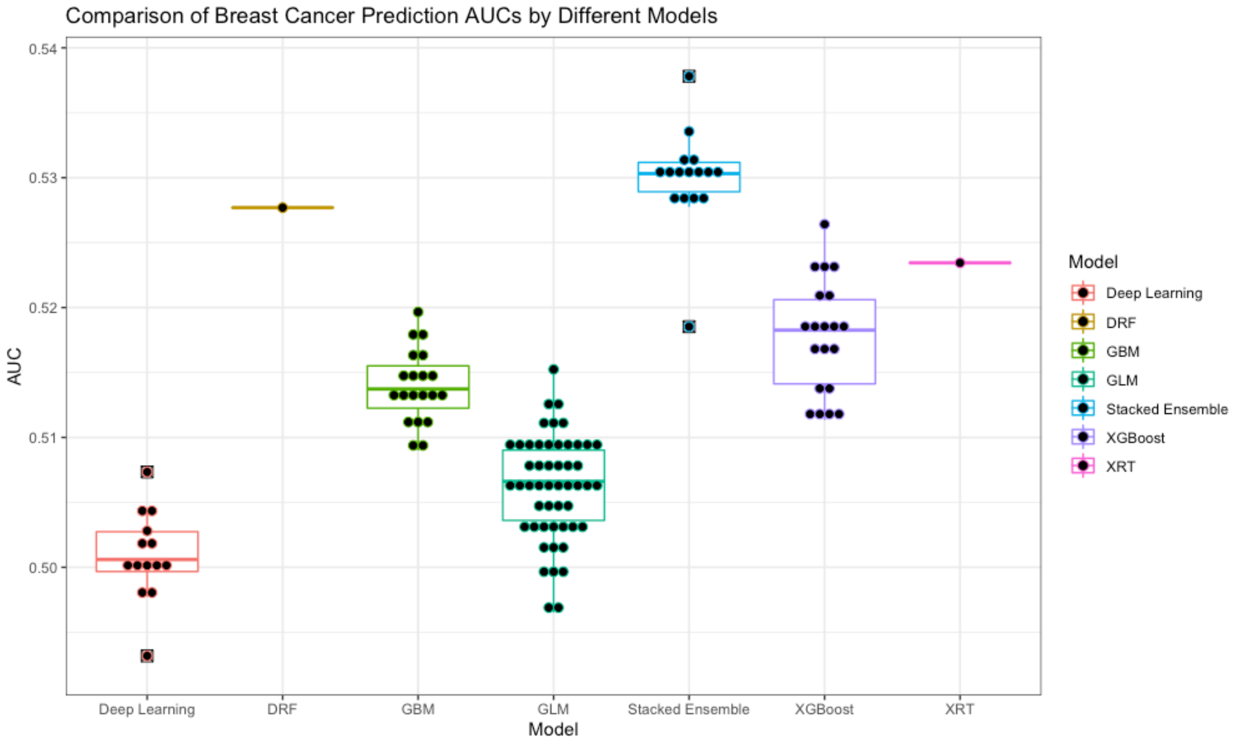

Figure 29: Comparison of Breast Cancer Prediction AUCs by Different Models
The predictive performance, represented by the metric AUC, of different machine learning algorithms. The dataset consists of breast-cancer patients with age- and gender-matched non-cancer individuals from the UK Biobank population as control. H2O was used to carry out a grid search for the best algorithms, and the top-performing models were selected and evaluated with AUCs. The algorithms tested were Deep Learning, Distributed Random Forests (DRF), Gradient Boosting Machine (GBM), General Linear Model (GLM), XGBoost, Extreme Randomized Tree (XRT), and Stacked Ensemble (combination of all the different models).

The AUC value of the ROCs for all machine learning models was 0.512, with a standard deviation of 0.010 and 95% confidence interval of (0.510, 0.514). The AUC differs from 0.5, which is equivalent to an AUC of random chance, with $p<0.00001$.

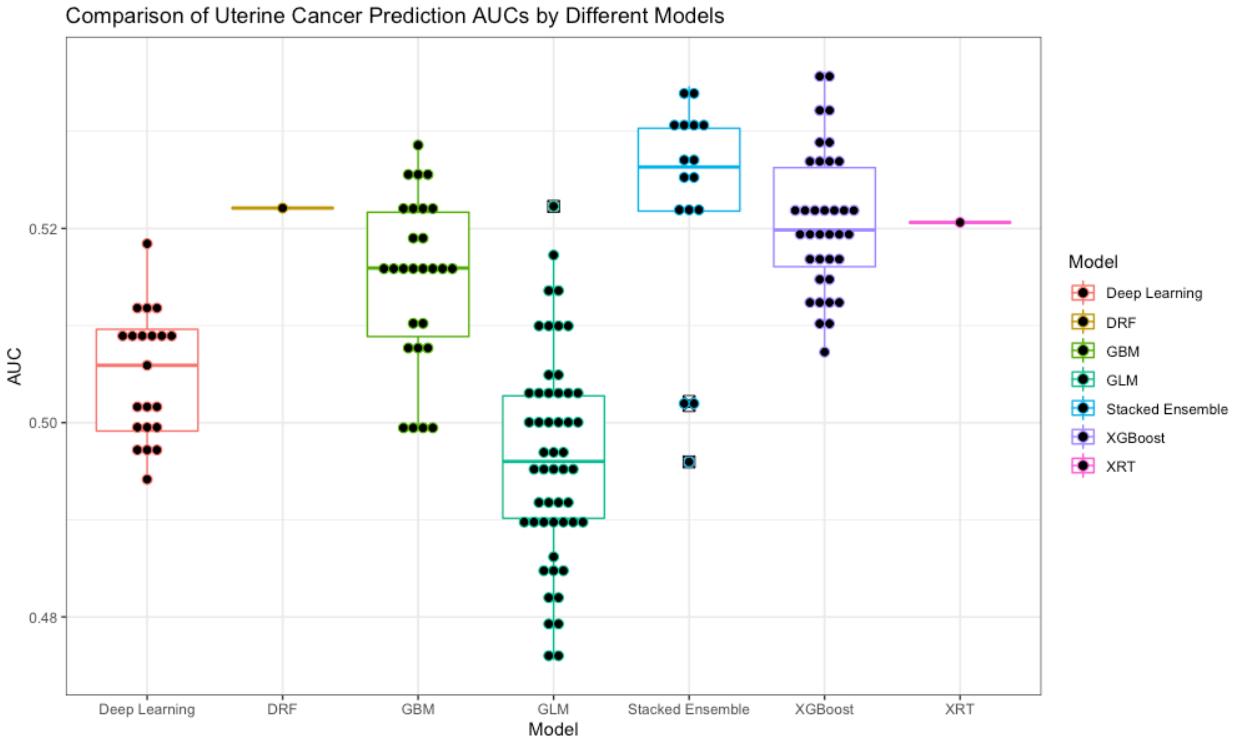

Figure 30: Comparison of Uterine Cancer Prediction AUCs by Different Models
The predictive performance, represented by the metric AUC, of different machine learning algorithms. The dataset consists of uterine-cancer patients with age- and gender-matched non-cancer individuals from the UK Biobank population as control. H2O was used to carry out a grid search for the best algorithms, and the top-performing models were selected and evaluated with AUCs. The algorithms tested were Deep Learning, Distributed Random Forests (DRF), Gradient Boosting Machine (GBM), General Linear Model (GLM), XGBoost, Extreme Randomized Tree (XRT), and Stacked Ensemble (combination of all the different models).

The AUC value of the ROCs for all machine learning models was 0.510, with a standard deviation of 0.014 and 95% confidence interval of (0.507, 0.512). The AUC differs from 0.5, which is equivalent to an AUC of random chance, with p<0.00001.

We also tested whether there was a proportionality between the number of CSLV splits and features and model performance. Five datasets were constructed: 4 splits, 4splits with standard deviation, 8 splits, 8 splits with standard deviation, and TCGA unmasked top 100 CNVs. The 4-splits dataset contains the average l2r value across a quarter of the entire chromosomes (chromosome 1-22, X, Y, and XY, which contains the pseudoautosomal region), resulting in a total of 100 numbers. The 4-splits-with-standard-deviation dataset consists of the 4-splits dataset and the standard deviations of the l2r values of the entire chromosomes, resulting in a total of 125 numbers. The 8-splits dataset contains the average of an eighth of the l2r values of the chromosome, resulting in a total of 200 numbers. The 8-splits-with-standard-deviation dataset consists of the 8-split dataset and the standard deviations of the l2r values of the entire chromosomes, resulting in a total of 225 numbers. The TCGA-unmasked-top-100-CNVs dataset was constructed by mapping the chromosomal location of the top 100 most common CNVs identified in the unmasked data from TCGA to UK Biobank l2r data, then computing the average of each segment. The first four datasets would be referred to as the CSLV sets, and the last dataset would be the CNV set, since it is based on CNVs called in TCGA. We focused on the gradient boosting machine, XGBoost, and stacked ensemble models due to their consistent performances.

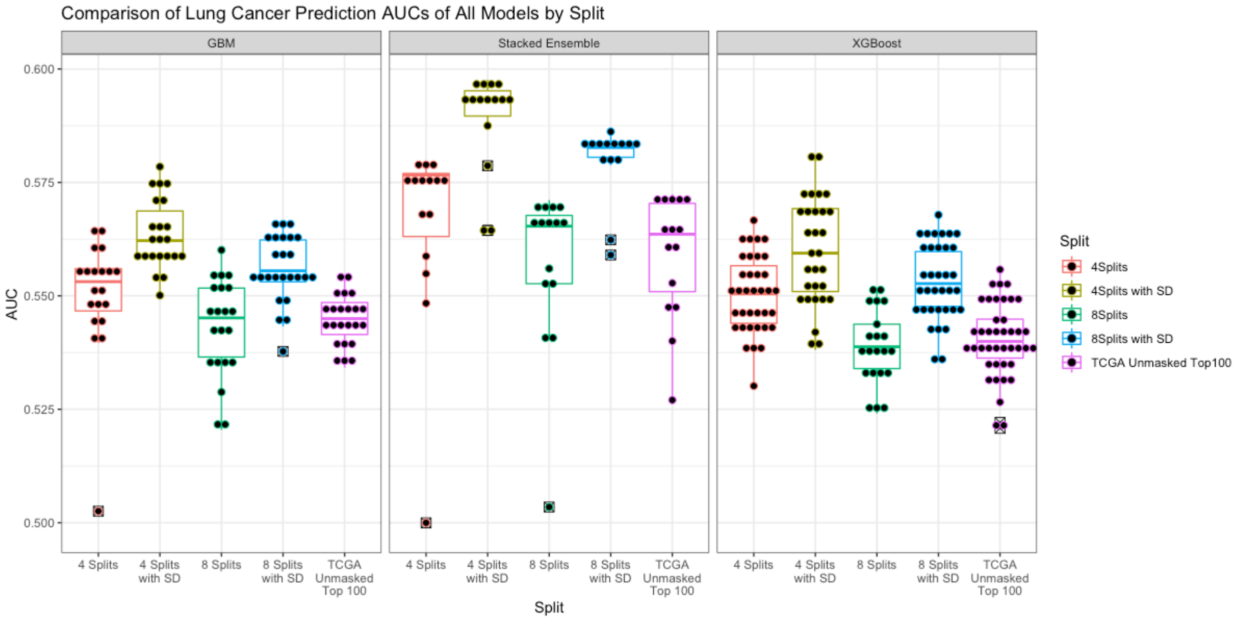

Figure 31: Comparison of Lung Cancer Prediction AUCs by Different Models and by Split

We tested whether the number of splits and features of each chromosome affect predictive performance and how CSLV sets compare to CNV set. We built five datasets by splitting each chromosome into four or eight segments and combining with the standard deviation, and mapping TCGA CNVs to UK Biobank l2r data. The dataset was used to predict whether an individual had lung cancer. The prediction was evaluated by the metric AUC. The plot presents the differences in predictive performance between models and chromosomal scale length variation combinations. The Stacked Ensemble model on the 4-splits-with-standard-deviation dataset performed the best.

Figure 31 demonstrates how these models compare on the five different datasets. The stacked ensemble achieved the best performance. The addition of standard deviation to the split sets improved predictability, but there does not appear to be a proportionality between the number of splits and lung cancer prediction. This still holds true for all models, as shown in Figure 32. We tested whether the performance of the CSLV sets differs significantly from the TCGA-CNV set, and the p-values are recorded in Table 9. The differences in performance were all significant, showing that the CSLV sets outperformed the TCGA-CNV set, and the best dataset was 4-splits-with-standard-deviation.

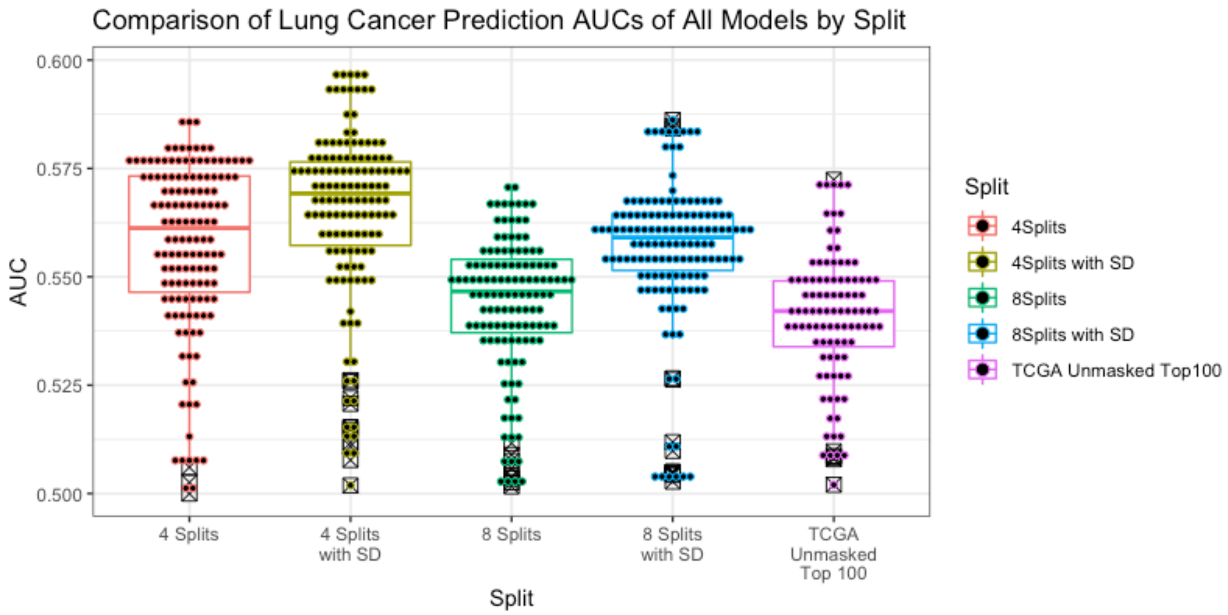

Figure 32: Comparison of Lung Cancer Prediction AUCs of All Models by Split
The average performance of all models for each combination of CSLV features.

Split Numbers	**Mean AUC**	**Standard Deviation**	**95% Confidence Interval**	**p-value vs. TCGA-CNV**
4 Splits	0.557	0.020	(0.554, 0.561)	1.344×10^{-17}
4 Splits with Standard Deviation	0.565	0.020	(0.562, 0.568)	5.149×10^{-28}
8 Splits	0.544	0.016	(0.541, 0.547)	0.018
8 Splits with Standard Deviation	0.556	0.017	(0.553, 0.559)	1.713×10^{-20}

Table 9: Lung-Cancer Prediction AUCs by Split
The mean, standard deviation, and p-values of the cross-validated AUCs of 4-splits, 4-splits-with-standard-deviation, 8-splits, 8-splits-with-standard-deviation datasets.

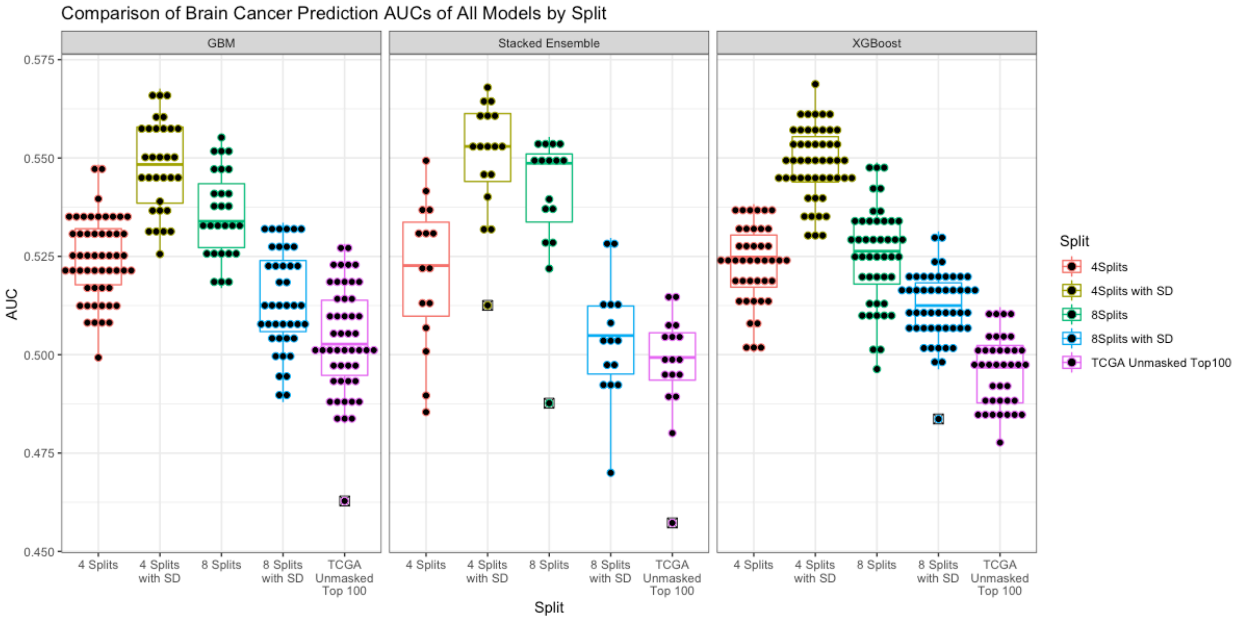

Figure 33: Comparison of Brain Cancer Prediction AUCs by Different Models and by Split

We tested whether the numbers of splits and features of each chromosome affect predictive performance and how CSLV sets compare to CNV set. We built five datasets by splitting each chromosome into four or eight segments and combining with the standard deviation, and mapping TCGA CNVs to UK Biobank l2r data. The dataset was used to predict whether an individual had brain cancer. The prediction was evaluated by the metric AUC. The plot presents the differences in predictive performance between models and chromosomal scale length variation combinations. The Stacked Ensemble model on the 4-splits-with-standard-deviation dataset performed the best.

Figure 33 demonstrates how these models compare on the five different datasets. The stacked ensemble achieved marginally better performance than the other models. The increase in the number of splits leads to greater predictability. The addition of standard deviation to the split sets only improved predictability for the 4-splits set, with 4-splits-with-standard-deviation performing the best. This still holds true for all models, as shown in Figure 34. We tested whether the performance of the CSLV sets differs significantly from the TCGA-CNV set, and the p-values are recorded in Table 10. The differences in performance were all significant, showing that the CSLV sets outperformed the TCGA-CNV set, and the best dataset was 4-splits-with-standard-deviation.

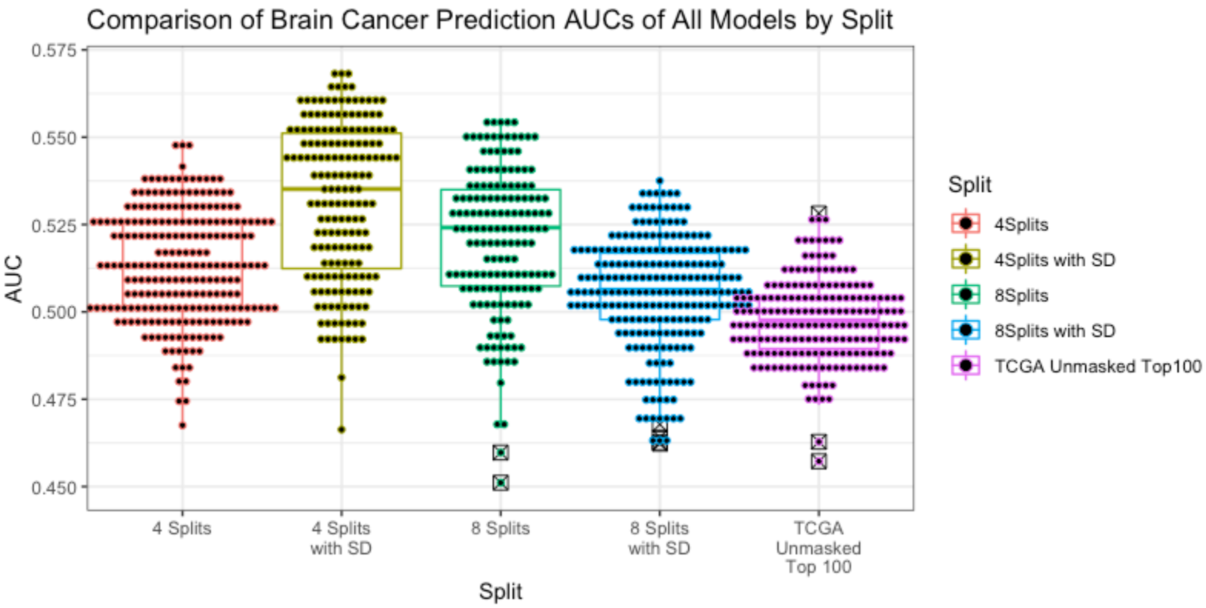

Figure 34: Comparison of Brain Cancer Prediction AUCs of All Models by Split
The average performance of all models for each combination of CSLV features.

Split Numbers	Mean AUC	Standard Deviation	95% Confidence Interval	p-value vs. TCGA-CNV
4 Splits	0.513	0.016	(0.511, 0.515)	2.382×10^{-38}
4 Splits with Standard Deviation	0.532	0.022	(0.529, 0.535)	2.524×10^{-48}
8 Splits	0.520	0.021	(0.517, 0.524)	3.644×10^{-30}
8 Splits with Standard Deviation	0.506	0.016	(0.504, 0.508)	1.112×10^{-13}

Table 10: Brain-Cancer Prediction AUCs by Split
The mean, standard deviation, and p-values of the cross-validated AUCs of 4-splits, 4-splits-with-standard-deviation, 8-splits, 8-splits-with-standard-deviation datasets.

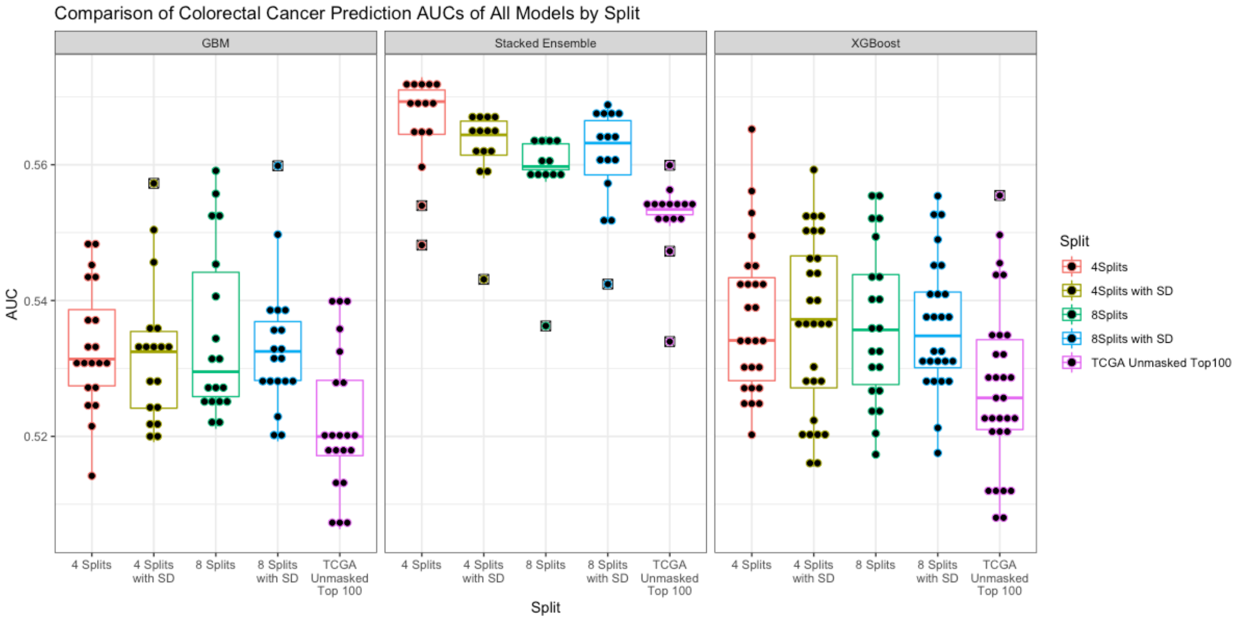

Figure 35: Comparison of Colorectal Cancer Prediction AUCs by Different Models and by Split
We tested whether the numbers of splits and features of each chromosome affect predictive performance and how CSLV sets compare to CNV set. We built five datasets by splitting each chromosome into four or eight segments and combining with the standard deviation, and mapping TCGA CNVs to UK Biobank l2r data. The dataset was used to predict whether an individual had colorectal cancer. The prediction was evaluated by the metric AUC. The plot presents the differences in predictive performance between models and chromosomal scale length variation combinations. The Stacked Ensemble model performed the best, but there was no difference between performances of CSLV- and CNV- sets.

Figure 35 demonstrates how these models compare on the five different datasets. The stacked ensemble achieved better performance than the other models. The increase in the number of splits or the addition of standard deviation did not improve predictability. This still holds true for all models, as shown in Figure 36. We tested whether the performance of the CSLV sets differs significantly from the TCGA-CNV set, and the p-values are recorded in Table 11. The differences in performance were not significant, showing that the CSLV sets performed as well as the TCGA-CNV set.

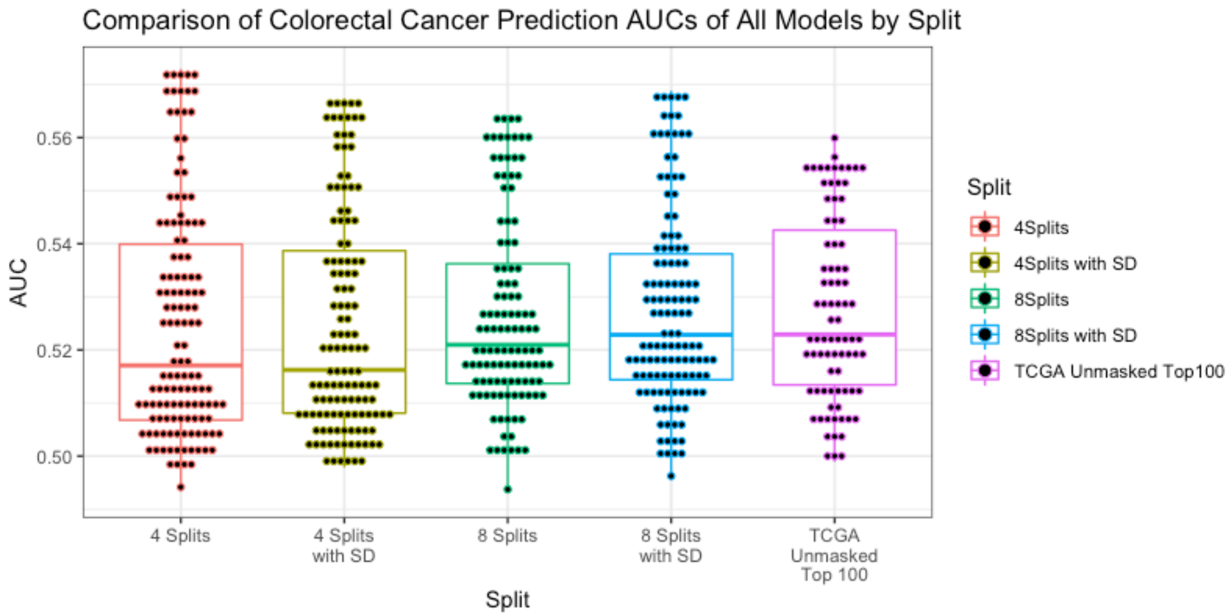

Figure 36: Comparison of Colorectal Cancer Prediction AUCs of All Models by Split
The average performance of all models for each combination of CSLV features.

Split Numbers	**Mean AUC**	**Standard Deviation**	**95% Confidence Interval**	**p-value vs. TCGA-CNV**
4 Splits	0.525	0.022	(0.521, 0.529)	0.901
4 Splits with Standard Deviation	0.525	0.021	(0.521, 0.528)	0.933
8 Splits	0.527	0.018	(0.523, 0.530)	0.710
8 Splits with Standard Deviation	0.528	0.019	(0.524, 0.531)	0.450

Table 11: Colorectal-Cancer Prediction AUCs by Split
The mean, standard deviation, and p-values of the cross-validated AUCs of 4-splits, 4-splits-with-standard-deviation, 8-splits, 8-splits-with-standard-deviation datasets.

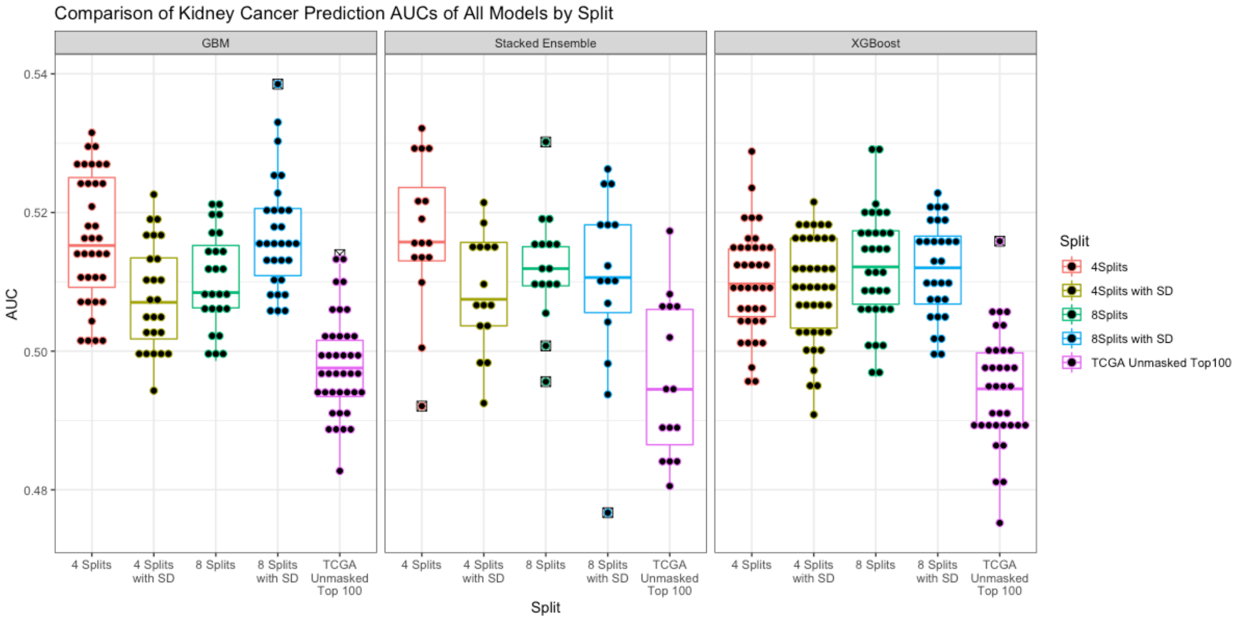

Figure 37: Comparison of Kidney Cancer Prediction AUCs by Different Models and by Split

We tested whether the numbers of splits and features of each chromosome affect predictive performance and how CSLV sets compare to CNV set. We built five datasets by splitting each chromosome into four or eight segments and combining with the standard deviation, and mapping TCGA CNVs to UK Biobank l2r data. The dataset was used to predict whether an individual had kidney cancer. The prediction was evaluated by the metric AUC. The plot presents the differences in predictive performance between models and chromosomal scale length variation combinations. The Stacked Ensemble model on the 8-splits dataset performed the best.

Figure 37 demonstrates how these models compare on the five different datasets. There was no difference between model performance. The increase in the number of splits or the addition of standard deviation did not consistently improve predictability. The 4-splits model performed marginally better than the other datasets. Figure 38 shows the differences for all models for the four datasets. We tested whether the performance of the CSLV sets differs significantly from the TCGA-CNV set, and the p-values are recorded in Table 12. The differences in performance were significant, showing that the CSLV sets performed better than the TCGA-CNV set.

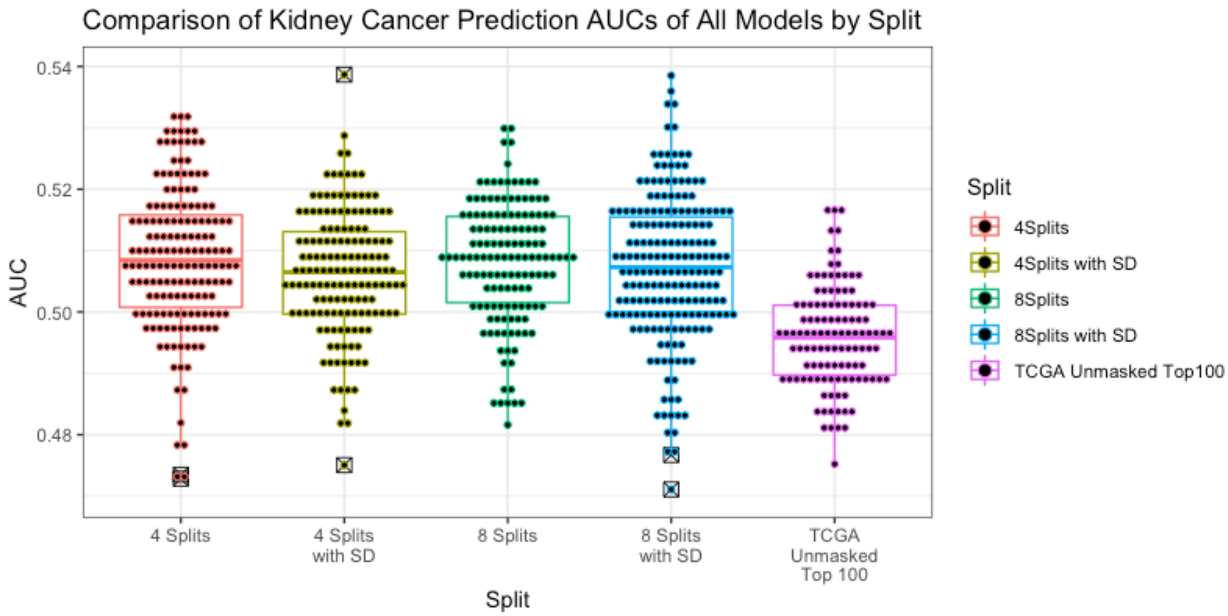

Figure 38: Comparison of Kidney Cancer Prediction AUCs of All Models by Split
The average performance of all models for each combination of CSLV features.

Split Numbers	Mean AUC	Standard Deviation	95% Confidence Interval	p-value vs. TCGA-CNV
4 Splits	0.509	0.012	(0.507, 0.510)	2.280×10^{-30}
4 Splits with Standard Deviation	0.506	0.010	(0.505, 0.508)	1.299×10^{-26}
8 Splits	0.508	0.010	(0.507, 0.510)	2.411×10^{-30}
8 Splits with Standard Deviation	0.507	0.012	(0.505, 0.509)	4.746×10^{-29}

Table 12: Kidney Cancer Prediction AUCs by Split
The mean, standard deviation, and p-values of the cross-validated AUCs of 4-splits, 4-splits-with-standard-deviation, 8-splits, 8-splits-with-standard-deviation datasets.

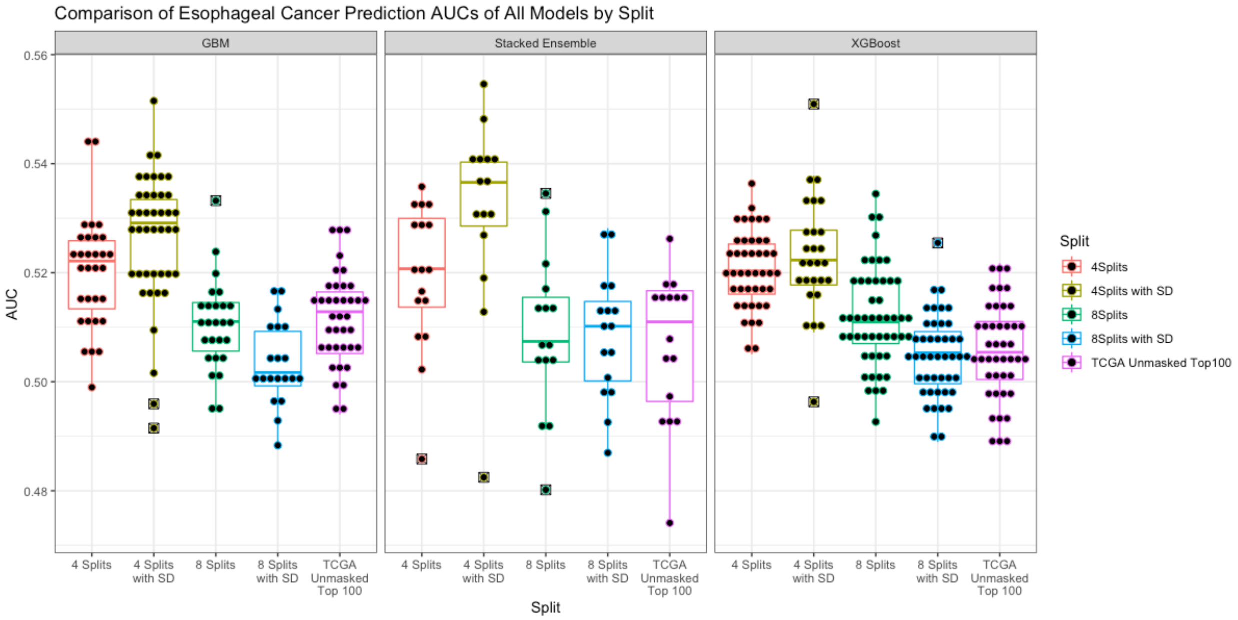

Figure 39: Comparison of Esophageal Cancer Prediction AUCs by Different Models and by Split

We tested whether the numbers of splits and features of each chromosome affect predictive performance and how CSLV sets compare to CNV set. We built five datasets by splitting each chromosome into four or eight segments and combining with the standard deviation, and mapping TCGA CNVs to UK Biobank l2r data. The prediction was evaluated by the metric AUC. The plot presents the differences in predictive performance between models and chromosomal scale length variation combinations. The Stacked Ensemble model on the 4-splits-with-standard-deviation dataset performed the best.

Figure 39 demonstrates how these models compare on the five different datasets. The stacked ensemble achieved marginally better performance than the other models. The increase in the number of splits did not improve predictability. The addition of standard deviation to the split sets improved predictability for the 4-splits model. This still holds true for all models, as shown in Figure 40. We tested whether the performance of the CSLV sets differs significantly from the TCGA-CNV set, and the p-values are recorded in Table 13. The differences in performance were significant for all datasets except 8-splits-with-SD set, showing that most of the CSLV sets performed better than the TCGA-CNV set.

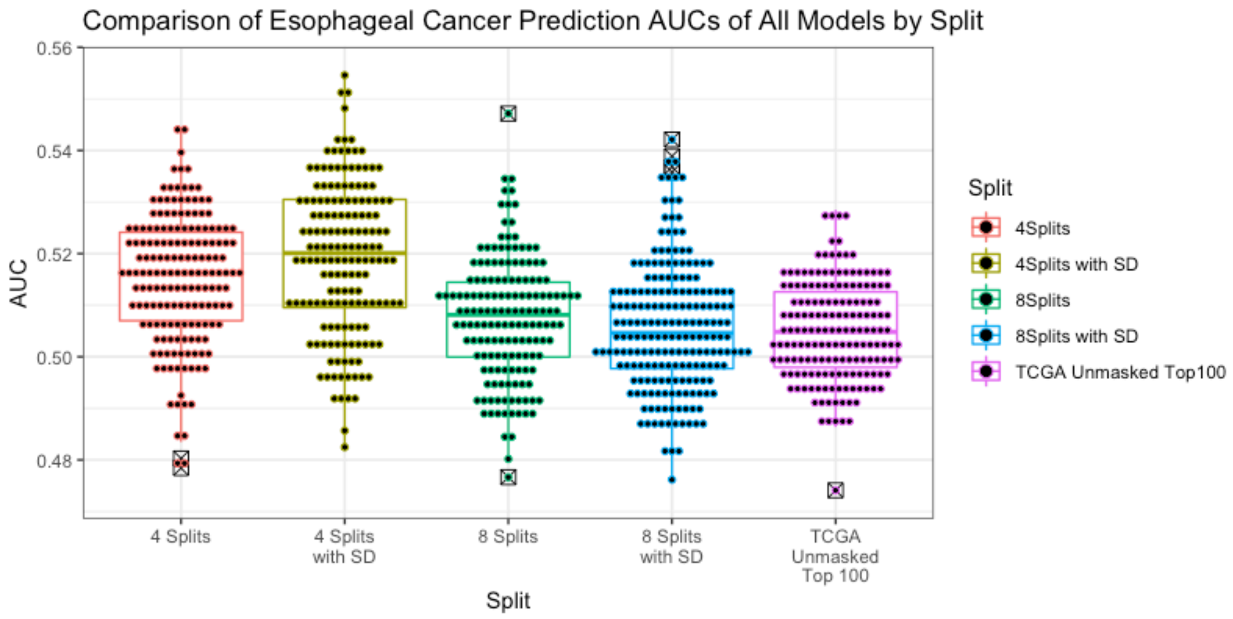

Figure 40: Comparison of Esophageal Cancer Prediction AUCs of All Models by Split
The average performance of all models for each combination of CSLV features.

Split Numbers	Mean AUC	Standard Deviation	95% Confidence Interval	p-value vs. TCGA-CNV
4 Splits	0.515	0.012	(0.513, 0.517)	4.600×10^{-21}
4 Splits with Standard Deviation	0.519	0.015	(0.517, 0.522)	7.937×10^{-26}
8 Splits	0.508	0.012	(0.506, 0.509)	0.004
8 Splits with Standard Deviation	0.506	0.012	(0.504, 0.507)	0.246
TCGA Unmasked Top 100	0.505	0.010	(0.504, 0.506)	

Table 13: Esophageal Cancer Prediction AUCs by Split
The mean, standard deviation, and p-values of the cross-validated AUCs of 4-splits, 4-splits-with-standard-deviation, 8-splits, 8-splits-with-standard-deviation, and TCGA-unmasked-top-100 datasets.

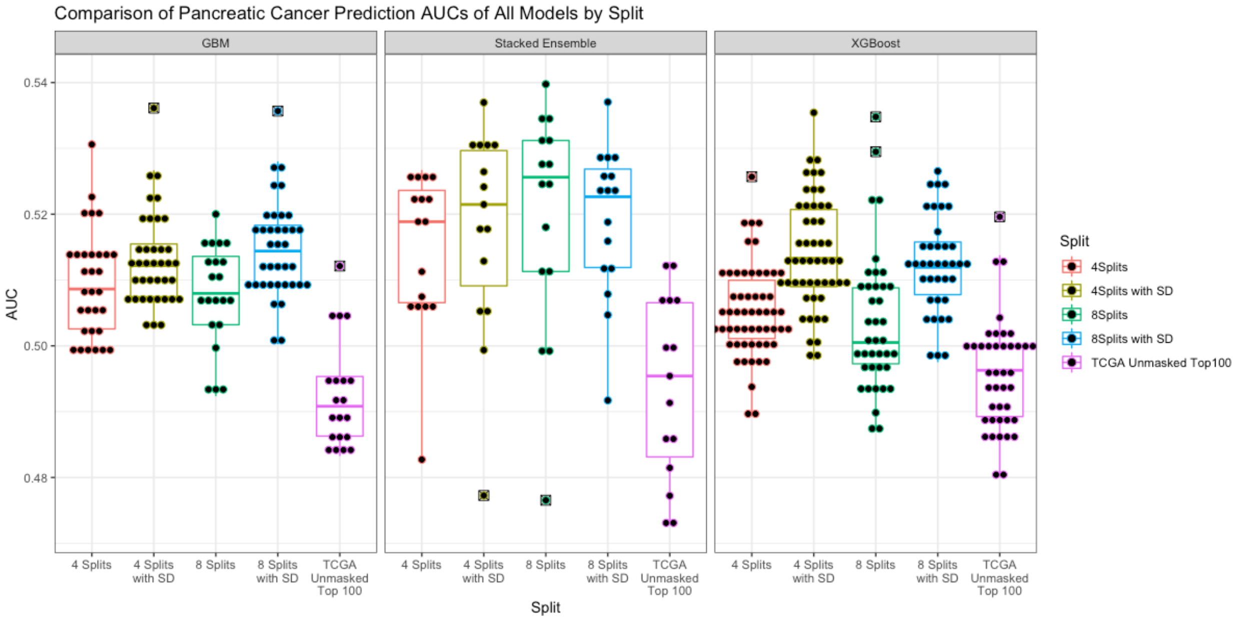

Figure 41: Comparison of Pancreatic Cancer Prediction AUCs by Different Models and by Split
We tested whether the numbers of splits and features of each chromosome affect predictive performance and how CSLV sets compare to CNV set. We built five datasets by splitting each chromosome into four or eight segments and combining with the standard deviation, and mapping TCGA CNVs to UK Biobank l2r data. The dataset was used to predict whether an individual had pancreatic cancer. The prediction was evaluated by the metric AUC. The plot presents the differences in predictive performance between models and chromosomal scale length variation combinations. The Stacked Ensemble model on the 4-splits-with-standard-deviation dataset performed the best.

Figure 41 demonstrates how these models compare on the five different datasets. The stacked ensemble achieved marginally better performance than the other models. The increase in the number of splits did not improve predictability, but the addition of standard deviation to the split sets did. This holds true for all models, as shown in Figure 42. We tested whether the performance of the CSLV sets differs significantly from the TCGA-CNV set, and the p-values are recorded in Table 14. The differences in performance were significant for all datasets except 8-splits-with-SD set, showing that most of the CSLV sets performed better than the TCGA-CNV set.

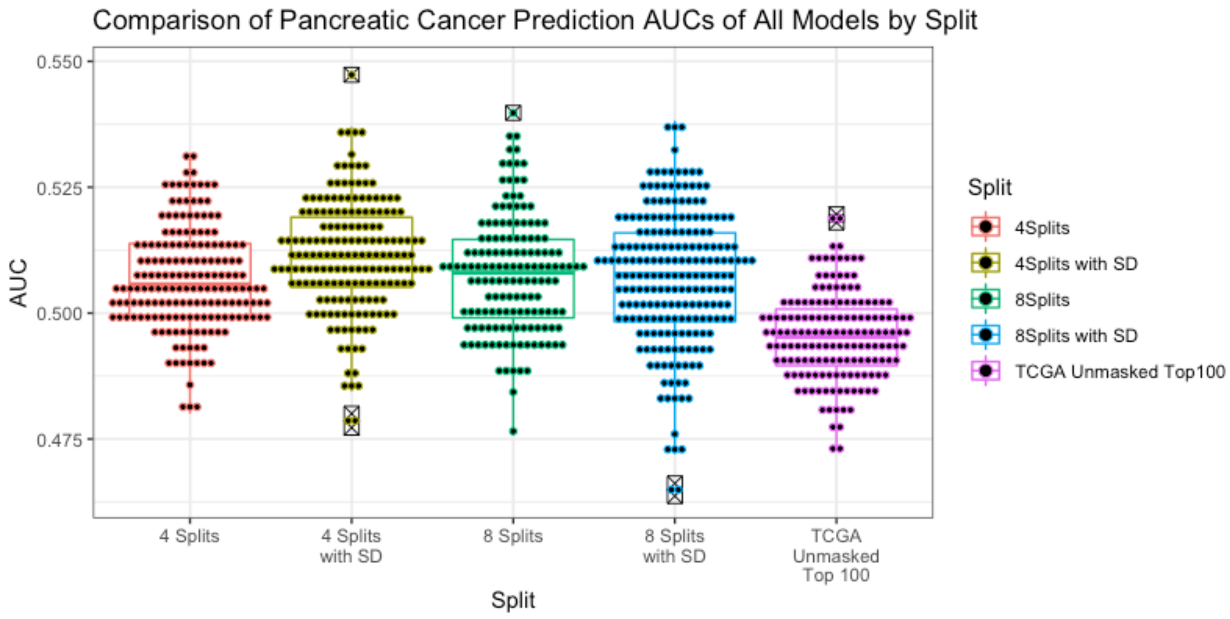

Figure 42: Comparison of Pancreatic Cancer Prediction AUCs of All Models by Split
The average performance of all models for each combination of CSLV features.

Split Numbers	Mean AUC	Standard Deviation	95% Confidence Interval	p-value vs. TCGA-CNV
4 Splits	0.507	0.010	(0.505, 0.508)	1.301×10^{-32}
4 Splits with Standard Deviation	0.511	0.011	(0.510, 0.513)	2.853×10^{-44}
8 Splits	0.508	0.012	(0.506, 0.510)	1.129×10^{-25}
8 Splits with Standard Deviation	0.507	0.013	(0.505, 0.509)	1.923×10^{-27}
TCGA Unmasked Top 100	0.496	0.009	(0.494, 0.497)	

Table 14: Pancreatic Cancer Prediction AUCs by Split
The mean, standard deviation, and p-values of the cross-validated AUCs of 4-splits, 4-splits-with-standard-deviation, 8-splits, 8-splits-with-standard-deviation, and TCGA-unmasked-top-100 datasets.

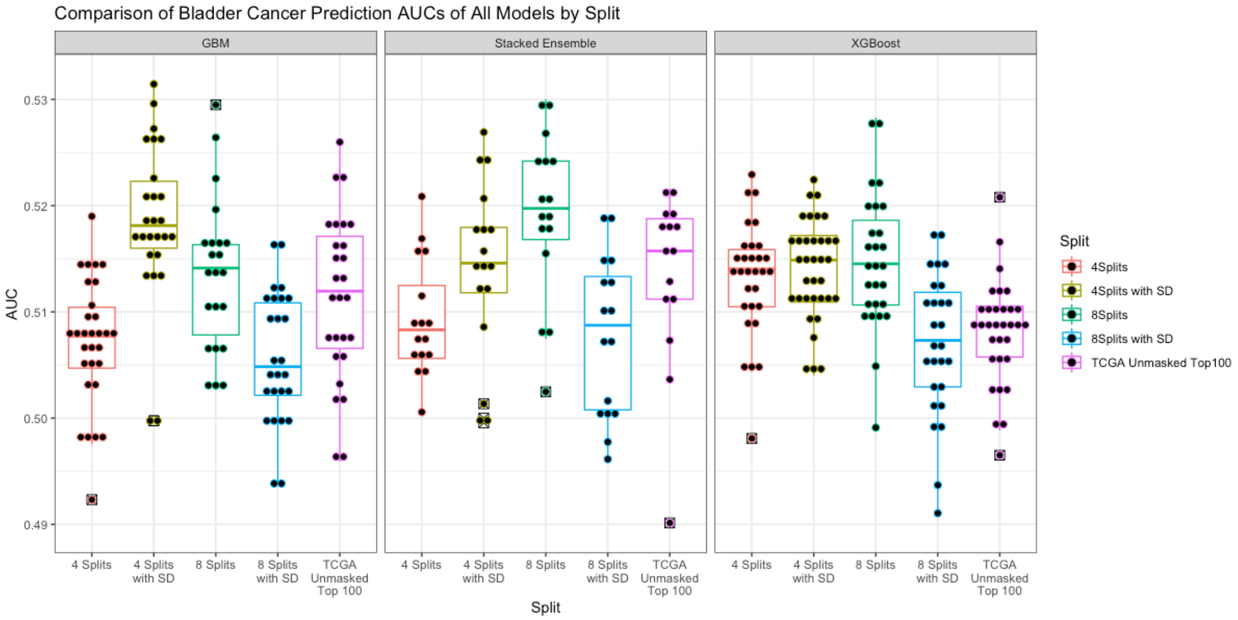

Figure 43: Comparison of Bladder Cancer Prediction AUCs by Different Models and by Split
We tested whether the numbers of splits and features of each chromosome affect predictive performance and how CSLV sets compare to CNV set. We built five datasets by splitting each chromosome into four or eight segments and combining with the standard deviation, and mapping TCGA CNVs to UK Biobank l2r data. The dataset was used to predict whether an individual had bladder cancer. The prediction was evaluated by the metric AUC. The plot presents the differences in predictive performance between models and chromosomal scale length variation combinations. The Stacked Ensemble model on the 4-splits-with-standard-deviation dataset performed the best.

Figure 43 demonstrates how these models compare on the five different datasets. The stacked ensemble achieved marginally better performance with certain datasets. The increase in the number of splits and the addition of standard deviation did not improve predictability consistently. Figure 44 shows the differences in performance of all models for the five datasets. We tested whether the performance of the CSLV sets differs significantly from the TCGA-CNV set, and the p-values are recorded in Table 15. The differences in performance were significant for all datasets except 8-splits-with-SD set, showing that most of the CSLV sets performed better than the TCGA-CNV set.

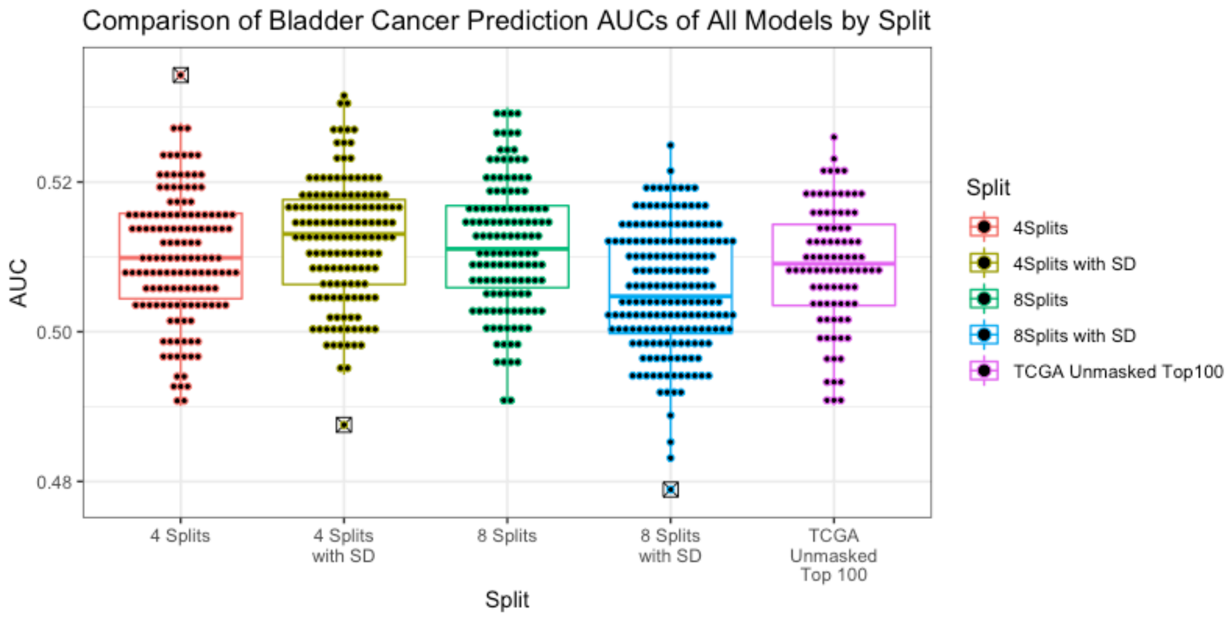

Figure 44: Comparison of Bladder Cancer Prediction AUCs of All Models by Split
The average performance of all models for each combination of CSLV features.

Split Numbers	Mean AUC	Standard Deviation	95% Confidence Interval	p-value vs. TCGA-CNV
4 Splits	0.510	0.009	(0.509, 0.511)	0.030
4 Splits with Standard Deviation	0.512	0.008	(0.511, 0.513)	5.858×10^{-7}
8 Splits	0.511	0.009	(0.510, 0.513)	1.269×10^{-4}
8 Splits with Standard Deviation	0.506	0.008	(0.504, 0.507)	1
TCGA Unmasked Top 100	0.509	0.008	(0.507, 0.510)	

Table 15: Bladder Cancer Prediction AUCs by Split
The mean, standard deviation, and p-values of the cross-validated AUCs of 4-splits, 4-splits-with-standard-deviation, 8-splits, 8-splits-with-standard-deviation, and TCGA-unmasked-top-100 datasets.

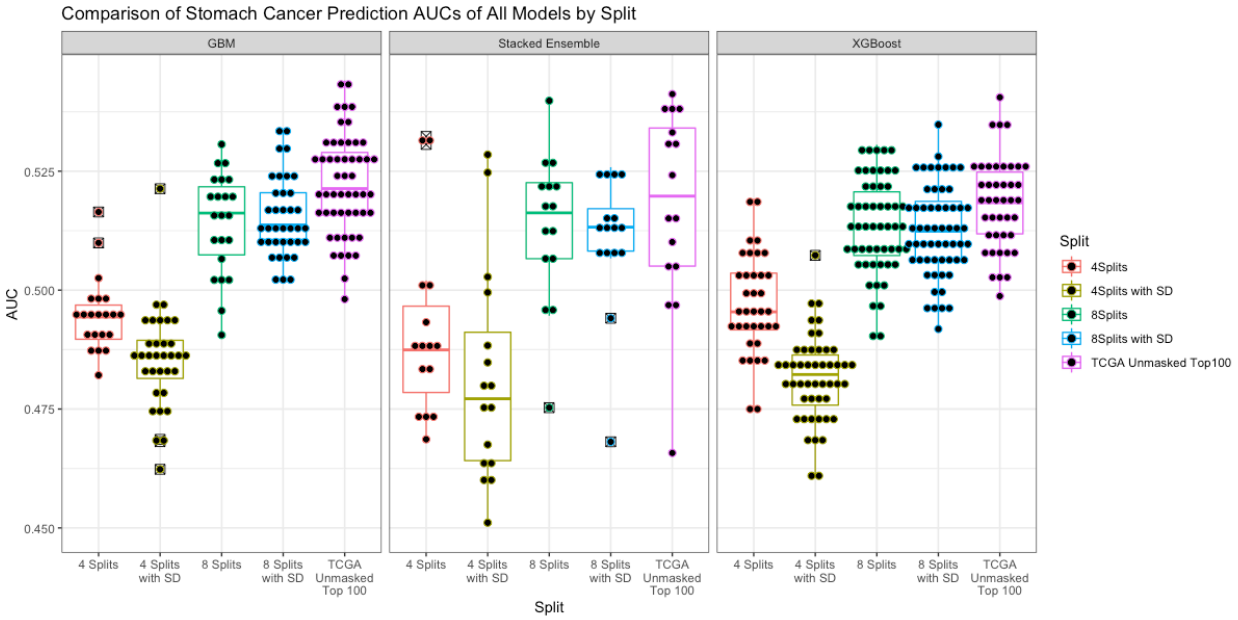

Figure 45: Comparison of Stomach Cancer Prediction AUCs by Different Models and by Split
We tested whether the numbers of splits and features of each chromosome affect predictive performance and how CSLV sets compare to CNV set. We built five datasets by splitting each chromosome into four or eight segments and combining with the standard deviation, and mapping TCGA CNVs to UK Biobank l2r data. The dataset was used to predict whether an individual had stomach cancer. The prediction was evaluated by the metric AUC. The plot presents the differences in predictive performance between models and chromosomal scale length variation combinations. The CNV dataset performed the best, regardless of machine learning algorithms.

Figure 45 demonstrates how these models compare on the five different datasets. The performances of the selected models were comparable between all datasets. The increase in the number of splits improved predictability consistently, but the TCGA-CNV set appeared to outperform the CSLV sets. Figure 46 shows the differences in performance of all models for the five datasets. We tested whether the performance of the CSLV sets differs significantly from the TCGA-CNV set, and the p-values are recorded in Table 16. The differences in performance are significant for all datasets, showing that the CSLV sets were outperformed by the TCGA-CNV set.

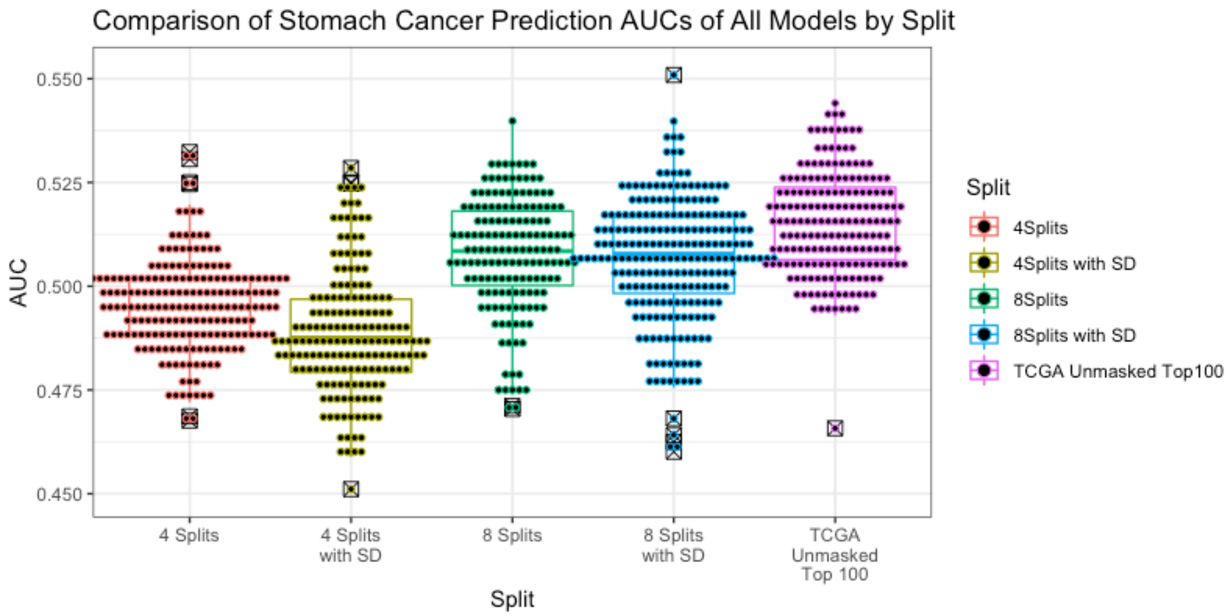

Figure 46: Comparison of Stomach Cancer Prediction AUCs of All Models by Split
The average performance of all models for each combination of CSLV features.

Split Numbers	Mean AUC	Standard Deviation	95% Confidence Interval	p-value vs. TCGA-CNV
4 Splits	0.495	0.011	(0.494, 0.497)	1.264×10^{-64}
4 Splits with Standard Deviation	0.488	0.015	(0.486, 0.491)	6.387×10^{-57}
8 Splits	0.508	0.014	(0.506, 0.510)	1.635×10^{-10}
8 Splits with Standard Deviation	0.507	0.015	(0.505, 0.508)	1.274×10^{-16}
TCGA Unmasked Top 100	0.515	0.012	(0.513, 0.517)	

Table 16: Stomach Cancer Prediction AUCs by Split
The mean, standard deviation, and p-values of the cross-validated AUCs of 4-splits, 4-splits-with-standard-deviation, 8-splits, 8-splits-with-standard-deviation, and TCGA-unmasked-top-100 datasets.

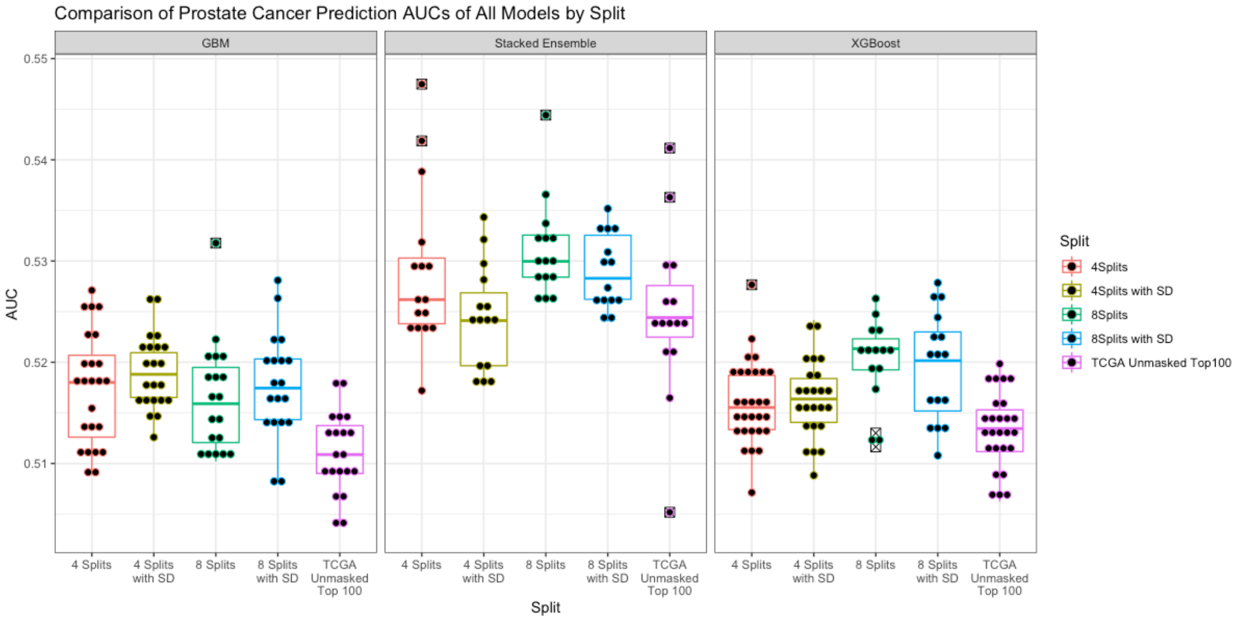

Figure 47: Comparison of Prostate Cancer Prediction AUCs by Different Models and by Split
We tested whether the numbers of splits and features of each chromosome affect predictive performance and how CSLV sets compare to CNV set. We built five datasets by splitting each chromosome into four or eight segments and combining with the standard deviation, and mapping TCGA CNVs to UK Biobank l2r data. The dataset was used to predict whether an individual had prostate cancer. The prediction was evaluated by the metric AUC. The plot presents the differences in predictive performance between models and chromosomal scale length variation combinations. The Stacked Ensemble model performed the best, while the CSLV and CNV sets achieved similar performance.

Figure 47 demonstrates how these models compare on the five different datasets. The stacked ensemble achieved the best performance. The increase in the number of splits and the addition of standard deviation did not consistently improve predictability. Figure 48 shows the differences in performance of all models for the five datasets. We tested whether the performance of the CSLV sets differs significantly from the TCGA-CNV set, and the p-values are recorded in Table 17. The differences in performance are not significant for all datasets, showing that the CSLV sets performed as well as the TCGA-CNV set.

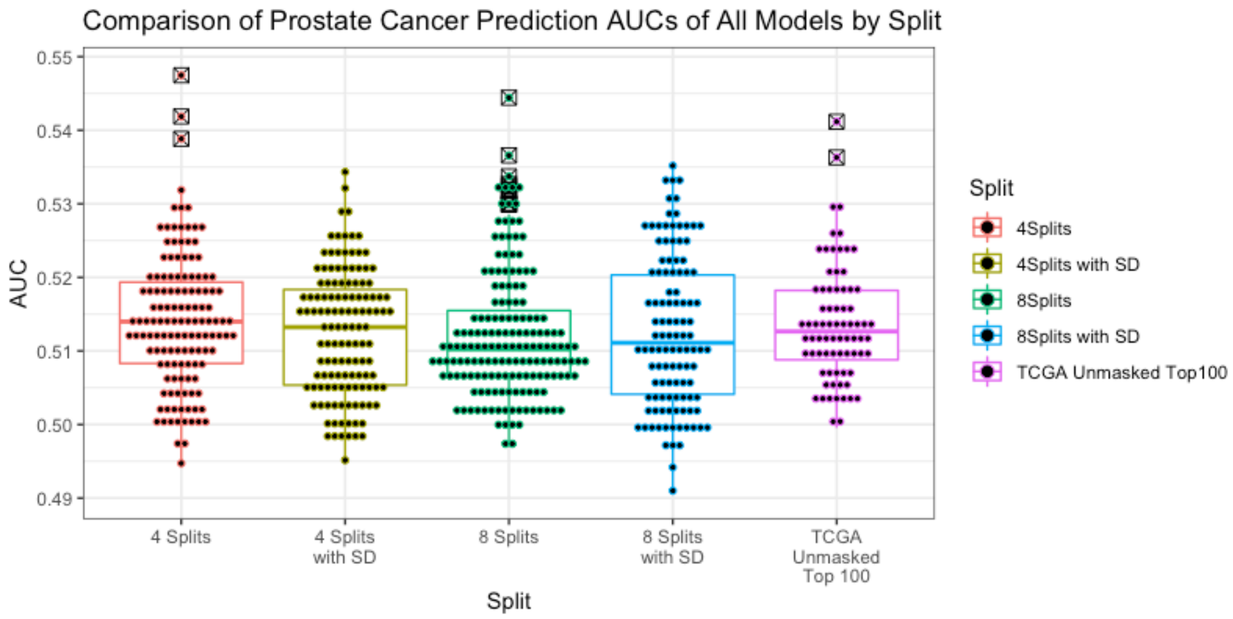

Figure 48: Comparison of Prostate Cancer Prediction AUCs of All Models by Split
The average performance of all models for each combination of CSLV features.

Split Numbers	**Mean AUC**	**Standard Deviation**	**95% Confidence Interval**	**p-value vs. TCGA-CNV**
4 Splits	0.514	0.009	(0.513, 0.516)	0.343
4 Splits with Standard Deviation	0.513	0.008	(0.511, 0.514)	0.957
8 Splits	0.512	0.009	(0.511, 0.513)	0.993
8 Splits with Standard Deviation	0.513	0.010	(0.511, 0.515)	0.894
TCGA Unmasked Top 100	0.514	0.008	(0.512, 0.516)	

Table 17: Prostate Cancer Prediction AUCs by Split
The mean, standard deviation, and p-values of the cross-validated AUCs of 4-splits, 4-splits-with-standard-deviation, 8-splits, 8-splits-with-standard-deviation, TCGA-unmasked-top-100 datasets.

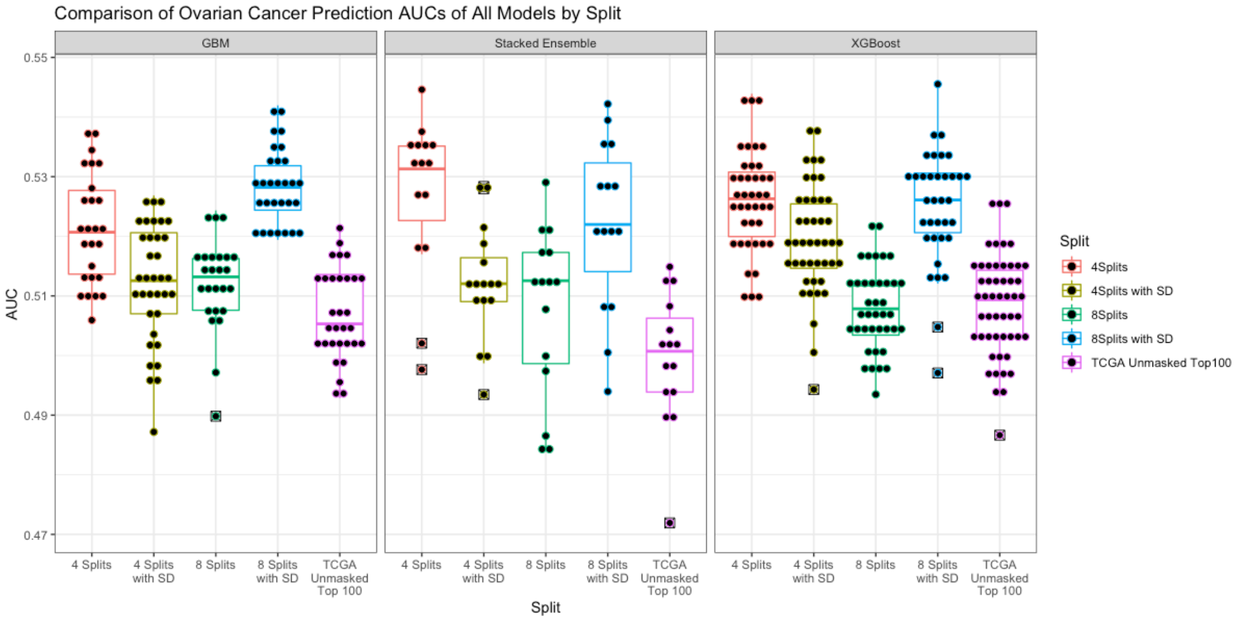

Figure 49: Comparison of Ovarian Cancer Prediction AUCs by Different Models and by Split
We tested whether the numbers of splits and features of each chromosome affect predictive performance and how CSLV sets compare to CNV set. We built five datasets by splitting each chromosome into four or eight segments and combining with the standard deviation, and mapping TCGA CNVs to UK Biobank l2r data. The dataset was used to predict whether an individual had ovarian cancer. The prediction was evaluated by the metric AUC. The plot presents the differences in predictive performance between models and chromosomal scale length variation combinations. The 4-splits dataset performed the best.

Figure 49 demonstrates how these models compare on the five different datasets. There is no significant difference between models. The increase in the number of splits and the addition of standard deviation impacted predictability. Figure 50 shows the differences in performance of all models for the five datasets. We tested whether the performance of the CSLV sets differs significantly from the TCGA-CNV set, and the p-values are recorded in Table 18. The differences in performance are not significant for all datasets, showing that the CSLV sets performed as well as the TCGA-CNV set.

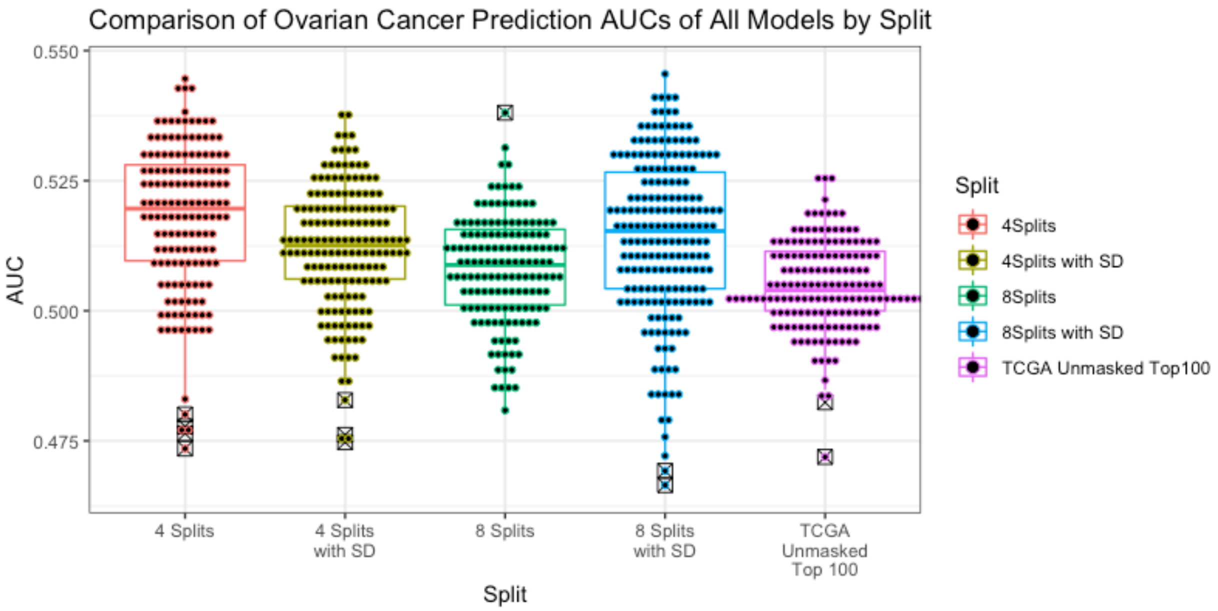

Figure 50: Comparison of Ovarian Cancer Prediction AUCs of All Models by Split
The average performance of all models for each combination of CSLV features.

Split Numbers	Mean AUC	Standard Deviation	95% Confidence Interval	p-value vs. TCGA-CNV
4 Splits	0.518	0.014	(0.515, 0.520)	9.399×10^{-22}
4 Splits with Standard Deviation	0.512	0.012	(0.510, 0.514)	1.671×10^{-13}
8 Splits	0.508	0.010	(0.506, 0.510)	6.551×10^{-5}
8 Splits with Standard Deviation	0.514	0.015	(0.512, 0.517)	8.061×10^{-16}
TCGA Unmasked Top 100	0.505	0.009	(0.504, 0.506)	

Table 18: Ovarian Cancer Prediction AUCs by Split
The mean, standard deviation, and p-values of the cross-validated AUCs of 4-splits, 4-splits-with-standard-deviation, 8-splits, 8-splits-with-standard-deviation, and TCGA-unmasked-top-100 datasets.

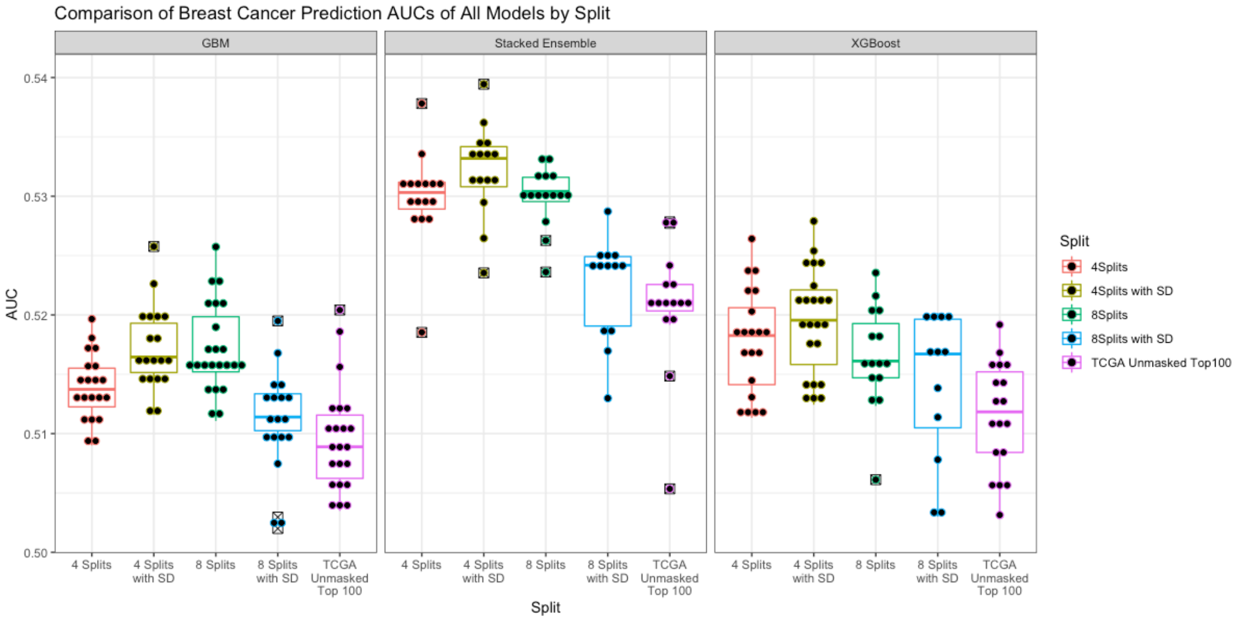

Figure 51: Comparison of Breast Cancer Prediction AUCs by Different Models and by Split
We tested whether the numbers of splits and features of each chromosome affect predictive performance and how CSLV sets compare to CNV set. We built five datasets by splitting each chromosome into four or eight segments and combining with the standard deviation, and mapping TCGA CNVs to UK Biobank l2r data. The dataset was used to predict whether an individual had breast cancer. The prediction was evaluated by the metric AUC. The plot presents the differences in predictive performance between models and chromosomal scale length variation combinations. The Stacked Ensemble model on the 4-splits-with-standard-deviation dataset performed the best.

Figure 51 demonstrates how these models compare on the five different datasets. The stacked ensemble achieved the best performance. The increase in the number of splits and the addition of standard deviation impact predictability. Figure 52 shows the differences in performance of all models for the five datasets. We tested whether the performance of the CSLV sets differs significantly from the TCGA-CNV set, and the p-values are recorded in Table 19. The differences in performance are significant for 4-splits-with-standard-deviation and 8-splits datasets, showing that half of the CSLV sets performed as well as the TCGA-CNV set.

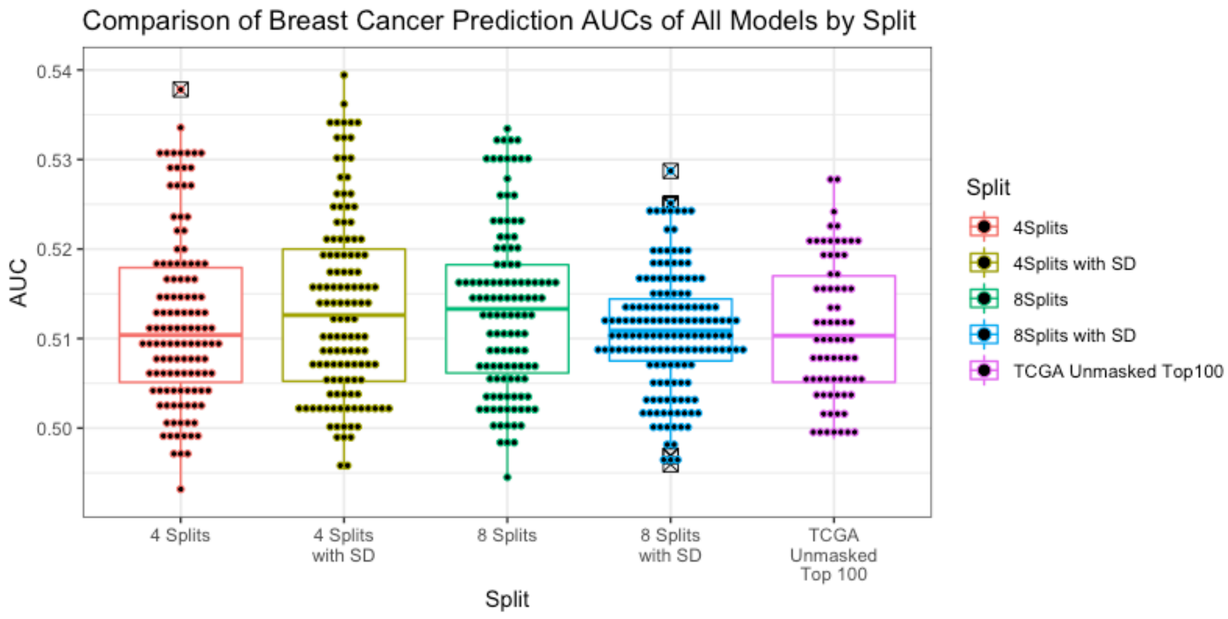

Figure 52: Comparison of Breast Cancer Prediction AUCs of All Models by Split
The average performance of all models for each combination of CSLV features.

Split Numbers	Mean AUC	Standard Deviation	95% Confidence Interval	p-value
4 Splits	0.512	0.010	(0.510, 0.514)	0.087
4 Splits with Standard Deviation	0.514	0.010	(0.512, 0.515)	0.003
8 Splits	0.513	0.009	(0.512, 0.515)	0.004
8 Splits with Standard Deviation	0.511	0.007	(0.510, 0.512)	0.580
TCGA Unmasked Top 100	0.511	0.008	(0.509, 0.513)	

Table 19: Breast Cancer Prediction AUCs by Split
The mean, standard deviation, and p-values of the cross-validated AUCs of 4-splits, 4-splits-with-standard-deviation, 8-splits, 8-splits-with-standard-deviation, TCGA-unmasked-top-100 datasets.

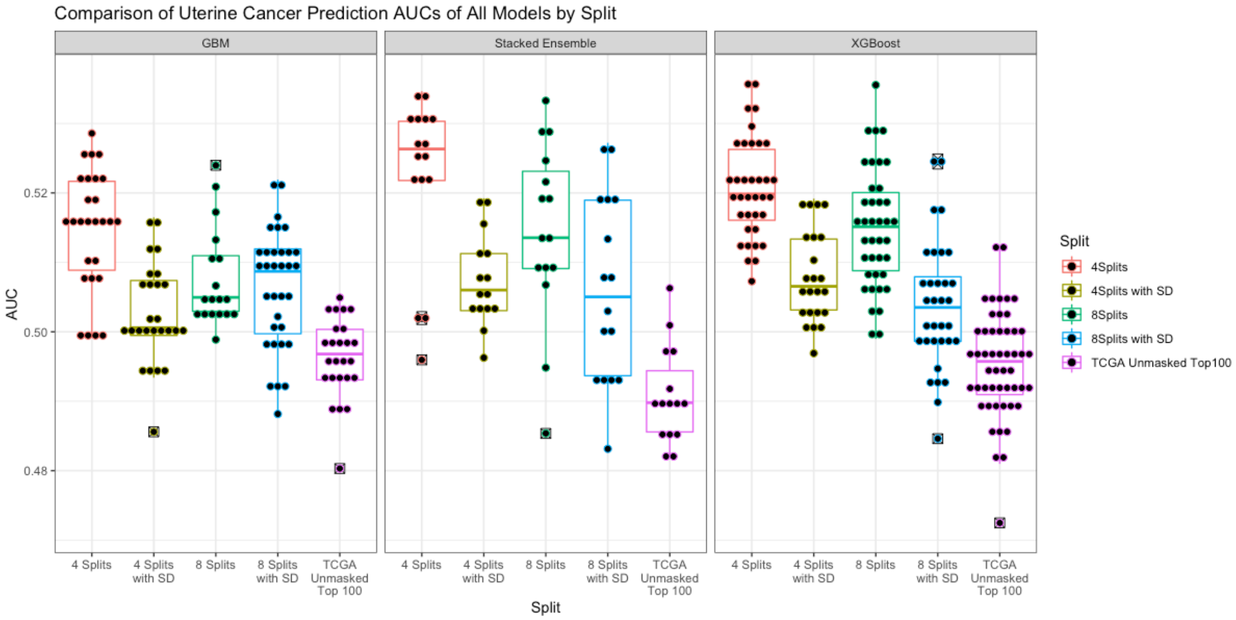

Figure 53: Comparison of Uterine Cancer Prediction AUCs by Different Models and by Split
We tested whether the numbers of splits and features of each chromosome affect predictive performance and how CSLV sets compare to CNV set. We built five datasets by splitting each chromosome into four or eight segments and combining with the standard deviation, and mapping TCGA CNVs to UK Biobank l2r data. The prediction was evaluated by the metric AUC. The plot presents the differences in predictive performance between models and chromosomal scale length variation combinations. The Stacked Ensemble model on the 4-splits-with-standard-deviation dataset performed the best.

Figure 53 demonstrates how these models compare on the five different datasets. There is no difference between model performance. The addition of standard deviation impacts predictability. Figure 54 shows the differences in performance of all models for the four datasets. We tested whether the performance of the CSLV sets differs significantly from the TCGA-CNV set, and the p-values are recorded in Table 20. The differences in performance were significant for all datasets except 8-splits-with-SD set, showing that most of the CSLV sets performed better than the TCGA-CNV set.

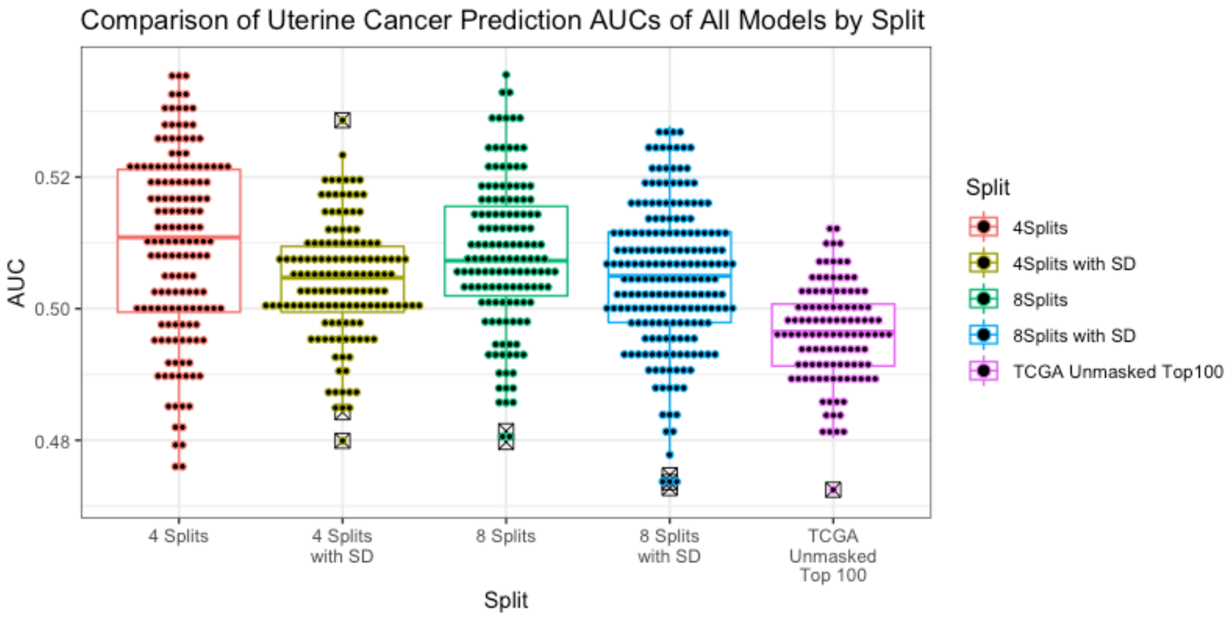

Figure 54: Comparison Uterine Cancer Prediction AUCs of All Models by Split
The average performance of all models for each combination of CSLV features.

Split Numbers	Mean AUC	Standard Deviation	95% Confidence Interval	p-value vs. TCGA-CNV
4 Splits	0.510	0.014	(0.507, 0.512)	1.325×10^{-23}
4 Splits with Standard Deviation	0.504	0.009	(0.503, 0.506)	2.555×10^{-20}
8 Splits	0.508	0.011	(0.506, 0.510)	1.246×10^{-25}
8 Splits with Standard Deviation	0.505	0.011	(0.503, 0.506)	2.423×10^{-22}
TCGA Unmasked Top 100	0.496	0.007	(0.495, 0.498)	

Table 20: Uterine Cancer Prediction AUCs by Split
The mean, standard deviation, and p-values of the cross-validated AUCs of 4-splits, 4-splits-with-standard-deviation, 8-splits, 8-splits-with-standard-deviation, and TCGA-unmasked-top-100 datasets.

There was no definitive consensus on which dataset and model achieved the best performance. Finer split does not significantly improve the predictability, though the addition of standard deviation impacts predictive performance more consistently, with stacked ensemble model on 4-splits-with-standard-deviation dataset most often outperforming other models. For most types of cancer, the CSLV sets performed better than the TCGA-CNV set.

We decided to further investigate a number of cancers with higher predictability more comprehensively to better understand the relative risk distribution and how the models arrive at their conclusions through H2O's explainability framework.

We first examined lung cancer. The predicted results were compared to the known lung-cancer status of the patients, who were first ranked by their scores, from the least likely to most likely to have lung cancer. As shown in Table 21, there was increasing risk of lung cancer by decile, and the highest decile has an approximately 2.6-fold of relative risk in comparison to the individuals in the lowest decile.

Decile	Number of Patients without Lung Cancer	Number of Patients with Lung Cancer	Total Number of Patients	Odds Ratio	95% Confidence Interval
1	472	352	824	0.74	0.64 – 0.86
2	462	362	824	0.78	0.68 – 0.90
3	439	385	824	0.87	0.76 – 1.00
4	456	367	823	0.80	0.69 – 0.93
5	413	410	823	0.99	0.86 – 1.14
6	425	398	823	0.93	0.81 – 1.08
7	424	399	823	0.94	0.81 – 1.08
8	375	448	823	1.19	1.03 – 1.37
9	360	463	823	1.28	1.11 – 1.48
10	281	542	823	1.92	1.65 – 2.23

Table 21: Odds Ratio of Lung Cancer Risk by Decile
The odds ratio between deciles of predicted results from the cross validation. The top 10% is 2.6 times as likely to be classified as lung cancer as the lowest decile.

A heatmap was created to present variable importance across all the generated models in Figure 55. Some of the most significant features were standard deviations and segments of chromosome 6, 7, 9, 13, X, and XY. This was then confirmed with a Shapley Additive explanation (SHAP) plot, as shown in Figure 56.

Second, we examined brain cancer. The odds ratio was calculated from the cross-validation predictions and recorded in Table 22. The risk of brain cancer increases by decile, and the top 10% has an approximately 2.4-fold of relative risk in comparison to the individuals in the lowest decile.

Decile	Number of Patients without Brain Cancer	Number of Patients with Brain Cancer	Total Number of Patients	Odds Ratio	95% Confidence Interval
1	127	78	205	0.61	0.46 – 0.82
2	114	91	205	0.80	0.60 – 1.06
3	110	95	205	0.86	0.65 – 1.15
4	104	101	205	0.97	0.73 – 1.29
5	104	100	204	0.96	0.72 – 1.28
6	106	98	204	0.92	0.69 – 1.23
7	92	112	204	1.22	0.91 – 1.62
8	96	108	204	1.13	0.84 – 1.50
9	87	117	204	1.34	1.01 – 1.80
10	82	122	204	1.49	1.11 – 1.99

Table 22: Odds Ratio of Brain Cancer Risk by Decile

The odds ratio between deciles of predicted results from the cross validation. The top 10% is 2.4 times as likely to be classified as brain cancer as the lowest decile.

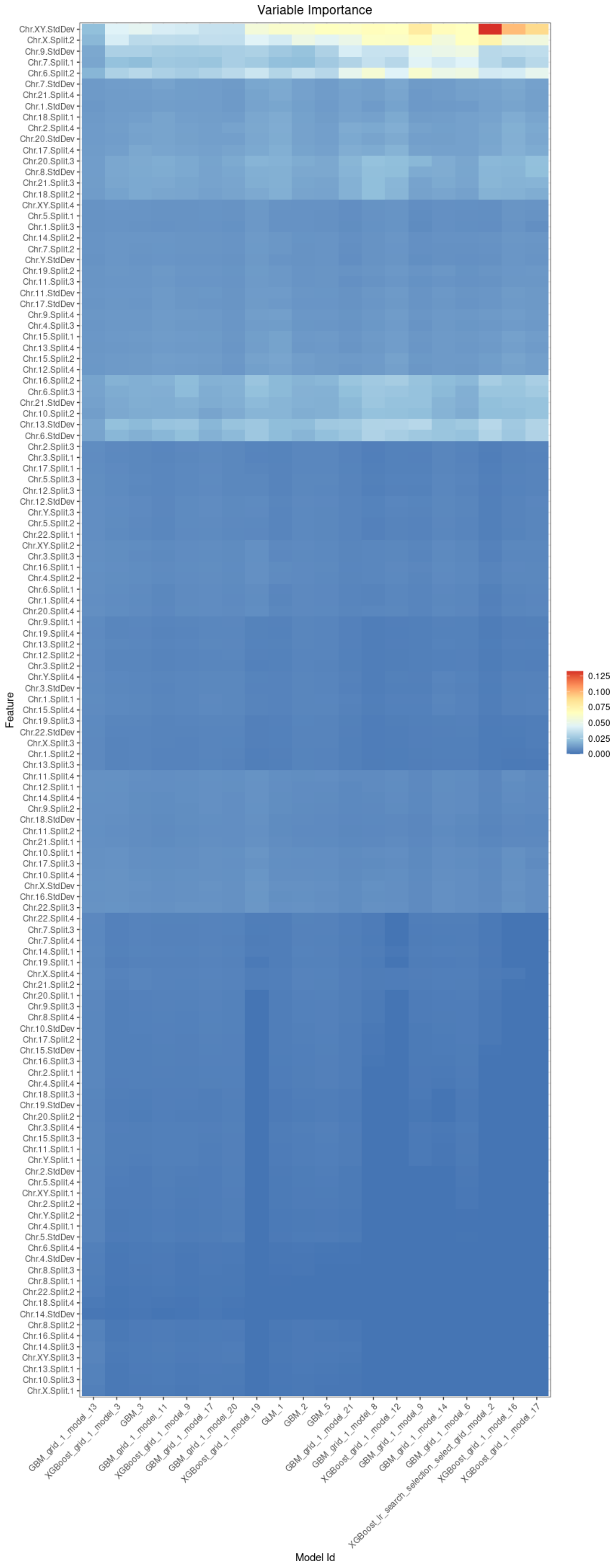

Figure 55: Variable Importance Heatmap of 4-Splits-with-SD Lung Cancer CSLV Models
The variables most influential to the predictive performance of the specified models. A value of 1.0 indicates the highest importance.

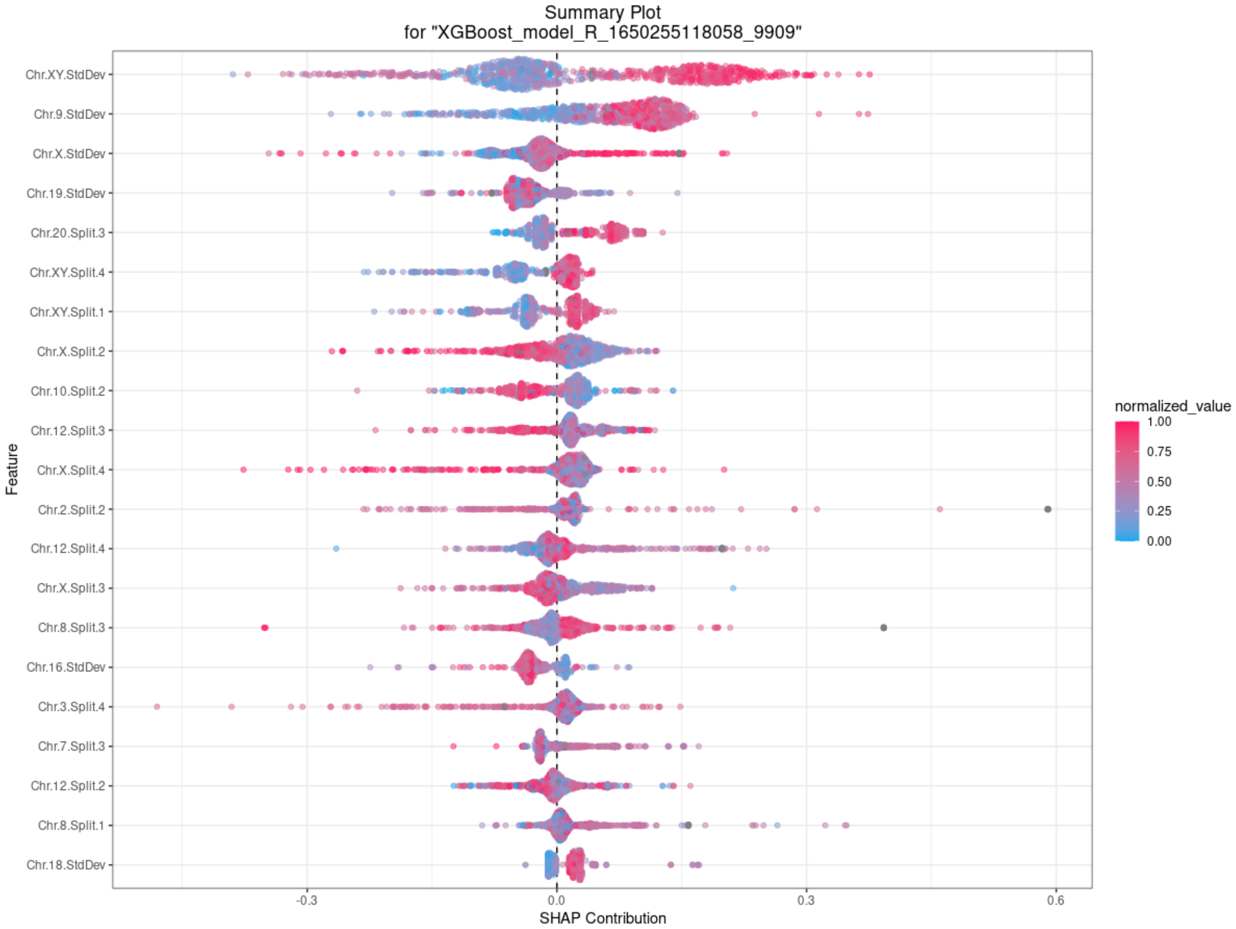

Figure 56: SHAP Plot of Top-Performing 4-Splits-with-SD Lung Cancer CSLV Model
The features with the highest contribution to the lung cancer prediction in the leading model. Some features identified in the plot originate from chromosome 6, 7, 9, 13, X, and XY,

A heatmap was created to present variable importance across all the generated models in Figure 57. Some of the most significant features were standard deviations and segments of chromosome 10, 12, 15, and X. This was then confirmed with a SHAP plot in Figure 58.

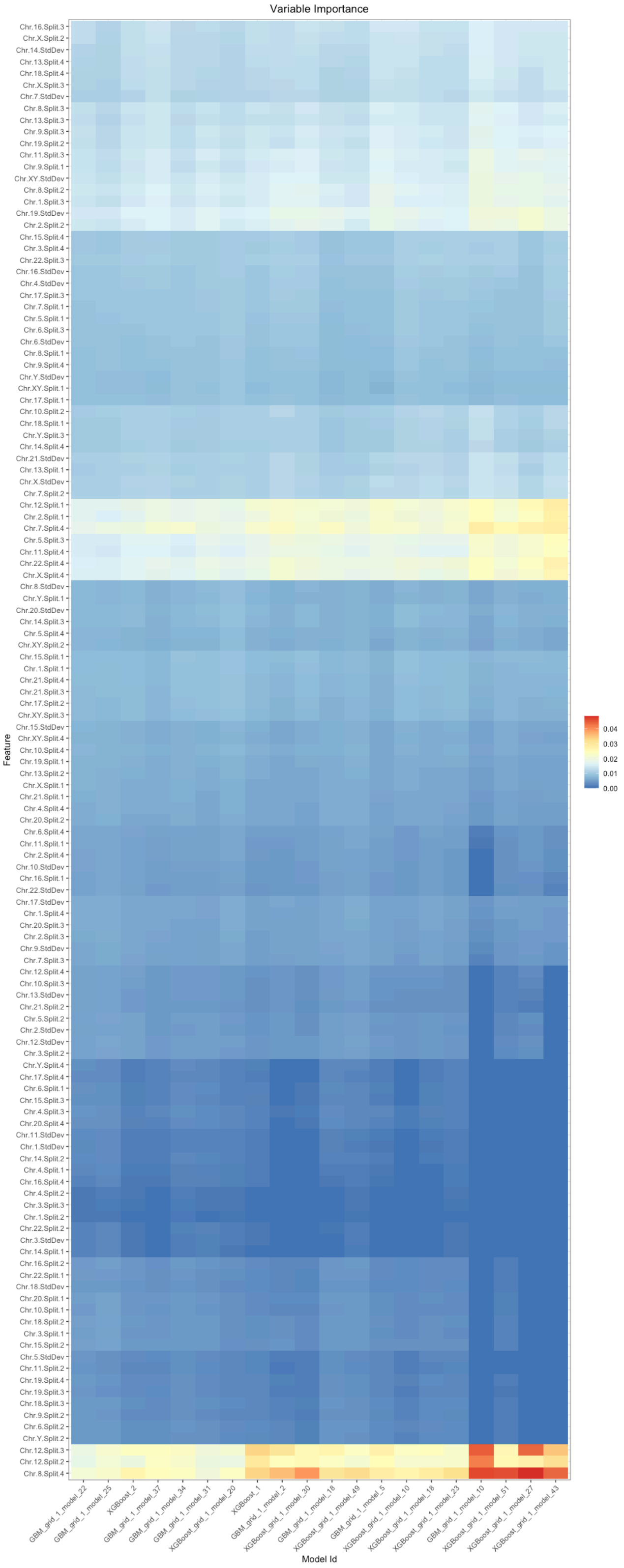

Figure 57: Variable Importance Heatmap of 4-Splits-with-SD Brain Cancer CSLV Models

The variables most influential to the predictive performance of the specified models. A value of 1.0 indicates the highest importance.

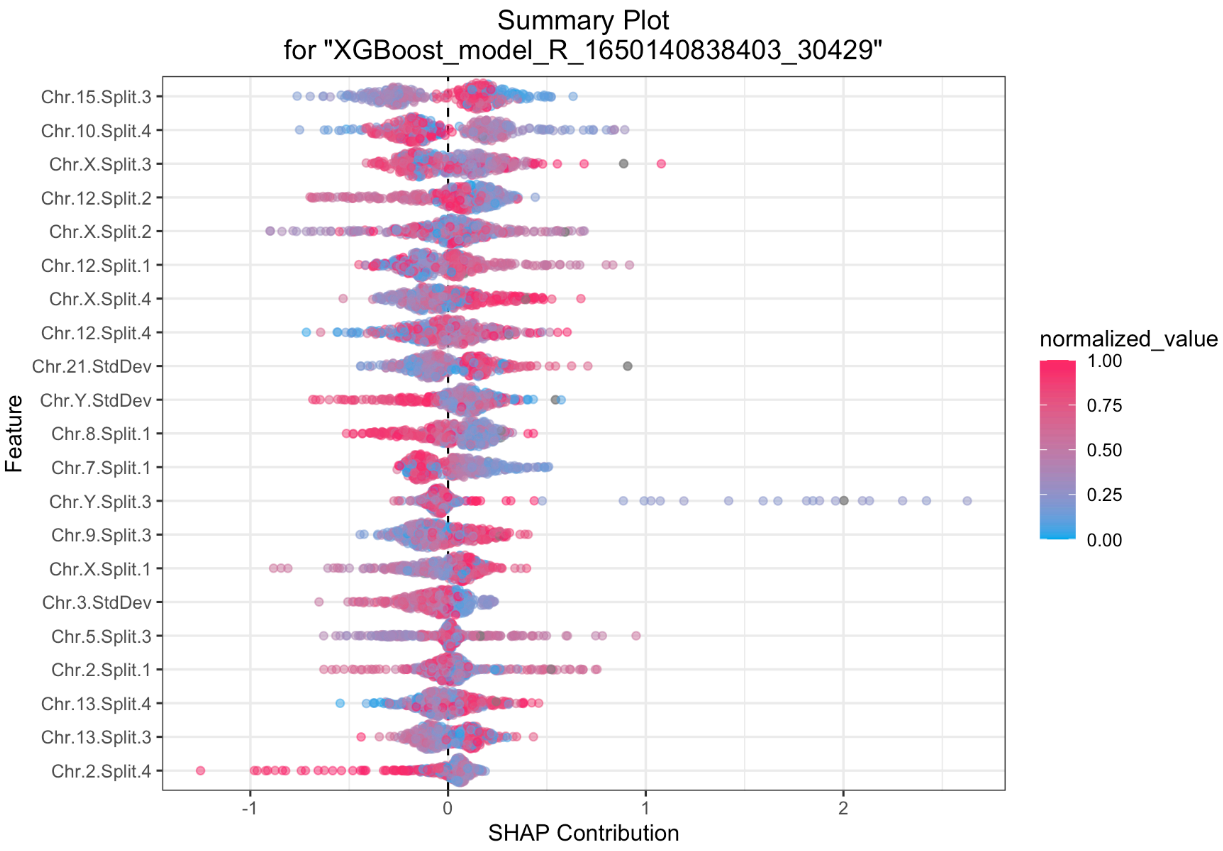

Figure 58: SHAP Plot of Top-Performing 4-Splits-with-SD Brain Cancer CSLV Model
The features with the highest contribution to the lung cancer prediction in the leading model. Some features identified in the plot originate from chromosome 10, 12, 15, and X.

We then examined colorectal cancer. The odds ratio was calculated from the cross-validation predictions and recorded in Table 23. The risk of colorectal cancer increases by decile, and the top 10% has an approximately 1.8-fold of relative risk in comparison to the individuals in the lowest decile.

Decile	Number of Patients without Brain Cancer	Number of Patients with Brain Cancer	Total Number of Patients	Odds Ratio	95% Confidence Interval
1	1009	722	1731	0.72	0.65 – 0.79
2	887	844	1731	0.95	0.86 – 1.05
3	892	838	1730	0.94	0.85 – 1.04
4	854	876	1730	1.03	0.93 – 1.13
5	873	857	1730	0.98	0.89 – 1.08
6	855	875	1730	1.02	0.93 – 1.13
7	862	868	1730	1.02	0.91 – 1.11
8	812	918	1730	1.13	1.02 – 1.25
9	841	889	1730	1.06	0.96 – 1.17
10	766	964	1730	1.26	1.14 – 1.39

Table 23: Odds Ratio of Colorectal Cancer Risk by Decile

The odds ratio between deciles of predicted results from the cross validation. The top 10% is 1.8 times as likely to be classified as colorectal cancer as the lowest decile.

A heatmap was created to present variable importance across all the generated models in Figure 59. Some of the most significant features were segments of chromosome 5, 19, and X. This was then confirmed with a SHAP plot in Figure 60.

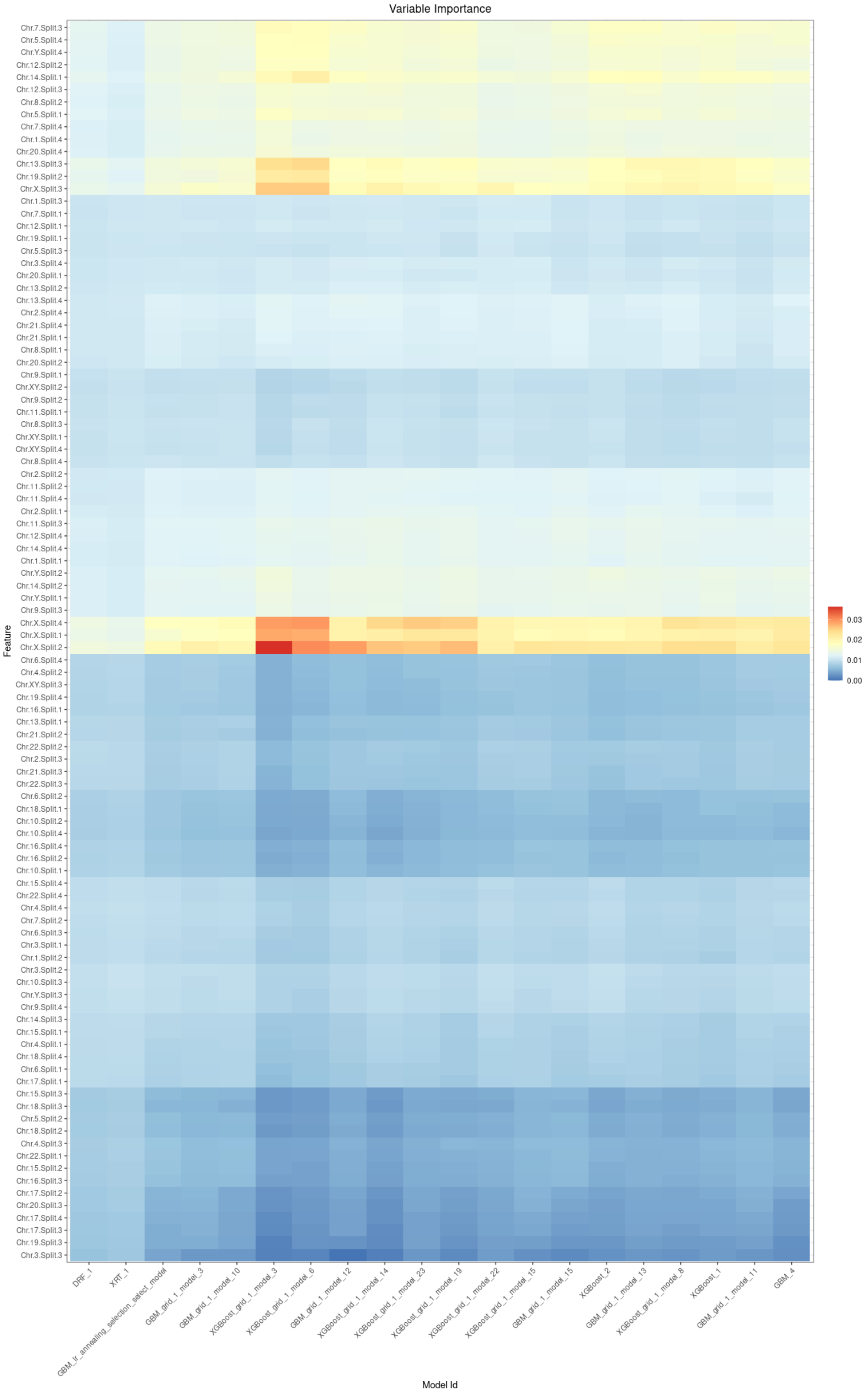

Figure 59: Variable Importance Heatmap of 4-Splits-with-SD Colorectal Cancer CSLV Models
The variables most influential to the predictive performance of the specified models. A value of 1.0 indicates the highest importance.

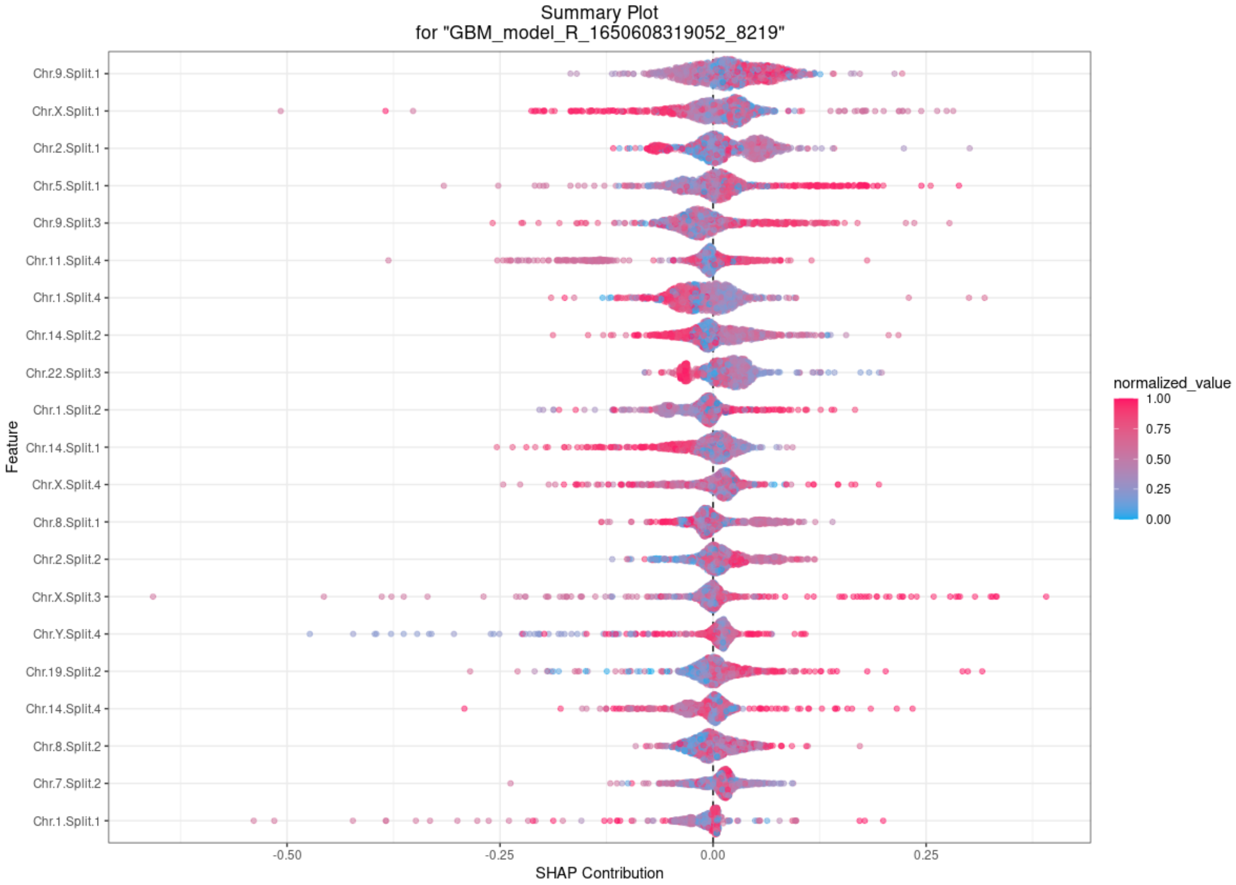

Figure 60: SHAP Plot of Top-Performing 4-Splits Colorectal Cancer CSLV Model
The features with the highest contribution to the lung cancer prediction in the leading model. Some features identified in the plot originate from chromosome 5, 19, and X.

These results demonstrate that germline genetic variations contribute to risk determination of various cancer types. Utilizing the structural difference across the genetic landscape alone provides sufficient information to predict whether an individual would have certain cancer better than random chance. The datasets were constructed with age- and gender-matched control to the diseased set to limit the variability derived from the phenotypic differences. Our analysis also revealed standard deviations of l2r values across chromosomes to be important factors in predicting cancer types. Currently, there are numerous studies dedicated to develop polygenic risk scores for many types of cancer[91], and our method may offer alternative means in such effort.

This study has several limitations. First, Chromosomal Scale Length Variation was developed for dimensionality reduction by averaging the l2r values across large segments of chromosomes, and the standard deviation feature was constructed to provide more information, i.e. the spread, of the dataset. The variable importance was explored with the SHAP analysis, but it does not provide an explanation for the underlying mechanism. Second, the UK Biobank suffers from a lack of diversity. Its population is primarily Caucasian in the United Kingdom. Lastly, it is uncertain that the non-cancer individuals in the database would remain so in the future, such that the "healthy" individuals might not actually belong in the control set after all.

We were able to build machine learning models based solely on germline chromosomal scale length variation for cancer risk determination, resulting in various levels of predictability. Lung cancer achieving the top performance, with an average AUC of 0.565. Although the AUC values obtained from these models would not be clinically useful at the current stage of development, they all significantly differ from chance and indicate the existence of a difference in structural genomics of the cancer patients in UK Biobank and the general UK Biobank population. The SHAP analysis performed on lung, brain cancers revealed potential regions and novel features, i.e. standard deviations of chromosomes, of importance.

Summary

The rapid advent in sequencing and computational technologies have advanced cancer research in recent years; however, much of the specific mechanism for cancer development remain uncertain. As cancer is a multifactorial disease, it results from interaction between genetic and environmental factors. It is important to first set apart hereditary genetic features from environment-induced mutations, an endeavor aided by the study of germline DNA, which solely comprise of inherited factors.

Currently, most genome-wide association study (GWAS) methods focus on somatic single nucleotide polymorphism (SNP) instead of interactions between more than two SNPs. Therefore, it is difficult to determine whether the SNPs are the sole cause or part of a group of genetic contributors. The exact role of SNPs in cancer development process often remains unclear. We utilized chromosomal scale length variants (CSLVs) of germline DNA to study the epistatic interactions between genes and the degrees hereditary factors contribute to specific cancers.

This study has shown that there is likely an epigenetic network effect of CNVs within an individual's genome, and such effect, once quantified, can be used to determine cancer risks. We developed Chromosomal Scale Length Variation to utilize germline CNVs in an efficient fashion while maintaining pertinent information. We have demonstrated that this germline genetic information can be used to distinguish between certain types of cancer and between healthy and cancer patients. This finding may serve as potential biomarkers for blood-based cancer diagnostics, and it provides another means to construct genetic risk scores for specific cancers.

The awareness of a hereditary predisposition to certain types of cancer is valuable, for screening increases the chance of early cancer detection, when it is more likely to be curable. For instance, the incidence rate of early-onset colorectal cancer is on the rise, and the cancer is often at the advanced stage of development when discovered, those with higher likelihood of colorectal cancer would benefit greatly with early screening. For other cancer types with environmental or lifestyle influences, e.g. lung cancer, those with high inherited cancer risks will be more likely to recognize and practice caution to avoid those harmful factors.

To evaluate the applicability of our findings, we compared the results to current published genetic risk scores, following the ICD 10 Cancer Code and validated in UK Biobank, for common cancers. The two metrics commonly used for reporting risk scores are the area-under-curve (AUC) value of the receiver operating characteristic curve and odds ratios between top and bottom percentage of the sample groups. Table 24 shows that the CSLV-based predictive models achieve comparable performance to previously reported risk scores for lung cancer and brain cancer. This establishes CSLV as a promising factor in studying the interaction between cancer germline genomics and inherited risk determination.

Cancer Type	**CSLV AUC**	**CSLV OR; Top 10%**	**Reported AUC**	**Reported OR; Top 10%**
Lung Cancer	0.598	2.594	0.51[91,92]	1.36[91,92]
Brain Cancer	0.567	2.443	0.546[91,93]	1.88[91,93]
Colorectal Cancer	0.565	1.759	0.545[91,92]	1.6[91,92]

Table 24: Comparison of Cancer Risk Scores

Comparison between CSLV-based risk scores and reported risk scores built and validated using UK Biobank GWAS and ICD 10 Cancer Code for Lung Cancer, Brain Cancer, and Colorectal Cancer.

We examined a large selection of the most common and top performing machine learning algorithms, such as tree-based methods like gradient boosting machines and XGBoost, but there are still many other techniques to be explored. For instance, more specific hyperparameter tuning may further improve our results.

CSLV also has some potential for improvement, concerning better feature selection, though CSLV-models performed better than models based on CNV segments called in TCGA for most cancer types. Currently, CSLVs are calculated from simple average of l2r values and standard deviation across an entire genome. This average may instead be constructed based on genomic location or the frequency of SNP probes, and the standard deviation may be computed from smaller chromosomal segments.

The Cancer Genome Atlas and the UK Biobank are incredibly valuable resources for our analysis; however, they suffer from a lack of diseased sample sets, with many cancer types numbering well below 5000 individuals and the control set considerably larger than the cancer data. The UK Biobank also lacks diversity, since its population is primarily Caucasian in the United Kingdom. This issue may be remedied in near future, as other countries begin to collect their own samples, and these large-scale databases will further grow in scope and size.

Our study demonstrates the promising results of applying machine learning methods on copy number variations, in the form of chromosomal scale length variations, in cancer diagnosis prediction and the potential of uncovering influential factors in the genomic landscape that contribute to hereditary cancer risk. This information may aid physicians in determining more personalized diagnostics and even employing relevant preventive measures based on individual susceptibility to specific cancer.

Reference

1. Turnbull C, Hodgson S. Genetic predisposition to cancer. *Clin Med J R Coll Physicians Lond*. 2005;5(5):491-498. doi:10.7861/clinmedicine.5-5-491

2. Eccles D, Tapper W. The influence of common polymorphisms on breast cancer. *Cancer Treat Res*. 2010;155:15-32. doi:10.1007/978-1-4419-6033-7_2

3. Sud A, Kinnersley B, Houlston RS. Genome-wide association studies of cancer: Current insights and future perspectives. *Nat Rev Cancer*. 2017;17(11):692-704. doi:10.1038/nrc.2017.82

4. Ongen H, Andersen CL, Bramsen JB, et al. Putative cis-regulatory drivers in colorectal cancer. *Nature*. 2014;512(1):87-90. doi:10.1038/nature13602

5. Glodzik D, Morganella S, Davies H, et al. A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. *Nat Genet*. 2017;49(3):341-348. doi:10.1038/ng.3771

6. Sadikovic B, Al-Romaih K, Squire J, Zielenska M. Cause and Consequences of Genetic and Epigenetic Alterations in Human Cancer. *Curr Genomics*. 2008;9(6):394-408. doi:10.2174/138920208785699580

7. Behjati S, Tarpey PS. What is next generation sequencing? *Arch Dis Child Educ Pract Ed*. 2013;98(6):236-238. doi:10.1136/archdischild-2013-304340

8. Yakhini Z, Jurisica I. Cancer computational biology. *BMC Bioinformatics*. 2011;12:120. doi:10.1186/1471-2105-12-120

9. Lau JW, Lehnert E, Sethi A, et al. The cancer genomics cloud: Collaborative, reproducible, and democratized - A new paradigm in large-scale computational research. *Cancer Res*. 2017;77(21):e3-e6. doi:10.1158/0008-5472.CAN-17-0387

10. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13:8-17. doi:10.1016/j.csbj.2014.11.005

11. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008;26(10):1135-1145. doi:10.1038/nbt1486

12. Thomas RK, Nickerson E, Simons JF, et al. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nat Med*. 2006;12(7):852-855. doi:10.1038/nm1437

13. Kamps R, Brandão RD, van den Bosch BJ, et al. Next-generation sequencing in oncology: Genetic diagnosis, risk prediction and cancer classification. *Int J Mol Sci*. 2017;18(2). doi:10.3390/ijms18020308

14. Freeman JL, Perry GH, Feuk L, et al. Copy number variation: New insights in genome diversity. *Genome Res*. 2006;16(8):949-961. doi:10.1101/gr.3677206

15. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet*. 2006;7(2):85-97. doi:10.1038/nrg1767

16. Kleinjan DA, Van Heyningen V. Long-range control of gene expression: Emerging mechanisms and disruption in disease. *Am J Hum Genet*. 2005;76(1):8-32. doi:10.1086/426833

17. Stranger BE, Forrest MS, Dunning M, et al. Relative impact of nucleotide and copy number variation on gene phenotypes. *Science*. 2007;315(5813):848-853. doi:10.1126/science.1136678

18. Buchanan JA, Scherer SW. Contemplating effects of genomic structural variation. *Genet Med*. 2008;10(9):639-647. doi:10.1097/GIM.0b013e318183f848

19. Kuiper RP, Ligtenberg MJL, Hoogerbrugge N, Geurts van Kessel A. Germline copy number variation and cancer risk. *Curr Opin Genet Dev*. 2010;20(3):282-289. doi:10.1016/j.gde.2010.03.005

20. Zack TI, Schumacher SE, Carter SL, et al. Pan-cancer patterns of somatic copy number alteration. *Nat Genet*. 2013;45(10):1134-1140. doi:10.1038/ng.2760

21. Krepischi ACV, Achatz MIW, Santos EMM, et al. Germline DNA copy number variation in familial and early-onset breast cancer. *Breast Cancer Res*. 2012;14(1):R24. doi:10.1186/bcr3109

22. Welcome to the Pan-Cancer Atlas. Accessed April 8, 2020. https://www.cell.com/pb-assets/consortium/pancanceratlas/pancani3/index.html

23. Hoadley KA, Yau C, Hinoue T, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*. 2018;173(2):291-304.e6. doi:10.1016/j.cell.2018.03.022

24. Malta TM, Sokolov A, Gentles AJ, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell*. 2018;173(2):338-354.e15. doi:10.1016/j.cell.2018.03.034

25. Huang K lin, Mashl RJ, Wu Y, et al. Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell*. 2018;173(2):355-370.e14. doi:10.1016/j.cell.2018.03.039

26. The Cancer Genome Atlas Program - National Cancer Institute. Accessed April 7, 2020. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

27. Genomics | Broad Institute. Accessed April 11, 2020. https://www.broadinstitute.org/genomics

28. Genome Characterization Pipeline - Center for Cancer Genomics - National Cancer Institute. Accessed April 11, 2020. https://www.cancer.gov/about-nci/organization/ccg/research/genomic-pipeline#collection-processing

29. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet*. 2008;40(10):1253-1260. doi:10.1038/ng.237

30. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004;5(4):557-572. doi:10.1093/biostatistics/kxh008

31. Bioinformatics Pipeline: Copy Number Variation Analysis - GDC Docs. Accessed April 11, 2020. https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/

32. About UK Biobank | UK Biobank. Accessed April 8, 2020. https://www.ukbiobank.ac.uk/about-biobank-uk/

33. Ollier W, Sprosen T, Peakman T. UK Biobank: From concept to reality. *Pharmacogenomics*. 2005;6(6):639-646. doi:10.2217/14622416.6.6.639

34. UKB : Search. Accessed April 8, 2020. http://biobank.ctsu.ox.ac.uk/crystal/search.cgi?wot=0&srch=hospitalization&sta0=on&sta1=on&sta2=on&sta3=on&str0=on&str3=on&fit0=on&fit10=on&fit20=on&fit30=on&fvt11=on&fvt21=on&fvt22=on&fvt31=on&fvt41=on&fvt51=on&fvt61=on&fvt101=on

35. *Genotyping and Quality Control of UK Biobank, a Large---Scale, Extensively Phenotyped Prospective Resource Information for Researchers*. Accessed April 14, 2020. http://biobank.ctsu.ox.ac.uk.

36. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z

37. UK Biobank — Oxford Big Data Institute. Accessed April 15, 2020. https://www.bdi.ox.ac.uk/research/uk-biobank

38. Shavlik JW, Dietterich TG Editors. Readings in machine learning. *Los Altos CA Morgan Kaufmann*. Published online 1990.

39. Kononenko I. Machine learning for medical diagnosis: History, state of the art and perspective. *Artif Intell Med*. 2001;23(1):89-109. doi:10.1016/S0933-3657(01)00077-X

40. Hunt EB, Marin J, Stone PJ. *Experiments in Induction.* Academic Press; 1966.

41. Nilsson N. Learning Machines: Foundations of Trainable Pattern-Classifying Systems. *McGraw-Hill N Y*. Published online 1965.

42. F. R. Principles of neurodynamics. *Wash DC Spartan Books*. Published online 1962.

43. Michie D. "Memo" Functions and Machine Learning. *Nature*. 1968;218(5136):19-22. doi:10.1038/218019a0

44. Love BC. Comparing supervised and unsupervised category learning. *Psychon Bull Rev*. 2002;9(4):829-835. doi:10.3758/BF03196342

45. Generalized Linear Model (GLM) — H2O 3.36.0.3 documentation. Accessed March 11, 2022. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html

46. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33(1):1. doi:10.18637/jss.v033.i01

47. Caruana R. *An Empirical Comparison of Supervised Learning Algorithms*. Accessed April 9, 2020. www.cs.cornell.edu

48. Caruana R, Karampatziakis N, Yessenalina A. *An Empirical Evaluation of Supervised Learning in High Dimensions*. Accessed April 9, 2020. http://yann.lecun.com/exdb/mnist/

49. Friedman JH. *Greedy Function Approximation: A Gradient Boosting Machine*. Vol 29.; 2001:1189-1232.

50. McCullagh P NJA. Generalized Linear Models, Second Edition (Monographs on Statistics and Applied Probability). *Lavoisierfr*. Published online 1989.

51. Breiman L. Random Forests. 2001;45:5-32.

52. Mason L, Bartlett P, Baxter J, Frean M. *Boosting Algorithms as Gradient Descent*.

53. Gradient Boosting Machine (GBM) — H2O 3.30.0.1 documentation. Accessed April 13, 2020. http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/gbm.html

54. Deep Learning (Neural Networks) — H2O 3.36.0.4 documentation. Accessed April 6, 2022. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/deep-learning.html

55. Stacked Ensembles — H2O 3.36.0.3 documentation. Accessed March 14, 2022. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/stacked-ensembles.html

56. Hanscombe K. *Ukbtools: Manipulate and Explore UK Biobank Data*.; 2019. Accessed April 6, 2022. https://CRAN.R-project.org/package=ukbtools

57. Tidyverse. Accessed April 2, 2022. https://www.tidyverse.org/

58. A Grammar of Data Manipulation • dplyr. Accessed April 2, 2022. https://dplyr.tidyverse.org/

59. Tidy Messy Data • tidyr. Accessed April 2, 2022. https://tidyr.tidyverse.org/

60. Create Elegant Data Visualisations Using the Grammar of Graphics • ggplot2. Accessed April 2, 2022. https://ggplot2.tidyverse.org/

61. package:ggthemes • All Your Figure Are Belong To Us. Accessed April 2, 2022. https://yutannihilation.github.io/allYourFigureAreBelongToUs/ggthemes/

62. Introduction to Boosted Trees — xgboost 1.1.0-SNAPSHOT documentation. Accessed May 22, 2020. https://xgboost.readthedocs.io/en/latest/tutorials/model.html

63. *LightGBM Release 2.3.2 Microsoft Corporation*.; 2020.

64. Parsons DW, Jones S, Zhang X, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science*. 2008;321(5897):1807-1812. doi:10.1126/science.1164382

65. Johnson DR, O'Neill BP. Glioblastoma survival in the United States before and during the temozolomide era. *J Neurooncol*. 2012;107(2):359-364. doi:10.1007/s11060-011-0749-4

66. Salcman M, Solomon L. Occurrence of glioblastoma multiforme in three generations of a cancer family. *Neurosurgery*. 1984;14(5):557-561. doi:10.1227/00006123-198405000-00006

67. Schwartzbaum JA, Fisher JL, Aldape KD, Wrensch M. Epidemiology and molecular pathology of glioma. *Nat Clin Pract Neurol*. 2006;2(9):494-503; quiz 1 p following 516. doi:10.1038/ncpneuro0289

68. Melin BS, Barnholtz-Sloan JS, Wrensch MR, et al. Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. *Nat Genet*. 2017;49(5):789-794. doi:10.1038/NG.3823

69. Janssens ACJW, Aulchenko YS, Elefante S, Borsboom GJJM, Steyerberg EW, van Duijn CM. Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med Off J Am Coll Med Genet*. 2006;8(7):395-400. doi:10.1097/01.gim.0000229689.18263.f4

70. Kinnersley B, Mitchell JS, Gousias K, et al. Quantifying the heritability of glioma using genome-wide complex trait analysis. *Sci Rep*. 2015;5:17267. doi:10.1038/srep17267

71. McLendon R, Friedman A, Bigner D, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nat 2008 4557216*. 2008;455(7216):1061-1068. doi:10.1038/nature07385

72. Brennan CW, Verhaak RGW, McKenna A, et al. The Somatic Genomic Landscape of Glioblastoma. *Cell*. 2013;155(2):462. doi:10.1016/J.CELL.2013.09.034

73. Jia D, Li S, Li D, Xue H, Yang D, Liu Y. Mining TCGA database for genes of prognostic value in glioblastoma microenvironment. *Aging*. 2018;10(4):592. doi:10.18632/AGING.101415

74. Kim YW, Koul D, Kim SH, et al. Identification of prognostic gene signatures of glioblastoma: a study based on TCGA data analysis. *Neuro-Oncol*. 2013;15(7):829-839. doi:10.1093/NEUONC/NOT024

75. Toh C, Brody JP. *Genetic Risk Score for Ovarian Cancer Based on Chromosomal-Scale Length Variation*. Genetic and Genomic Medicine; 2020. doi:10.1101/2020.07.18.20156976

76. Toh C, Brody JP. *Chromosomal Scale Length Variation of Germline DNA Can Predict Individual Cancer Risk*. Genomics; 2018. doi:10.1101/303339

77. Reid BM, Permuth JB, Ann Chen Y, et al. Genome-wide Analysis of Common Copy Number Variation and Epithelial Ovarian Cancer Risk. *Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol*. 2019;28(7):1117-1126. doi:10.1158/1055-9965.EPI-18-0833

78. Szulkin R, Whitington T, Eklund M, et al. Prediction of Individual Genetic Risk to Prostate Cancer Using a Polygenic Score. *The Prostate*. 2015;75(13):1467. doi:10.1002/PROS.23037

79. Ho WK, Tan MM, Mavaddat N, et al. European polygenic risk score for prediction of breast cancer shows similar performance in Asian women. *Nat Commun 2020 111*. 2020;11(1):1-11. doi:10.1038/s41467-020-17680-w

80. Hughes E, Tshiaba P, Gallagher S, et al. Development and Validation of a Clinical Polygenic Risk Score to Predict Breast Cancer Risk. *JCO Precis Oncol*. 2020;4(4):585-592. doi:10.1200/PO.19.00360

81. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet*. 2010;11(7):459-463. doi:10.1038/NRG2813

82. Xi Y, Xu P. Global colorectal cancer burden in 2020 and projections to 2040. *Transl Oncol*. 2021;14(10):101174. doi:10.1016/j.tranon.2021.101174

83. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424. doi:10.3322/caac.21492

84. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. *Gut*. 2017;66(4):683-691. doi:10.1136/GUTJNL-2015-310912

85. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015. *JAMA Oncol*. 2017;3(4):524-548. doi:10.1001/jamaoncol.2016.5688

86. Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB. Colorectal cancer. *Lancet Lond Engl*. 2019;394(10207):1467-1480. doi:10.1016/S0140-6736(19)32319-0

87. Siegel RL, Torre LA, Soerjomataram I, et al. Global patterns and trends in colorectal cancer incidence in young adults. *Gut*. 2019;68(12):2179-2185. doi:10.1136/GUTJNL-2019-319511

88. Lui RN, Tsoi KKF, Ho JMW, et al. Global Increasing Incidence of Young-Onset Colorectal Cancer Across 5 Continents: A Joinpoint Regression Analysis of 1,922,167 Cases. *Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol*. 2019;28(8):1275-1282. doi:10.1158/1055-9965.EPI-18-1111

89. Weigl K, Chang-Claude J, Knebel P, Hsu L, Hoffmeister M, Brenner H. Strongly enhanced colorectal cancer risk stratification by combining family history and genetic risk score. *Clin Epidemiol*. 2018;10:143. doi:10.2147/CLEP.S145636

90. Hsu L, Jeon J, Brenner H, et al. A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology*. 2015;148(7):1330-1339.e14. doi:10.1053/J.GASTRO.2015.02.010

91. Fritsche LG, Patil S, Beesley LJ, et al. Cancer PRSweb: An Online Repository with Polygenic Risk Scores for Major Cancer Traits and Their Evaluation in Two Independent Biobanks. *Am J Hum Genet*. 2020;107(5):815-836. doi:10.1016/j.ajhg.2020.08.025

92. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet*. 2018;50(9):1335-1341. doi:10.1038/s41588-018-0184-y

93. Xiao Y, Decker PA, Rice T, et al. SSBP2 variants are associated with survival in glioblastoma patients. *Clin Cancer Res Off J Am Assoc Cancer Res*. 2012;18(11):3154-3162. doi:10.1158/1078-0432.CCR-11-2778

Appendix

Study Abbreviation	Study Name
LAML	Acute Myeloid Leukemia
ACC	Adrenocortical carcinoma
BLCA	Bladder Urothelial Carcinoma
LGG	Brain Lower Grade Glioma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
LCML	Chronic Myelogenous Leukemia
COAD	Colon adenocarcinoma
CNTL	Controls
ESCA	Esophageal carcinoma
FPPP	FFPE Pilot Phase II
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
MESO	Mesothelioma
MISC	Miscellaneous
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular Germ Cell Tumors
THYM	Thymoma
THCA	Thyroid carcinoma
UCS	Uterine Carcinosarcoma
UCEC	Uterine Corpus Endometrial Carcinoma
UVM	Uveal Melanoma

Appendix: Table of TCGA study names and corresponding abbreviations