Interpretable Machine Learning for Biomarker Identification in RNA Seq Cancer Data

Jeremy Newton

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2025

Reading Committee:

Wooyoung Kim

Douglas Wacker

Dharma Dailey

Program Authorized to Offer Degree:
Computer Science & Software Engineering

University of Washington

Abstract

Interpretable Machine Learning for Biomarker Identification in RNA Seq Cancer Data

Jeremy Newton

Chair of the Supervisory Committee:
Wooyoung Kim
Department of Computing and Software Systems

Existing research on RNA Seq gene expression biomarkers has provided various methods to select a small list of genes as cancer biomarkers from a large number of gene expression data. Previous methods for identifying potential gene expression cancer biomarkers have focused on statistical analysis, but other methods have incorporated machine learning, often including Interpretable Machine Learning (iML) techniques. On 16 cancer types from TCGA data, we used inherently interpretable machine learning models: Logistic Regression, Random Forest, and Linear Support Vector Machine to narrow down subsets of potential genes as biomarkers using the trained models' feature importance rankings. We subsequently applied model-agnostic iML techniques, such as Shapley Additive Explanations (SHAP) and Permutation Importance, to narrow down the subsets even further. We compared classification performance between machine learning models trained on iML selected features with features selected by statistical methods, and biomarkers from external research. We found that iML biomarker selection methods lead to comparable or better classification performance on these datasets than the biomarkers from outside research, or from statistical analysis alone. Mutual Information estimation (MI) was a surprisingly useful technique for initial feature selection, and iML techniques improved the MI selected features for classification. We cross-checked potential biomarkers with biomedical annotations and gene pathway analysis, finding some support for the validity of the biomarkers.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY AND ABBREVIATIONS

AKAIKE INFORMATION CRITERION (AIC): Information theory based method for evaluating quality of statistical models.

BIOMARKER: "Biomarkers are cellular, biochemical, and molecular (proteomic, genetic, and epigenetic) alterations to recognize or monitor a normal, abnormal, or simply a biological process" [57]. Here biomarkers refer to genes of interest relating to their expression and cancer classification. In this paper "gene-set" is a collection of gene expression features as potential biomarkers.

CHI-SQUARED TEST: A statistical hypothesis test for testing whether two categorical variables are dependent or independent.

CLASSIFICATION PERFORMANCE: Performance of methods and models in this paper refers always to classification performance, not computational efficiency, which was not a main subject of study.

COHORT: Cohort in reference to TCGA data refers to the specific cancer type patients in the study were diagnosed with. The dataset of all available samples from this study cohort is also referred to as "cohort."

DECISION TREE: Decision Tree (Classifier) is a simple and inherently interpretable ML method. Consisting of one single tree, samples are split at each node attempting to best separate classes, based on a a feature and a threshold value for that feature. A Random Forest consists of multiple decision trees with specific properties.

DIFFERENTIAL GENE EXPRESSION (DGE): Differential Gene Expression analysis involves the difference of gene expression values across samples by category, with statistical tests for significance. For a single gene at a time, of which expression data is available, in the form of RNA-seq data or otherwise. Categories could be Tumor or Normal samples, or other categories according to the study.

EVALUATION MODEL: (EVM) an ML model used strictly for classification and not for iML in this experiment.

FEATURE: A dimension of input data for machine learning models.

FEATURE IMPORTANCE: (FI) Via multiple different methods, assigning a relative score of data features used by a machine learning model relative to performance.

FEATURE INTERACTION: How the relationship between the values of different features affects machine learning model predictions.

FOLD: see *k-fold*

FORMALIN-FIXED, PARAFFIN EMBEDDED (FFPE): A method for preserving tissue samples, as opposed to fresh frozen.

GENE ONTOLOGY (GO): "The mission of the GO Consortium is to develop a comprehensive, **computational model of biological systems**, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life." [8]

GENE EXPRESSION OMNIBUS (GEO): National Library of Medicine - National Center for Biotechnology Information (NCBI) "GEO is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community." [4]

GSL: gene-set length (GSL) refers to the number of gene expression features in a set of features.

HYPERPARAMETER: Hyperparameters are adjustable parameters in ML algorithms. Available hyperparameters depend on the architecture type, and implementation, and can have large impacts in model behavior. Hyperparameters are chosen and defined before training begins, in contrast to internal parameters which change automatically, or are "learned," during training.

HYPERPLASIA: Increased number of cells, when benign, distinct from cancer or tumor.

INHERENT FEATURE IMPORTANCE (IFI): Feature importance derived post-hoc from a model-specific inherently interpretable technique.

INHERENT FEATURE IMPORTANCE 1 (IFI_1): Refers to the Step 1 rank of all 20,530 genes used to select the top 250 genes by various methods. As a Method 2, IFI_1 refers to leaving the 250 features in this original order. This also includes statistical methods, which are not truly iML inherent feature importance.

INHERENT FEATURE IMPORTANCE 2 (IFI_2): Refers to Step 2, model specific inherent feature importance again when the same model was trained with the top 250 genes from IFI_1. Statistical lists of top 250 genes were used to train Logistic Regression 0 (LR_0) for this purpose.

INTERPRETABLE MACHINE LEARNING (IML): A field in machine learning that seeks to explain the predictions and inner workings of machine learning models [66].

K-FOLD: k-fold refers to data splits using a scikit-learn module. Data is split between training and testing data. K-folds (k=10 in this experiment) means the data is split into k sections, and the model is trained and evaluated k times, with an alternating test set each time, and the rest of the splits as training data. Stratified k-fold splits keeps the proportion of sample types the same in each split e.g. 10 percent Normal Tissue 90 percent Tumor.

K-NEAREST NEIGHBORS (KNN): KNN is a supervised machine learning algorithm. For classification, data-points are predicted to be the same class of the majority of k nearest neighbors in feature space. KNN was not an algorithm used in this experiment, but is used internally for scikit-learn's implementation of estimation of Mutual Information.

KYOTO ENCYCLOPEDIA OF GENES AND GENOMES (KEGG): An online Encyclopedia of gene annotations and information [52].

LEAVE ONE OUT CROSS VALIDATION (LOOCV): For the entire dataset, the ML model is trained on all samples except one, which is used as the sole test-set sample. This is repeated for every sample.

LEAVING-ONE-COVARIATE-IN (LOCI): A model-agnostic iML technique, a special case of Leaving-One-Covariate-Out (LOCO) which involves retraining ML models with features removed [50, 38].

LINEAR SUPPORT VECTOR CLASSIFIER (LSVC): See Support Vector Classifier (SVC).

LOG FOLD CHANGE (LFC): Log base 2 of mean group 1 / mean group 2, in the experiment LFC is calculated from gene expression counts between Primary Tumor (group 1) and Solid Tissue Normal (group 2) samples.

LOGISTIC REGRESSION: Here refers to Logistic Regression Classifier, scikit-learn implementation. A machine learning model calculating probability of target class.

LONG INTERGENIC NON-CODING RNA (LINC): LINCs do not code for proteins but can be measured in RNA-seq and some genes in the TCGA dataset are designated as LINC. LINCs have been studied in relation to cancer [76, 92].

MACHINE LEARNING: Class of algorithms or statistical methods that self-adjust parameters without explicit instructions, in this case for the task of classification.

MATHEWS CORRELATION COEFFICIENT (MCC): MCC is a classification metric, which can range from -1 to 1, with 1 being equivalent to accuracy of 1, or perfect classification with no errors.

METHOD 1, METHOD 2: Method 1 refers to the first feature ranking method applied to all features, and method 2 to subsequently re-ranking the top 250 features from method 1 or leaving them in their original order.

MODEL-AGNOSTIC: Interpretable techniques that can be applied to any kind of machine learning model.

MUTUAL INFORMATION (MI): Information and probability theory concept about the amount of information shared between two variables, or degree of randomness. In this paper, MI is the estimated MI score generated by scikit-learn's $mutual\_info\_classif$ algorithm [11, 30].

ONCOKB: A Memorial Sloan Kettering Cancer Center curated online database of cancer genes: known onco-genes, or cancer-causing or related genes [9]. FDA recognized. In this paper the prefix "onco-" means genes from this database, though the database also further labels some genes internally as onco-genes, or tumor suppressor genes.

PATHWAY ANALYSIS: A variety of methods for determining the biological context of a gene's function [40].

PERMUTATION IMPORTANCE: (PI) A model-agnostic method for calculating feature importance, by effectively removing one feature at a time by randomly shuffling the values, and tracking the changes in model performance.

PRIMARY TUMOR AND SOLID TISSUE NORMAL: Primary Tumor (PT) is a categorical label given to samples in TCGA data, referring to the origin of the gene expression sample. Solid Tissue Normal (STN) refers to samples taken from normal tissue, solid, not blood, from the same region of the body as the cancer type.

PROTEIN-CODING GENE: There are about 20,000 protein-coding genes in the human genome, which mean a section of DNA that is transcribed into mRNA which is then translated into a protein. Not all genes are protein coding, and not all DNA information in the gene is necessarily used for protein coding [8, 61].

RANDOM FOREST: A machine learning model based on an ensemble of decision trees, scikit-learn's Random Forest Classifier implementation was used for this experiment.

RANKS: Genes, as ML features, were ranked according to different systems. For example the ranking criteria could be Mutual Information Score, Point-Biserial Correlation, or ML model inherent feature importance. Any way of assigning a score to the genes or features, allows ordering by the score.

RNA: Ribonucleic acid, often single-stranded, information copied from DNA, often used to make proteins (mRNA is messenger RNA) [3, 15].

RNA-SEQ: RNA sequencing, RNA-seq data. RNA molecules originating from different genes are counted in a particular sample. This is a type of gene expression data, because the amount of RNA from a particular gene in a sample, shows to what degree that gene was expressed around the time the sample was taken [55].

SHAPLEY ADDITIVE EXPLANATIONS (SHAP): Model-agnostic iML technique. Algorithms implemented in this case by the python library shap for estimating an application of the Shapley value concept to machine learning models and features [12, 69].

SHAPLEY VALUE: A game theoretical concept developed by Lloyd Shapley for fairly assigning contribution in a cooperative setting [94, 80].

SUPPORT VECTOR CLASSIFIER (SVC): A Support Vector Machine (SVM) used for classification. Linear SVC (LSVC) uses a linear kernel, which does not remap features, and so preserves inherent interpretability, as opposed to other possible SVM kernels. SVMs work by separating labeled data positioned on a hyper-plane according to feature values, with a margin that maximizes the distance between the data points from different classes. The scikit-learn implementation was used in this experiment [5, 1].

TARGET: The target output feature for a machine learning model also known as the label. See Primary Tumor for labels in this experiment.

TCGA: The Cancer Genome Atlas, a multi-year intensive study collecting a variety of biological data from cancer patients, publicly available [16].

TRIAL: Specifically in this paper *trial* refers to the training/evaluation of ML models. Models were trained using *k-fold* splits. A separate trial is a new random k-fold split of the data, and the subsequent training/evaluating with each fold as the test set. With 2 trials, and 10 folds, each model was trained and evaluated 20 times in total for every training step in the experiment. Within each trial, the test-set is distinct in each fold.

WRAPPER METHOD: Wrapper methods are a feature selection method which use ML model performance to evaluate and guide feature selection. As opposed to filter methods, which use criteria ahead of time to select features for ML models.

XENA BROWSER: Provided by University of California Santa Cruz (UCSC) making TCGA RNA Seq data available in several preprocessed formats [14].

ACKNOWLEDGMENTS

I would like to acknowledge and thank my professors and mentors at UWB, as well as thank my friends and family who have supported and encouraged me throughout the process.

Chapter 1

INTRODUCTION

In this experiment, we evaluated the performance of Interpretable Machine Learning (iML) techniques for identifying gene expression biomarkers related to cancer. We compared potential biomarkers identified in this experiment with outside research biomarkers by using them as features for ML classification of Primary Tumor (PT) or Solid Tissue Normal (STN) in separate cancer datasets. We also compared biomarker identification methods by the rate at which previously established biomarkers were selected. The status of genes as known cancer biomarkers was based on cancer gene databases, and gene database annotations.

1.1 Motivation

Early detection of cancer is important for patient outcomes, as well as classifying the risk of recurrence, and predicting response to specific treatments. This has led the medical field to develop a variety of diagnostic tests for these purposes [29, 87]. Additionally, researchers are seeking to understand the mechanisms of different cancers and to develop new treatments [51]. Imaging, endoscopy, or physical examination have long been used for diagnosis and classification, along with increasingly sophisticated types of biomarkers. Biomarkers are metrics relating to biological processes, both normal, and abnormal, and can include cellular, molecular, genetic, epigenetic, chemical measurements, and more [57]. Besides detection and diagnosis, therapeutic drugs can be designed to target specific biomarkers, and overall understanding of the biological context of the disease can be enhanced when using biomarkers with bioinformatics methods [95]. There are then multiple uses for biomarkers: detection/classification, developing treatments, and research into the causes and mechanisms of the disease.

The best diagnostic biomarkers are accurate, minimally invasive, and cost-effective [37]. In this experiment, gene expression biomarkers were used for classification of patient samples as Primary Tumor (PT) or Solid Tissue Normal (STN). In other contexts gene expression biomarkers can be used for cancer stage classification and predicting clinical outcomes. Some methodologies for biomarker selection are applicable across contexts. Simple classification of Primary Tumor (PT) versus Normal Tissue Solid (STN) was chosen to make use of the most possible samples from the TCGA data source.

Gene expression is one type of biomarker, which can be ascertained from very small tissue or blood samples. Besides whole-genome sequencing, and diagnostic exome sequencing, gene expression based analysis is an established tool for determining if a sample is possibly cancerous [53]. Gene expression information can also help with studying the relationship of specific genes, proteins, or gene-pathways in cancer risk and development. RNA Sequencing is one way to measure gene expression, and has advantages both in cost and performance over DNA or protein based biomarkers. RNA sequence based gene expression tests have become more feasible and less expensive due to recent technological improvements [53].

RNA Sequencing is a method for measuring gene expression at a given point in time at the cellular level. There are 20,530 protein-coding genes and Long Intergenic Non-coding RNAs (LINCs) from the human genome in the TCGA datasets used in this experiment. RNA names match to both gene and protein names, as RNA is an intermediary molecule [20]. By counting specific RNA sequences in a sample, it is inferred that the specific gene matching the RNA was being expressed by definition. RNA biomarkers have already been studied for identifying potential therapeutic targets [79, 85]. Therapeutic targets can include specific proteins or gene expressions to be altered with a chemical/medical intervention [73].

1.2 Background

RNA Sequencing results in continuous counts for each gene measured, which become the dataset features, or the columns in a tabular dataset by convention. The number of feature columns in this experiment's data was 20,530, referring to 20,530 specific protein-coding genes

or Long intergenic non-coding RNAs (lincRNAs). The datasets were made up of samples from multiple patients, each sample was either from a tumor, or from solid normal tissue in the same region as the cancer type of the cohort. Each sample is a row by convention in a tabular dataset. The data is separated into datasets by cohort, which refers to the specific cancer type in question, e.g. *BRCA* is the abbreviation for Breast invasive carcinoma or "Breast Cancer", and *COAD* for Colon adenocarcinoma, "Colon Cancer." Each row has a unique sample ID, as well as a target label, which in this case is "Primary Tumor," or "Solid Tissue Normal." Using this data, the goal is to find a selection process for a small number of features out of the total 20,530 features, to be effectively used as biomarkers.

One common approach to finding biomarkers from gene expression data is to use statistical methods to determine Differential Gene Expression (DGE) for each gene in the dataset. The approach involves splitting the whole dataset into two groups: tumor samples and normal tissue samples. Then the values for each gene/feature, one at a time, are compared across the groups. Tests for statistical difference between the two distributions such as Student's t-test, or Welch's t-test, can be conducted. Then statistically different genes are selected based on the greatest average difference between them, or ratio such as Log Fold Change (LFC). This method is straightforward, but it does not capture interaction between genes as predictors [24, 71, 68]. For this reason, machine learning is a promising avenue because many models do account for feature interaction, and non-linear relationships between variables.

A classification machine learning model uses features, in this case gene expression levels through RNA Sequencing, to predict a target class, in this case clinical sample type: Primary Tumor or Solid Tissue Normal. The classifier combines information from some or all features (depending on the classifier) in any given sample during training or making a prediction (inference.) Often in machine learning models, the internal processes of the classifier are not well understood, and the model is considered a "black box." All that is known is the relationship between the model's predictions and the previously established labels for all samples in the dataset. Interpretable Machine Learning (iML) seeks to overcome this challenge by making the workings of the model more explainable.

One aspect of iML is establishing feature importance (FI), as well as trying to visualize or understand how a certain feature is used in making predictions [66]. For a given model, there are a variety of potential feature importance calculation methods. In this experiment, multiple iML techniques were used to create feature importance rankings. The next tasks in this experiment were to evaluate the different feature importance ranking methods, and also to synthesize results. Selected features were evaluated based on their usefulness in classification, using machine learning models training and predicting only with these few features rather than all of the originally available features. Feature overlap across different methods was also measured. Additionally, lists of gene expression features from other research papers, and lists of features chosen based on statistical methods in this experiment were also used for classification for comparison.

Whether the genes selected this way were actually the most biologically relevant in cancer development, or could possibly be therapeutic targets is still unknown. As with statistical Differential Gene Expression analysis, all that is known is a relationship or correlation between the value of the feature and the target variable in question. Additional analysis, such as gene pathway analysis, or online gene annotations, provided more information about selected potential biomarker genes [82]. Methods were also employed to attempt to measure the number of already known biomarkers selected by each method.

1.3 Experiment Goals

This experiment's goal is to identify potential biomarkers in cancer based on RNA-seq data using multiple iML approaches. The goal of using multiple approaches is to compare approaches for biomarker identification, including external methods published in peer-reviewed literature. Previous research in this area has often focused on statistical methods for finding potential biomarkers [71, 68, 99]. Other studies have incorporated complex computational algorithms and machine learning methods into the biomarker selection process [28, 93]. Or, some biomarkers are derived in part from scientific or medical literature [33]. The goal of this experiment is to compare these different approaches for identifying biomarkers. One question

this experiment attempted to address was: Since ML models can make classifications based on feature interactions, would iML methods be more successful in selecting biomarkers than statistical methods which do not account for feature interaction?

Chapter 2

RELATED WORK

Machine learning has been studied extensively for a variety of medical use cases including cancer research applications. Specifically, gene expression data has been studied for the purposes of classifying cancer tumors. There have been previous attempts to use the expression levels of a small subset of genes as biomarkers for classification or for further study.

2.1 Early Work in machine learning and cancer

In 1999, Golub et al. published their study where DNA microarray data, an earlier available form of measuring gene expression, was used with machine learning to classify patient samples as Acute lymphoblastic leukemia (ALL) or Acute myeloid leukemia (AML) [41]. This marked an early case of investigating gene expression data as a potential method for classification or diagnosis of cancer using machine learning.

Khan et al. used Artificial Neural Networks (ANNs) on DNA microarray data once again to classify small, round blue cell tumors (SRBCTs). This study also included an attempt to identify the top most important genes for further research [54]. Using machine learning models as methods for biomarker selection has been a well-studied [46, 86, 35, 18] subject in research, where the algorithms selected were suited to the computational and data resources available at the time and state of the art of the field [100, 75].

2.2 Current State of Research

With the advent of RNA-Seq technology, gene expression datasets are becoming more available due to the lower cost in data generation. Alharbi et al. cataloged 13 papers and 14 datasets relating to gene expression and cancer classification in a literature review. The

survey examines and compares multiple machine learning algorithms and feature selection methods, providing a baseline for current techniques. Oftentimes research is specific to a particular type of cancer, and with a target such as prognosis or survival analysis. A growing number of algorithms and methods are being developed for using gene expression data for machine learning cancer related use cases [19].

Das et al. published a review of existing biomarkers for cancer clinical use cases, including biomarker gene signatures based on gene expression [33]. This survey included *MammaPrint* (Used for risk of metastases in Breast Cancer,) *Prosigna (PAM50)* for risk of recurrence and predictive of response to hormone therapy or chemo-therapy in Breast Cancer, and *Oncotype DX*, which is used to predict recurrence of Breast cancer. Also included in the survey was *Decipher* a gene panel related to risk of recurrence of Prostate Cancer (PRAD) [47]. New methods for biomarker selection are being developed, and comparisons with known biomarkers is a strategy for validating the tested approach [32]. Given the many specific classification tasks related to cancer research, including individualized medicine, there will likely be no shortage of new data and use-cases for gene expression cancer biomarkers [48].

2.3 Biomarkers from Previous Studies

Peng et al. published a study on The Cancer Genome Atlas RNA Seq data from 12 different cohorts [71]. The researchers' methods produced a 14-gene signature, which was evaluated in classification performance on each cohort, separating tumor from normal tissue. The method began with separating all 20,500 genes in 3236 clusters based co-regulation in normal tissue samples, across all 12 cohorts. Among these gene clusters, a test for statistically significant difference between tumor and normal tissue expression was conducted. The test was done for each cohort separately, and then seven clusters were chosen that were significantly different in at least four of the cohorts. The researchers then conducted pathway enrichment analyses using Pathway Commons analysis [10], Gene Ontology (GO) analysis [8], and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis [52] which showed that all of these seven selected gene-clusters were closely related to cell-cycle regulation.

The final 14-gene signature was constructed by taking the top two most differentially expressed genes in each of the seven clusters. This 14-gene signature was then evaluated by using it as features for a Support Vector Machine (SVM) to classify normal tissue or tumor on the seven cohorts which had at least one significant gene cluster associated with it. Leave One Out Cross Validation (LOOCV) was used, and accuracy was the metric employed. As an example of a statistical and biomedical knowledge based method, these 14 genes were used to compare with potential biomarkers identified in this experiment using iML methods.

In another study, Nikitina et al. used RNA-Seq gene expression levels in formalin-fixed, paraffin embedded (FFPE) tissue samples from Russian patients to classify prostate cancer and benign prostatic hyperplasia with data not from The Cancer Genome Atlas [68]. The researchers also used statistical methods to find differentially expressed genes individually and also included non-protein coding genes in the analysis. The researchers then compared their analysis with differential gene expression in The Cancer Genome Atlas data. With the differentially expressed genes, an analysis of transcription factor recognition sites and microRNAs allowed them identify eight further differentially expressed genes. This method is also a combination of statistical and biological knowledge based methods, though with different data and approaches than Peng et al. The 12 total genes published by Nikitina et al. were another example of statistical analysis and were used to compare with iML methods as well.

Two other papers reviewed also published short lists of potential biomarkers based on statistical and biological analysis with gene expression cancer data. Wan et al. created a database and tool for differential analysis of genes in tumor or normal samples with a statistical Differential Gene Expression basis. Wan et al. used The Cancer Genome Atlas data as well as many other sources, and 64 cancer types giving the study a much wider range. The study was completely focused on differential expression as individual statistical tests per gene. The researchers provided a list of the top five over-expressed and top five under-expressed genes (in tumor samples vs. normal samples) based on the number of cancer types that the gene was significantly differentially expressed in over 50 percent of patients. They

also provided another list by the sorting method of if the gene was significantly differentially expressed in at least one patient in all of 23 selected cancer type cohorts [91]. Zheng et al. conducted a study based on Esophageal squamous cell carcinoma (ESCC). This study used a combination of Gene Expression Omnibus (GEO) and The Cancer Genome Atlas data, and began with statistical differential analysis per gene. The focus of this study was both on prognosis and diagnosis. Among differentially expressed genes, hub genes based on pathway and interactions were selected. Receiver Operating Characteristic (ROC) Curve Area Under Curve (AUC) per each hub gene individually for diagnosis, as well as expression levels in a 3rd data set were used to validate the results [99].

ML techniques have also been used: Wei et al. used an "RNA-Seq biomarker-generating algorithm" in the case of classifying cancer of origin of metastatic cancers. Their algorithm was step-wise Logistic Regression, where features are iteratively added or subtracted and the model is evaluated on Akaike Information Criterion (AIC) [93]. Kallah-Dagadu et al. published an experiment using The Cancer Genome Atlas' Breast Cancer (BRCA) cohort data. The researchers employed a variety of Interpretable Machine Learning (iML) techniques to better explain predictions. Their experiment also yielded a list of top most important features by the methods employed. The study's methodology included a wrapper method based on three different machine learning models, and concluded with SHAP and Leaving-One-Covariate-In (LOCI), and other iML techniques for analysis of the final sets of selected genes as predictors [50]. de la Guardia-Bolívar et al. arrived at a list of 27 cross-cancer biomarkers, 26 gene names of which were present in their paper. To arrive at this initial list of genes they used a statistical method, paired differential analysis, meaning the patient's normal tissue sample was compared to that same patient's tumor sample. The study used eight The Cancer Genome Atlas cohorts, and selected genes based on significant paired differential expression in at least 7/8 cohorts. The next step of the experiment involved training classifiers with the selected features and using iML techniques to rank feature importance. The features selected by de la Guardia-Bolívar et al. were evaluated by researchers for classifying normal or tumor samples both on the source cohorts used and unseen cohorts not

involved in compiling the biomarkers [34].

A study published by Coleto-Alcudia et al. provided another method for identification of biomarkers from RNA-seq datasets [28]. Rather than creating a fixed-length gene-set as potential biomarkers, this study used a multi-objective optimization algorithm to find the smallest number of genes while maximizing classification accuracy of tumor or normal tissue. The first step in this approach was a preliminary filtering of the candidate genes, due to the high computation cost of high-dimensional data. The first filter pass involved combining Information Gain (Mutual Information), Chi-Squared, a categorical statistical method, and gain ratio (a Decision-Tree related feature importance method.) With the selected genes, the Artificial Bee Colony based on Dominance (ABCD) algorithm was conducted. ABCD is a multi-objective optimization algorithm seeking to minimize the number of features and to maximize the classification accuracy of an R implementation of Support Vector Machine. Three cancer-type specific outputs were provided by the paper, and two outputs based on the combination of 6 types of cancer. These potential biomarkers were also then investigated by the researchers from a biological lens for relevance to confirm their results. The datasets in the researchers' study came from The Cancer Genome Atlas, Gene Expression Omnibus (GEO), and other sources. This unique algorithm, along with the other iML technique based methods from previous published works, is an opportunity to compare diverse methods for identifying biomarkers.

The full list of genes in each of these lists of biomarkers from external research papers, along with the cancer type cohorts they were intended for use with can be found in Table A.2. Coleto-Alcudia et al. ABCD 4 genes for BRCA [28], and "Decipher" for 25 genes for PRAD [33] each contained one gene that could not be found in The Cancer Genome Atlas data by any other alias. These two sets were modified to three and 24 gene-set lengths (GSL) and compared with experimental sets of the same length.

Among the selected reviewed papers with biomarkers, it is of interest whether the same genes are often selected by different methods. Among these 13 lists of biomarkers of varying lengths, from different research studies external to this experiment, and intended for different

cancer type cohorts, 15 genes overlapped or appeared in more than one list in Figure 2.1.

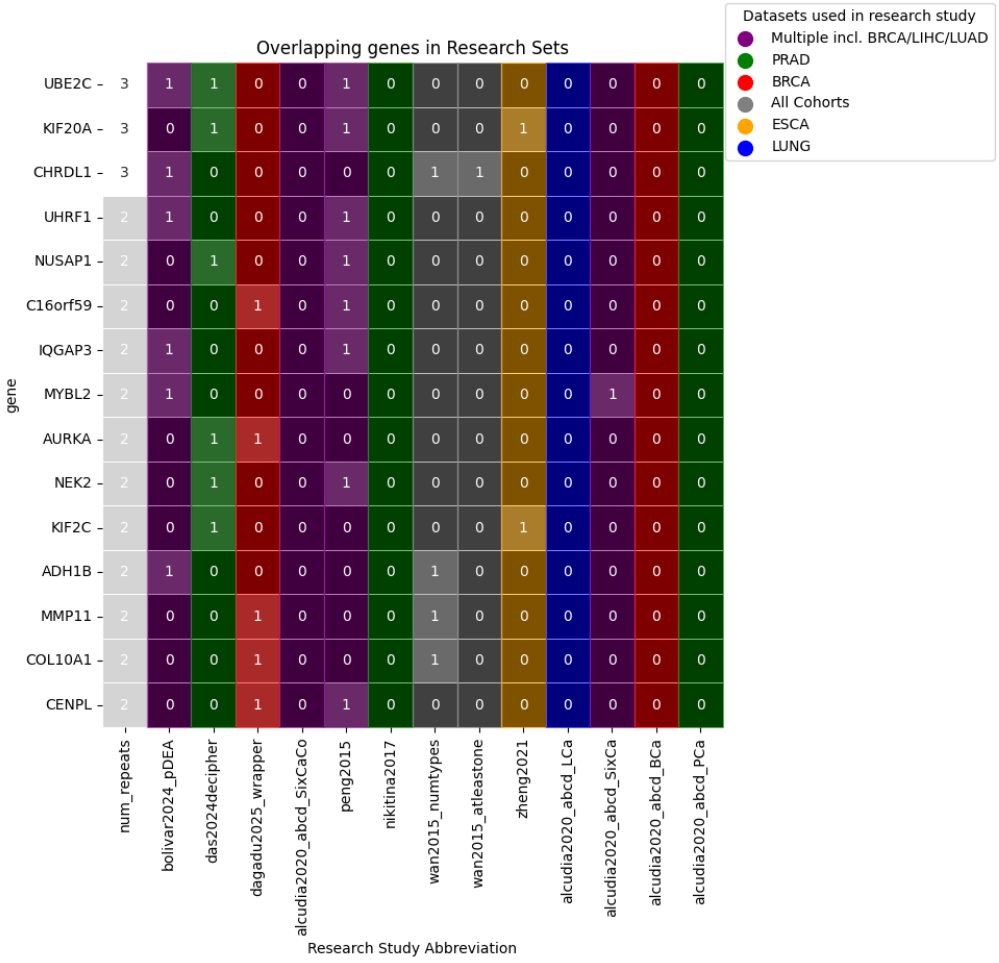

Figure 2.1: 15 genes appeared in more than one of these published biomarker lists, three genes appeared in three lists, and 12 in two lists. Research lists are color-coded by the cancer cohort type data they were intended for. Overlap between combination datasets or the same cohort was expected, but in some cases, as in *KIF20A*, a gene appears in sets of seemingly unrelated cohorts such as Prostate adenocarcinoma (PRAD) and Esophageal carcinoma (ESCA).

Chapter 3

METHODOLOGY

To compare methods for identifying biomarkers, the methodology for this experiment followed a multi-step process to create and evaluate feature rankings. Feature importance ranks were calculated using multiple methods in Step 1 using 20,530 genes, and in Step 2 using the top 250 genes from Step 1. In Step 3, to evaluate feature rankings, Machine Learning (ML) classifiers were trained and evaluated using smaller sets of the highest ranked features from Step 2 lists. In Step 4, biological analysis of the top repeating features across lists was conducted.

3.1 Overview

The data used in this experiment was originally gathered by The Cancer Genome Atlas [13], and was accessed through UCSC Xena Browser [14]. After data cleaning, Step 1 classification used all 20,530 genes to train simple inherently interpretable ML models and rank feature importance post-hoc (IFI_1). Mutual Information and statistical methods were also used to score and rank features. Then lists were created using subsets of the top 250 genes from each ranking method. Models trained in Step 2 generated more feature importance lists using SHapley Additive exPlanations (SHAP), Permutation Importance (PI), and post-hoc model-specific inherent feature importance again (IFI_2). OncoKB [9, 21, 83] only versions of generated subsets by feature importance were also included for comparison. In Step 3, the top features by importance of all methods employed so far were evaluated by training evaluation models (EVM) on between 26 and 2 top ranked genes. Gene-sets from external research were also included at this step as feature lists for evaluation and comparison. In Step 4, analysis of best performing sets and overlap of gene-sets was used to investigate

genes further for annotations related to cancer, and for pathway analysis. The methodology overview in steps is shown in Figure 3.1.

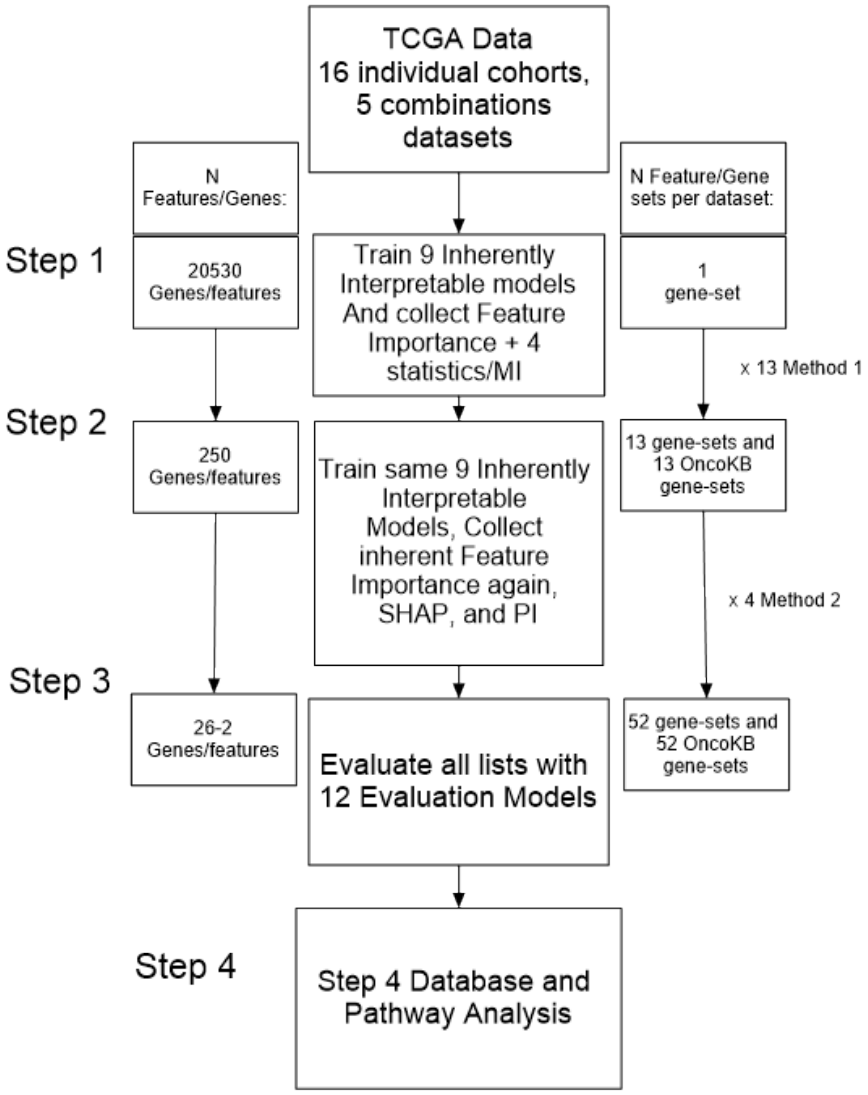

Figure 3.1: A simplified flowchart of the experiment methodology shows feature importance gathered in Steps 1 and 2, evaluation in Step 3, and in Step 4, biomarker analysis.

3.2 Data Source

Data was originally gathered and made available by the invaluable The Cancer Genome Atlas (TCGA) project. TCGA was a multi-year project by the National Cancer Institute (NCI) and the National Human Genome Research Institute. This publicly available data was drawn from more than 20,000 samples from patients with 33 different kinds of cancer [16]. For the purposes of this experiment, only RNA-Seq data of 20,530 protein-coding genes and LINCS is used, though there are far more types of data available from TCGA. The datasets for each cohort consist of RNA sequence reads of samples from "Primary Tumor" or "Solid Tissue Normal." The gene names are the same in each cohort dataset. A small number of samples with other labels such as Metastatic Tumor or Recurrent Tumor were discarded.

Secondly, XENA browser, provided by the University of California Santa Cruz provides the original TCGA data in a much more user-friendly format. The data is available by cohort which indicates the primary disease (cancer type) of the samples, and the corresponding normal tissue samples. Directly downloaded from Xena Browser, the data had already been log_2 transformed, and normalized. There are multiple normalization options available from Xena Browser; the two types used in this methodology were within-cohort normalization, and "PANCAN" normalization, or normalization across all cohorts. For each individual dataset, the within-cohort normalized data was used, and for combined-cohort datasets, the PANCAN normalized data was used. The data from USC Xena Browser still required some further minor data wrangling and data cleaning steps.

3.3 Exploratory Data Analysis (EDA)

After the data had been transformed there were 20,530 gene expression columns, (feature columns.) In all datasets the number of tumor samples exceeds the number of normal tissue samples, often times by a great deal. Some datasets included no normal samples at all, which were not used in this experiment. Out of all 33 downloadable cohorts, 16 cohorts were selected for suitability in the experiment. A cut-off of at least 10 normal tissue samples was

set so that a stratified k-fold split with *k=10* would contain at least one unique normal tissue sample per split. This way, possibly skewed results from cohorts with very little samples of one class would not affect the average of feature ranking methods across cohorts. The selected cohorts with number of samples and class composition are shown in Figure 3.2.

Five combinations of cohorts of cancer types were used to match the data that some of the gene-sets from external research were generated from. Pie charts in Figure 3.2 display the data composition of 16 selected cohorts and five combinations: LUNG is the combination of LUAD, and LUSC; PANCAN_selected is all 16 individual cohorts; [28] is BRCA, LIHC, ESCA, TGCT, THCA, LUAD, LUSC; [71] is BLCA, BRCA, COAD, HNSC, LIHC, LUAD, LUSC; [34] is COAD, BRCA, LUAD, KIRC, STAD, LIHC, THCA, UCEC.

Combined datasets were expected to be more difficult than individual cohort datasets, as the task involves separating normal tissue samples from primary tumor samples from all different regions of the body. The fact that there are limited numbers of normal tissue samples makes the classification task more difficult for all datasets. The imbalanced classes can also result in misleading performance evaluation, such as when using accuracy as a metric alone.

Since the data are high-dimensional with 20,530 feature columns, data visualization is challenging. One technique is the use of dimensionality reduction, three techniques of which were employed here: Principal Component Analysis (PCA), T-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP). Each of these techniques was applied to each dataset in two and three dimensions. Many datasets appear easily separable between primary tumor and normal tissue samples. Breast Adenocarcinoma (BRCA) has the most samples of any cohort and is fairly easily separable in lower dimensions. BRCA, Prostate adenocarcinoma (PRAD), and Stomach adenomcarcinoma (STAD) in PCA-2D are displayed in Figure 3.3.

One interesting observation from the dimensionality reduction analysis of the Breast invasive carcinoma (BRCA) data is that the tumor class itself seems to be split into two distinct clusters. This reflects the variability of possible cancer datasets or cancer types.

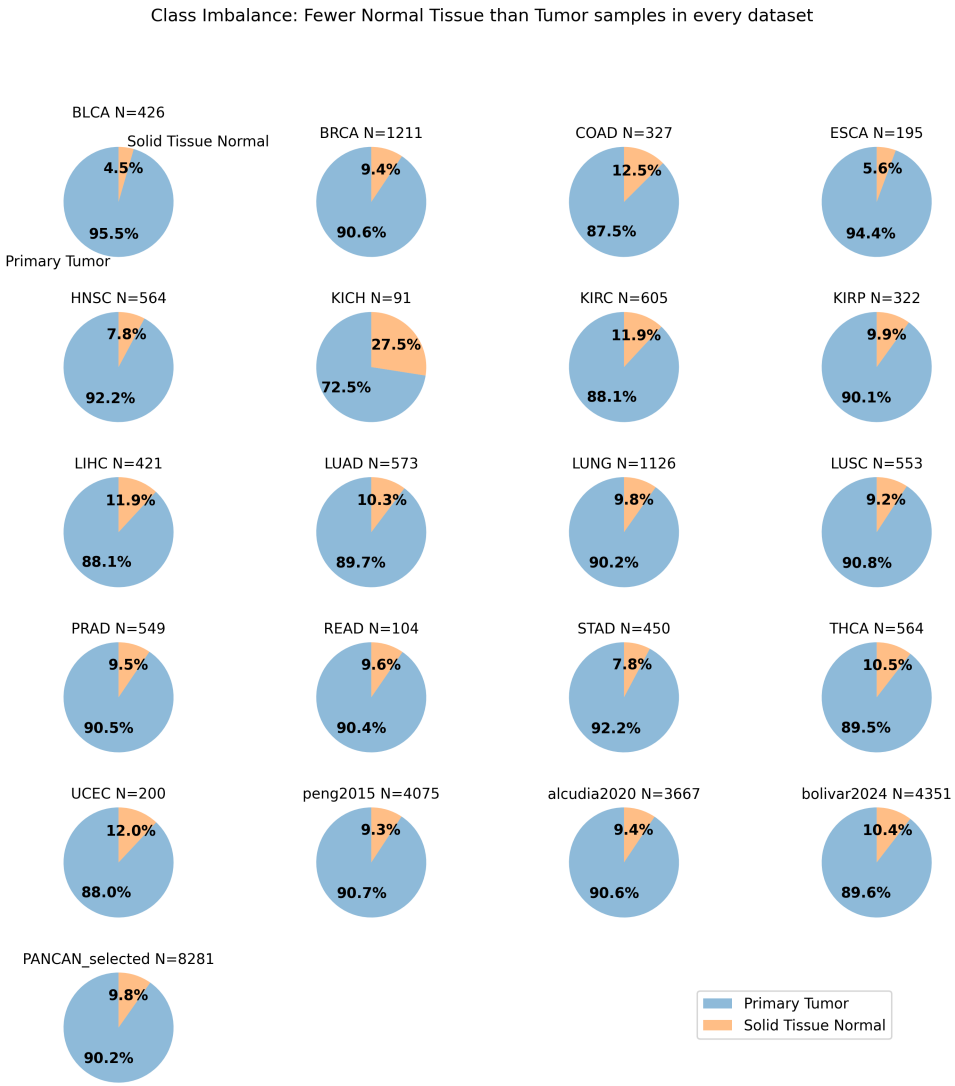

Figure 3.2: Each pie chart shows the percentage of Primary Tumor (PT) samples in blue, and Normal Tissue Solid (STN) samples in orange. The total number of samples along with the cohort abbreviation is above each chart with N= the total number of samples. All datasets are imbalanced, with more Primary Tumor samples than Solid Tissue Normal samples.

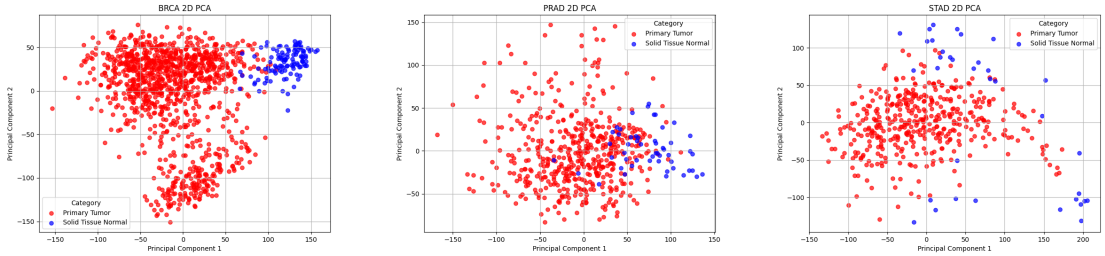

Figure 3.3: PCA 2-Dimensions (2D), Primary Tumor (PT) samples in red and Solid Tissue Normal (STN) samples in blue. Left: BRCA dataset PCA in 2D visualization, showing fairly clear separation between target classes, and two distinct centers for Primary Tumor. Center: PRAD dataset, showing that the normal tissue and primary tumor samples are not as easily separated in this dimensionality reduction view. Right: STAD dataset showing a more difficult separation, and unusually different cluster centers for the normal tissue samples.

There are also examples of other cohorts where the separation of classes is not as easy in lower dimensions. Stomach Adenocarcinoma (STAD) is unusual in having separate clusters for normal tissue samples as well. Prostate Adenocarcinoma (PRAD) seems to have less separable clusters, indicating a possibly challenging classification task. The fact that different cancer types have different data means some algorithms or methods may work well for some cohorts and not others.

The PANCAN combination dataset has multiple challenging clusters as was expected, since it combines multiple different normal tissue and tumor sources. PANCAN_selected is displayed in PCA-2D and t-SNE-2D in Figure 3.4.

3.4 Statistics

T-Test, Log Fold Change (LFC), Point-Biserial Correlation (PBC), and Mutual Information (MI) were calculated for all gene expression values in each dataset. Welch's T-Test was used which does not assume equal variance between the samples, and is better suited to unequal sample sizes. The SciPy implementation of Welch's T-test was used in this experiment.

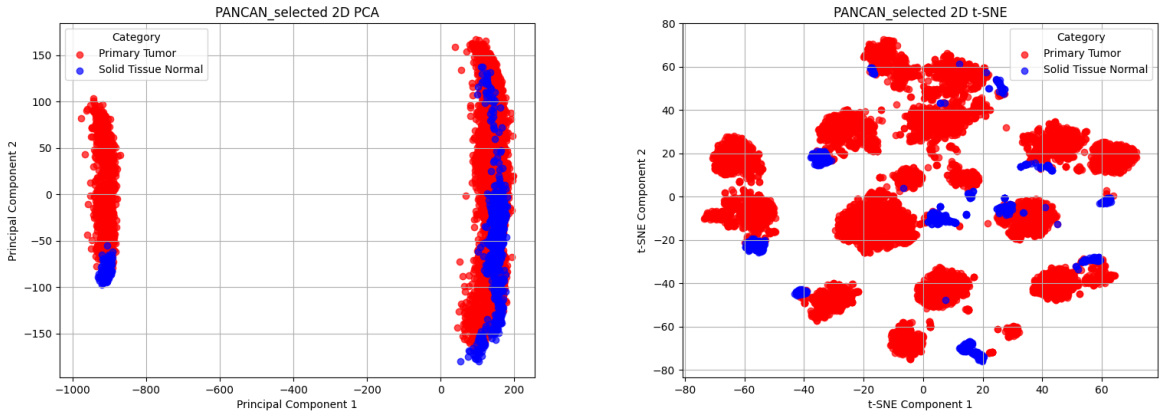

Figure 3.4: Left: 2D PCA of PANCAN_selected, a combination of all 16 individual cohorts (PANCAN normalized.) There are two distinct clusters, and Primary Tumor / Solid Tissue Normal samples are not easily separable. Right: 2D t-SNE visualization of PANCAN_selected, providing a different view showing multiple clusters of both sample types.

Welch's T-test still requires the assumption of normal distributions of each group, which was assessed with the Shapiro-Wilk test, of which the SciPy implementation was used. If the number of samples in each group was higher than 50, the test for normality was bypassed on the basis of assumptions from the central limit theorem. The p-values generated by Welch's T-Test were corrected for multiple trials using the Benjamini-Hochberg method implemented by the statsmodel Python package [72].

Log Fold Change (LFC) is defined as the mean of the tumor sample values divided by the mean of normal tissue samples. The log base 2 of the quotient is then taken. An algorithm, known as DESeq is commonly used to arrive at a better conclusions of the differential expression statistically of a gene, accounting for the dispersion of raw RNA sequence reads [62, 89]. The DESeq algorithm, an R package, was not used here in its full capacity, but was used by Peng et al. [71].

For each gene, Point-Biserial Correlation (PBC) was also calculated with the dichotomous variable as the labels of "Primary Tumor" or "Solid Tissue Normal" and the continuous

variable the gene's expression values. SciPy's implementation of Point-Biserial Correlation was again used [81]. Because PBC requires an assumption of equal variances, Levene's test, implemented by SciPy, was used. An example histogram of gene expression statistics in the BRCA cohort is displayed in Figure 3.6.

3.4.1 Volcano Plots for LFC EDA

A volcano plot displays Log Fold Change (LFC) information along with p-value information which is -log_{10} transformed for visualization, here BRCA and COAD volcano plots are provided for example and contrast in Figure 3.5.

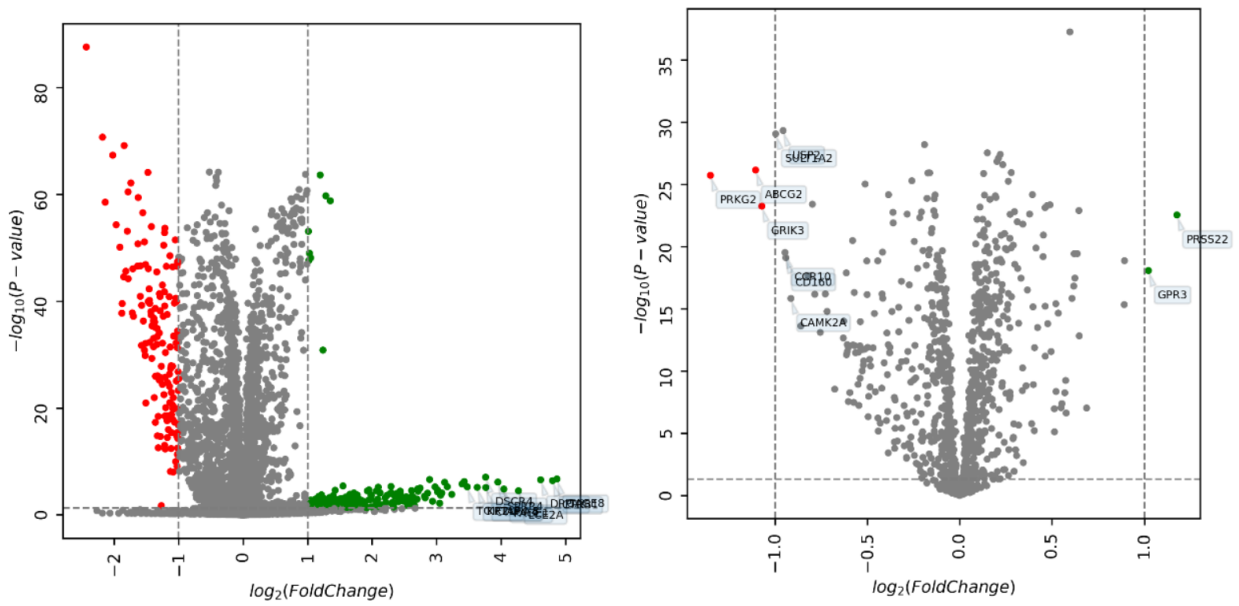

Figure 3.5: Left: BRCA Volcano Plot, with Log Fold Change threshold of –1, 1 and transformed p-values showing up and down regular significantly differently expressed genes in the dataset. Right: COAD Volcano Plot. There are fewer genes that pass both thresholds in the Colon adenocarcinoma (COAD) dataset than in Breast invasive carcinoma (BRCA.) COAD was one of the more difficult datasets to classify.

In the Volcano plots, adjusted p-values by the Benjamini-Hochberg method were used

with a threshold of .05 for statistical significance. If not enough genes were significantly expressed to make the plot, the p-value was adjusted to .1, and the LFC threshold to .5. Gray points in the volcano plot indicate genes that have smaller Log Fold Change, or non-significant p-values, the p-value threshold (of -log_{10} transformed adjusted p-values) is shown with a horizontal line, the higher the -log_{10} transform of the p-value, the lower the original p-value actually is [2].

3.5 Mutual Information

Mutual Information, also known as Information Gain, from probability and information theory, is a measure of dependence and shared information between two variables. Unlike Pearson correlation, Mutual Information does not only measure linear dependence [59]. Mutual Information is defined based on the joint probability distribution of two variables, and if the two random variables are statistically independent, then Mutual Information is 0 [58]. Mutual Information was estimated, rather than using the full equation, in this case between a continuous variable (the gene's expression levels over samples in the cohort dataset) and the categorical target variable (label of Tumor or Solid Tissue Normal) [78]. Scikit-learn's implementation of mutual information estimation, *mutual_info_classif*, was used, which is based on estimating entropy with a k-nearest neighbors algorithm [11, 56]. Two configurations of *mutual_info_classif* were used, one with n_neighbors=3, (MI) and the other with n_neighbors=5 (MI2). Mutual Information was estimated for one gene, or variable, at a time, and so like the other statistical methods, does not account for feature interaction. An example histogram of the Breast invasive carcinoma (BRCA) dataset gene expression Mutual Information estimation scores is shown in Figure 3.6.

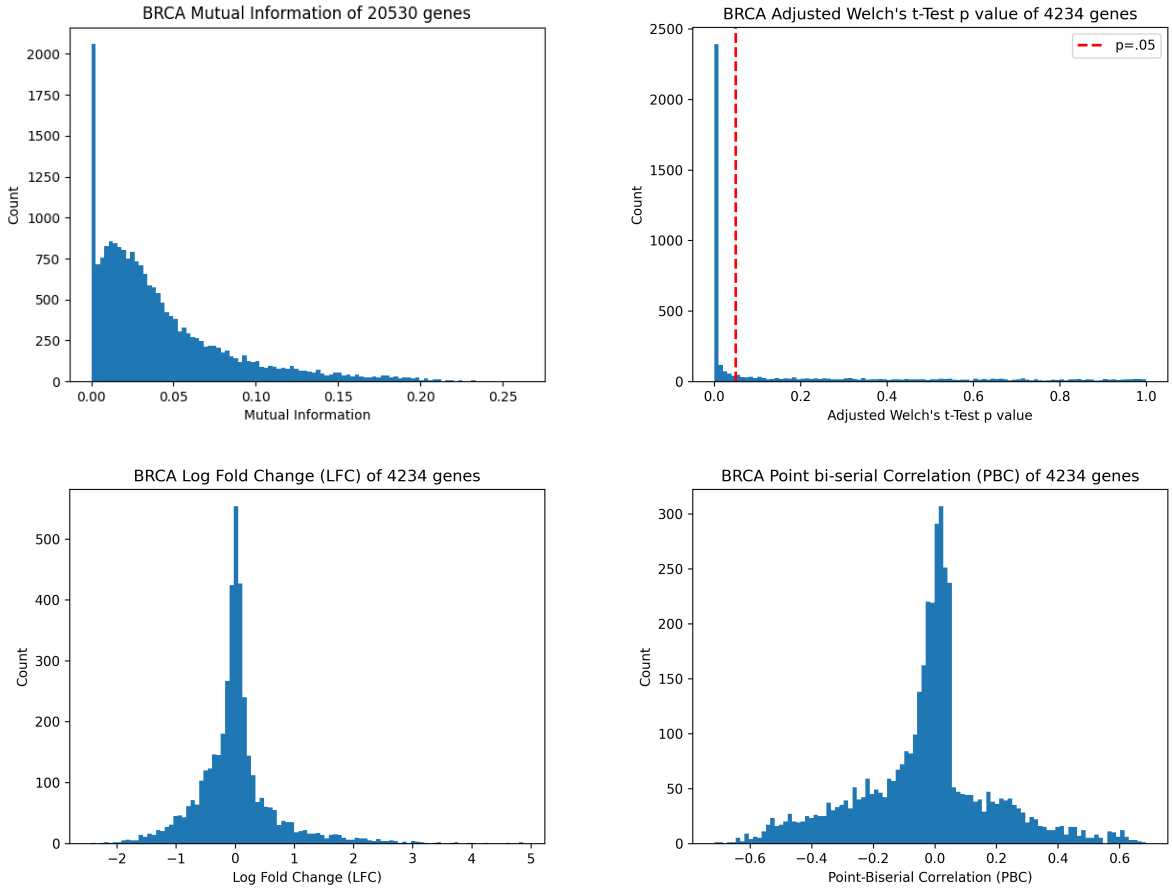

Figure 3.6: BRCA cohort example histograms: Top Left: Mutual Information. The MI estimator returns values between 0 and 1, with 0 being no measurable information. Top Right: Welch's t-test corrected p-value, Bottom Left: Log Fold Change (LFC). Bottom Right: Point-Biserial Correlation (PBC). A high number of genes which met conditions for Welch's t-Test are statistically significant in difference, with a corrected p-value threshold of .05.

3.6 Step 1: Classification Using All Genes

For each dataset, nine separate classifiers were trained and evaluated, and were used to create feature importance ranks. The nine classifiers as described in Table 3.1 were scikit-learn implementations of the following classifier models: Logistic Regression (x 3), Random Forest (x 4), and Linear Support Vector Classifier (LSVC) (x 2). All other hyperparameters are Scikit-learn defaults, which for Random Forest *num_features* is the square root of the number of features, $\sqrt{20,530} = 143$.

These model types were chosen for having inherent feature importance methods, and for being relatively quick to train. Logistic Regression contains coefficients based on assigning class probabilities, Random Forest has Mean Decrease Impurity (MDI), and the Linear SVC, as opposed to non-Linear SVC, also has interpretable feature coefficients for creating the SVM decision boundary. The models were also chosen for their efficiency when used with Permutation Importance and SHapley Additive exPlanations (SHAP) which are time-intensive calculations.

The hyperparameters were chosen based on a hyperparameter grid search using only the BRCA dataset, since it had the most features. Grid search hyperparameter optimization is time consuming, although performing the optimization on all datasets and averaging the results could possibly result in better performing models across all datasets. Different Random Forest hyperparameters impacted performance, and later on in the experiment, feature importance scores as well. The differences in classification performance of Logistic Regression and SVC (non-linear) across hyperparameter configurations were less pronounced.

Hyperparameters for the top performing model of each of the three Logistic Regression solvers were chosen. Hyperparameters for each of weighted and unweighted by class balance Linear Support Vector Classifiers were chosen. Hyperparameters for the best performing Random Forest models using 32, 64, 128, and 256 estimators were chosen. These decisions were made to create a diversity in configurations of models. There is a theoretical basis for using these models for feature importance. Random Forest feature importance has been

Table 3.1: The 9 inherently interpretable models used, Scikit-learn implementations and their non-default parameters

Model	ID	Hyperparameters
Logistic Regression	LR_0	solver='newton-cholesky', tol=.001, C=10
Logistic Regression	LR_1	solver='lbfgs', tol=1e-5, C=10
Logistic Regression	LR_2	solver='newton-cg', tol=1e-5, C=1
Linear SVC	LSVC_3	class_weight='balanced', tol=.001, C=.1
Linear SVC	LSVC_4	class_weight=None, tol=.001, C=10
Random Forest	RF_5	criterion='log_loss', max_depth=5, min_samples_split=2, min_samples_leaf=1, n_estimators=32
Random Forest	RF_6	criterion='log_loss', max_depth=10, min_samples_split=5, min_samples_leaf=1, n_estimators=64
Random Forest	RF_7	criterion='log_loss', max_depth=10, min_samples_split=2, min_samples_leaf=2, n_estimators=128
Random Forest	RF_8	criterion='entropy', max_depth=10, min_samples_split=2, min_samples_leaf=1, n_estimators=256

used before to iteratively find the best features among gene expression datasets for cancer classification [75, 17]. Linear Support Vector Machine coefficients are also an established method for scoring feature importance [22]. Logistic Regression coefficients have been part of biomarker search algorithms as well, and also often as an initial feature selection method [36].

3.6.1 Experiment Process

For each dataset, the feature values, already log_2 transformed and normalized, were further min-max scaled for both computational efficiency and to obtain scaled feature importance in linear models. Over two repeated trials, a 10-fold stratified k-fold split was created, each split using 90% of data for training, and a hold-out 10% test-set for evaluation. This resulted in training each model 20 times, each with a potentially different random split. Classification metrics were averaged over all 20 evaluations from trials and folds, as was the feature importance score generated for each feature.

Classification metrics are derived from True Positives, False Positives, True Negatives, and False Negatives. For this experiment, "Primary Tumor" was specified as the positive class by convention. For all datasets, there is a large class imbalance, where there are significantly more "Primary Tumor" samples than "Solid Tissue Normal" samples. Trial experiments with using class balancing techniques such as down-sampling and Synthetic Minority Oversampling Technique (SMOTE) on the BRCA dataset, showed no significant improvement or even a decrease in classification performance. In the interests of computational efficiency and simplicity of analysis of the final results, no class balancing was conducted in the main experiment.

Since the classes are significantly imbalanced in the test sets, regardless of any rebalancing techniques used on the training sets, simply predicting all samples as "Primary Tumor" could result in 90% accuracy or higher, which could deceivingly appear to be a well-calibrated classifier. For this reason, balanced accuracy, f1_score, and Mathews Correlation Coefficient (MCC) are potentially more helpful metrics. MCC is suggested as a better metric for some classification problems in research [27, 26]. In this experiment, MCC was used as the primary classification metric because the score is more sensitive to errors such as False Positives than f1_score. Here, TP is True Positives, TN is True Negatives, FP is False Positives, and FN means False Negatives.

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$\text{TPR (Recall)} = \frac{TP}{P}$$

$$\text{TNR (Specificity)} = \frac{TN}{N}$$

$$\text{Balanced Accuracy} = \frac{TPR + TNR}{2}$$

$$\text{PPV (Precision)} = \frac{TP}{TP + FP}$$

$$\text{F1 Score} = \frac{2 \cdot PPV \cdot TPR}{PPV + TPR}$$

Accuracy is simply the proportion of correct classifications out of all total predictions, which is potentially misleading with imbalanced datasets and biased predictions. Balanced accuracy is the average of True Positive Rate (TPR) and True Negative Rate (TNR) so is more sensitive to errors in the minority class, either positive or negative. F1 Score is the harmonic mean of Positive Predictive Value (PPV) or Precision, and True Positive Rate (TPR) or Recall. Both PPV and TPR are based only on positive predictions, which when the positive class is the majority and errors are few, as with this experiment, there is little difference in F1 score. The equation for Matthew's Correlation Coefficient (MCC) is sensitive to all errors, either False Positive or False Negative which bring down the score even with imbalanced classes. Thus, a small difference in the number of True Negatives or False Positives causes a wider discrepancy in the MCC metric. For this reason MCC is more useful with very imbalanced data and small difference in number of errors for evaluating classification performance.

Even when using these metrics that are more sensitive to errors in imbalanced datasets, classification performance was very high across these datasets and classifiers. This is probably due to the classification being a relatively easy problem for many cohorts, as is evident

sometimes by viewing the dimension-reduced plots of data points and the number of statistically different genes. Gene expression levels, measured by RNA Seq counts, are very different between normal tissue and primary tumor. As will be seen later, this makes the choice of genes as potential biomarkers difficult on the basis of classification performance.

Although classification is relatively easy, in the case where classification is being made between a potential tumor or normal tissue, even a small improvement in accurate prediction or confidence in the predictions can be very critical. For this reason, performance evaluation remains important in the selection of biomarkers for classification [60]. The secondary motivation was to attempt to narrow down potential genes of interest for further research in understanding the mechanisms of cancer or to investigate as potential therapeutic targets. In the second use case, similar classification performance evaluations give only little information as to which genes are better research targets.

3.6.2 Feature Importance (FI)

While baseline performance was being calculated using all genes, each model's inherent feature importance methods were also used for each fold in each trial and were averaged over all folds and trials. This feature importance rank is a model-specific inherently interpretable aspect of each model, and gathered post-hoc, meaning they are collected after the model is trained. It is relatively quick to access the coefficients in the Logistic Regression and Linear Support Vector Classifier, as they are calculated during training. Scikit-learn's Random Forest Classifier also has a quick method to return Mean Decrease Impurity (MDI) for each feature.

The Logistic Regression coefficients are the log odds for predicting labels based on feature values. A Linear Support Vector Classifier has coefficients for each feature adjusted for creating a division through the hyperplane to best separate data points between labels and does not remap the features into a different dimensional space as with other Support Vector Machine kernels. This means the magnitude of the coefficients relates linearly to feature values and hence feature importance in calculating the decision boundary. Mean Decrease

Impurity (MDI) is averaged over each split per each tree in the Random Forest that the target feature was involved in creating. Gini Impurity is a measure of the label distribution in the nodes resulting from a split. The average decrease in Gini Impurity by this feature is calculated over all splits in all estimators (trees) in the Random Forest.

3.6.3 Top 250 Lists

These feature importance results were then the basis for the next step of selecting a smaller subset of features based on feature rank. The reason for reducing the number of features first was so that SHapley Additive exPlanations (SHAP) and Permutation Importance (PI) could be employed without using excessive computational resources and time, which would be considerable using all 20,530 features. The number of 250 was chosen somewhat arbitrarily, but partly based on visualizations of the resulting initial feature importance from Step 1. The Feature Importance (FI) was plotted with bar graphs to determine if there were any patterns or clear thresholds for selecting top features. With BRCA as an example, a pattern emblematic of all cohorts is visible. Feature importance by Logistic Regression and Linear Support Vector Classifier smoothly declines well beyond the top 1000 features as illustrated in Figure 3.7.

Whereas with Random Forest models, the feature importance tend to drop more quickly as rank increases, and are even eclipsed by the standard deviation as in Figure 3.8. The same pattern was apparent for all other cohorts tested. Standard deviation is calculated for each feature's importance over trials and folds, then features are sorted, unnamed, by average feature importance per model. The features in each model's graph are ordered separately in Figures 3.7 and 3.8.

There was a higher relative standard deviation in all Random Forest (RF) Mean Decrease Impurity feature importance rankings than in the coefficients from Logistic Regression (LR) and Linear Support Vector Classifiers (LSVC). Logistic Regression had slightly less variation between trials than LSVC. Over each fold the same model is trained on a different subset of the data and can assign a different Feature Importance (FI) score to the same feature, the

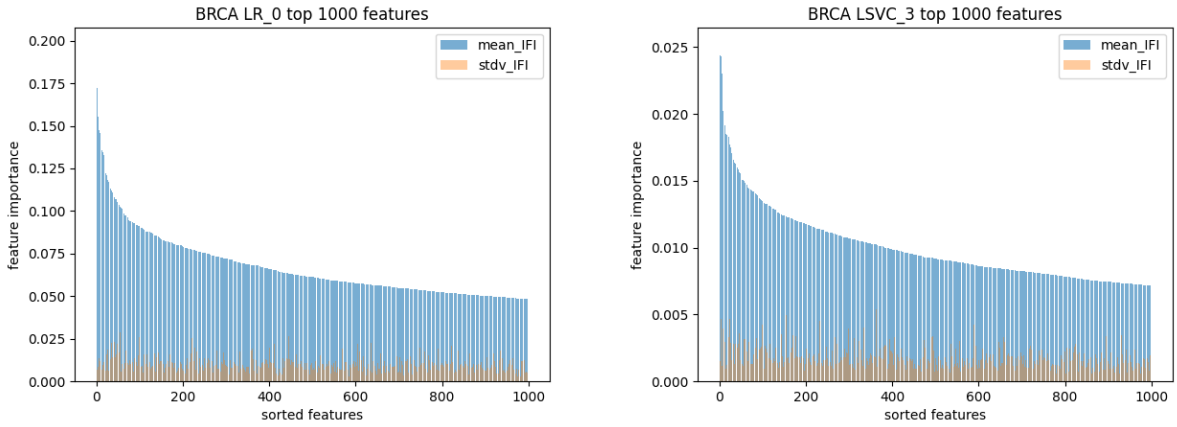

Figure 3.7: Left: Logistic Regression 0 (LR_0). Right: Linear Support Vector Classifier 3 (LSVC_3). Two models as examples of visualization of feature importance distributions on the Breast invasive carcinoma (BRCA) dataset. There does not appear to be a clear threshold for choosing top N features, but the slope is higher initially from 1 - 100 features by importance.

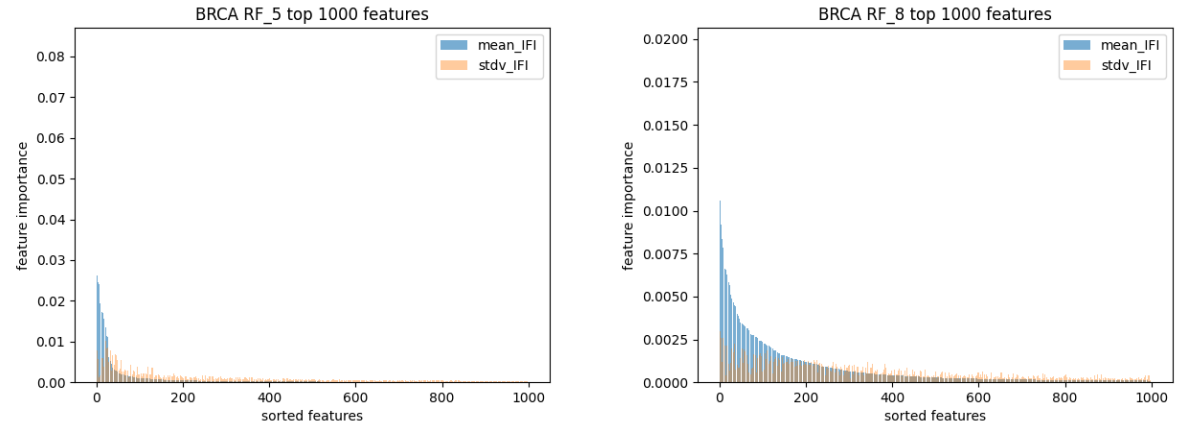

Figure 3.8: Left: Random Forest 8 (RF_8) with 256 estimators (trees). Right: Random Forest 5 (RF_5) with 32 estimators (trees). Both models trained on the Breast invasive carcinoma (BRCA) dataset. With the 32 estimator Random Forest (RF_5) the feature importance seems to significantly decline past the top 200 features.

mean and standard deviation of Feature Importance were then calculated. In the two trials, the 10 k-folds were chosen with a different random split.

Using the mean model feature importance generated in this step over trials and folds, the top 250 genes from each importance method were selected. The absolute value of the feature importance was used for Logistic Regression and Linear Support Vector Classifier, where the coefficients could be negative or positive indicating which class the feature tended to shift predictions towards.

Additionally, the top 250 genes by each Mutual Information estimation score were included as sets, along with the statistical methods. For Log Fold Change (LFC), first only the significantly different genes were considered, determined by the corrected p-value of a two-tailed Welch's t-test and Welch's t-test assumptions. For Point-Biserial Correlation (PBC) stats, also the genes considered were limited by Levene's test and adjusted PBC p-value. Among genes that met these conditions, the top 250 in order of absolute magnitude were used to create the lists. After selecting genes based on statistical significance the top 250 genes by absolute LFC and by Absolute Mean Difference (AMD) were identical, so AMD was discarded for the rest of the experiment.

3.6.4 OncoKB sets

OncoKB is a Memorial Sloan Kettering Cancer Center curated database of cancer genes, from various sources [9, 21, 83]. "The following genes are considered to be cancer genes by OncoKB™, based on their inclusion in various different sequencing panels, the Sanger Cancer Gene Census, or Vogelstein et al. (2013)" [9]. Using the OncoKB database, of which 1195 genes in these TCGA datasets belong to, an OncoKB-gene-only version of each list created by the above methods was compiled. In order of the ranks per method, the first genes that were also in the OncoKB list were added to a new set until 250 genes had been selected. In the case of statistics based lists, when not enough genes that had adjusted $p < .05$ were present in the OncoKB list, then the top ranked features not in the OncoKB list were added to make up the difference. This tests the question of whether known cancer-genes would

be better or worse than genes selected strictly based on ML feature importance rank or by other statistical or informational methods.

3.7 Step 2: Classification using 250 Genes, and Additional Feature Importance

Each of the nine original models were trained again, using the top 250 features by that model's feature importance rankings from Step 1. For the gene-sets that were generated using statistical or informational methods, Logistic Regression 0 (LR_0), as a simple and high performing model, was used for training. For each dataset, for each gene-set, classification metrics were gather again over two trials of 10 stratified k-fold splits. Permutation Importance (PI), and SHapley Additive exPlanations (SHAP) were also calculated on the new trained models.

After each model was trained, permutation importance (PI) was calculated, using 20 repeats per feature, although more repeats would be preferable it would add to experiment time. In previous versions of this experiment, using the BRCA dataset only, Permutation Importance (PI) did not compare well to other feature importance methods. Unlike with SHapley Additive exPlanations (SHAP), PI does not capture feature interactions. Over these 20 repeated permutations, an average feature importance, and a standard deviation over repeats are calculated. PI is calculated by randomly shuffling the values in the target feature column, and measuring the change in the model's classification accuracy from baseline. Both the importance and standard deviations over repeats were averaged and the standard deviation calculated over both trials of 10 k-fold splits.

Using the python shap library [12], an explainer for each trained model, at each trial, and at each k-fold was created. For Logistic Regression and Linear Support Vector Classifier, linear explainer was used, and for random forest, the tree based explainer was used. The shap library also supports a kernel explainer which takes much longer, so models were chosen ahead of time that supported these quicker types of explanations. The mean of absolute value of SHAP values was averaged over both trials of 10 k-folds. It would be of interest in future

work to experiment with more complicated models and explainers, to find if different top genes would be selected. Also, other more complex models might have better, or different, baseline classification performance.

Finally, the same model-specific inherently interpretable post-hoc methods were used again this time, on models trained with only 250 selected genes by the previous iteration of inherent feature importance. This results in four versions of the same 250 genes in different orders. Collectively referred to as Method-2s, the orders are by rank of either the original inherent feature importance/stats/MI (IFI_1), SHAP, Permutation Importance (PI), or inherent feature importance again (IFI_2).

3.8 Step 3: Gene-set Classification Performance Evaluation

To evaluate the gene-sets created in the previous step, models were trained again, this time only using 26 or fewer of the top highest-rated features. The number of features used, corresponded to number of features in the gene-sets from external research. For evaluation, all of the original nine models used in Step 1, with all 20,530 genes, were used again on every dataset. In addition, three new models were added as shown in Table 3.2. Performance metrics were again calculated and averaged over two trials and 10 folds per model per gene-set. Classification performance was measured in MCC due to the highest variation in scores and as being more sensitive to errors in imbalanced datasets. [26, 27] Subtracting the average random gene-set MCC from each gene-set's MCC is noted as MCC-R.

The three new models were chosen to have unbiased models with which to compare gene-sets created using feature importance from the original 9 models. A Decision Tree Classifier was used because it is simple with inherent interpretability of its own. The Support Vector Classifier (SVC) with a polynomial kernel was chosen, because SVCs with non-linear kernels were among the best performers in the original hyperparameter search using all genes. However, the non-linear Support Vector Classifier could not be used in Step 1, because it lacks inherent interpretability, and additionally, in Step 2, the shap kernel explainer required would be very computationally expensive. A non-linear SVC was also how Peng et al. evaluated

Table 3.2: Three additional models for evaluation besides the original 9 models. Non-default scikit-learn hyperparameters.

Model	ID	Hyperparameters
Support Vector Classifier - Polynomial Kernel	SVC_1	kernel='poly', C=1, gamma='scale', tol=0.001, probability=True, max_iter=20000
Decision Tree	DT_0	criterion='entropy', splitter='best', max_depth=6
Gradient Boosting Classifier	GBC_3	All Scikit-learn defaults

their own gene-set of 14 genes, the original inspiration for this experiment. The Gradient Boosting Classifier (GBC) was chosen as an alternative tree-based model with potential for high classification performance, and somewhat distinct from Random Forest.

The main objective of this experiment was to compare methods for biomarker identification, and Step 3 evaluated ranked lists of biomarkers as ML features. Overall, 52 distinct identification methods arose from ranking methods in Step 1 (on 20,530 genes) and Step 2 (on top 250 genes.) The combinations of 13 Method-1s from Step 1 and four Method-2s from Step 2 resulted in 52 separate rankings per dataset. The results of other researcher's biomarker identification processes were also used for comparison with these methods in Step 3. Method-1 was the first feature ranking in Step 1 with which the top 250 genes were selected. The 13 Method-1s were the nine inherently interpretable ML models, Mutual Information 1 and 2, Point-Biserial Correlation, and Log Fold Change. Method-2 re-ranked the 250 genes (or left them in the original order.) The four Method-2s were original order, SHAP, PI, and inherent feature importance again. With Step 3 results, each combination of Method-1s and Method-2s was ranked according to average performance over all gene-set lengths (GSLs) and all cohorts. The average of each Method-1 and Method-2 sets were also

compared. There were also 52 sets of OncoKB only genes which were used for comparison with all-gene sets bringing the total to 104. The average performance of all methods per cohort was also compared. A table of the methods is available in Table 3.3.

The re-ranking of 250 genes becomes significant only when using subsets of fewer than 250 features, such as the top 26 through top 2 genes from each list as features for ML evaluation models. In addition to using features from all 104 lists for each Gene-set-length (GSL) for each cohort, genes from published papers in the related works section were also used as ML features. The compared biomarker methods from external research are available in Table 3.4. The specific gene-names in each list can be found in the appendix in Table A.2.

Table 3.4: There were 12 lists of biomarkers from external research for specific datasets and with specific Gene-Set-Lengths (N Genes.) For comparison, the same number of genes from each of the 104 experimental lists for the specific dataset, were also used as features for training evaluation models.

Authors and Publication Date	N Genes	Dataset	Method Summary
de la Guardia-Bolívar et al. 2024 [34]	26	Combination of COAD, BRCA, LUAD, KIRC, STAD, LIHC, THCA, and UCEC	Paired Differential Gene Expression (pDEA)
Das et al. 2024 [33]	24	PRAD	"Decipher" Clinical gene panel for risk of Prostate Cancer Recurrence
Kallah-Dagadu et al. 2025 [50]	22	BRCA	iML wrapper method
Peng et al. 2015 [71]	14	Combination of BLCA, BRCA, COAD, HNSC, LIHC, LUAD, and LUSC	Differential Gene Expression (DGE), pathway analysis, co-regulated gene clusters

Nikitina et al. 2017 [68]	12	PRAD	Differential Gene Expression (DGE), Formalin-Fixed Paraffin-Embedded (FFPE) benign prostate hyperplasia vs. PRAD
Wan et al. 2015 [91]	10	All Cohorts	"Num Types" Differential Gene Expression (DGE) multiple cohorts
Wan et al. 2015 [91]	9	All Cohorts	"At Least One" Differential Gene Expression (DGE) multiple cohorts
Zheng et al. 2021 [99]	6	ESCA	Differential Gene Expression (DGE) and hub genes, Esophageal squamous cell carcinoma (ESCC)
Coleto-Alcudia et al. 2022 [28]	5	LUNG (LUAD and LUSC)	Artificial Bee Colony based on Dominance (ABCD) multi-objective optimization algorithm.
Coleto-Alcudia et al. 2022 [28]	4	Combination of BRCA, LIHC, ESCA, TGCT, THCA, LUAD, and LUSC	Artificial Bee Colony based on Dominance (ABCD) multi-objective optimization algorithm.
Coleto-Alcudia et al. 2022 [28]	3	BRCA	Artificial Bee Colony based on Dominance (ABCD) multi-objective optimization algorithm.

Coleto-Alcudia et al. 2022 [28]	2	PRAD	Artificial Bee Colony based on Dominance (ABCD) multi-objective optimization algorithm.

20 random gene-sets selected from all original feature genes were added to the list of feature-sets to evaluate. Additionally, 20 random gene sets selected only from the OncoKB database of 1195 genes were also included. The averages of these random gene-sets were used as benchmarks for comparison of performance with other feature ranking methods, given that the classification task was already very tractable. The random set results were averaged for each cohort and each gene-set length (GSL) tested.

3.9 Step 4 Biomarker Analysis

Besides classification performance, other forms of analysis were used to evaluate potential biomarkers: overlapping genes, pathway analysis, and automated annotation search.

3.9.1 Overlap Analysis of Gene-sets

Among constructed gene-sets in this experiment, the amount of appearances, and corresponding ranks of unique genes among all gene-sets was calculated. A score for each gene was calculated based on the number of appearances, rankings, and performance of the sets the gene appeared in. This scoring was done within each cohort specifically, and then over all cohorts in the attempt to find possible general cancer biomarkers. Top genes by this score were chosen for further analysis.

3.9.2 Pathway Analysis

Among the highest scoring features, pathway information was collected from the Kyoto Encyclopedia of Genes and Genomes (KEGG). For each gene, an automated web query of

Table 3.3: Thirteen Method-1s and four Method-2s. Method 1 is the first ranking system used in Step 1 to select 250 out of 20,530 genes. Method 2 re-ranks the top 250 genes. In combination this creates 52 lists. There is also an OncoKB-only version of each list for another 52 lists. All of the lists were created separately for each of the 21 datasets (16 individual cohorts, and 5 combinations.)

Method 1	**Abbreviation**
Logistic Regression 0	LR_0
Logistic Regression 1	LR_1
Logistic Regression 2	LR_2
Linear Support Vector Classifier 3	LSVC_3
Linear Support Vector Classifier 3	LSVC_4
Random Forest 5	RF_5
Random Forest 6	RF_6
Random Forest 7	RF_7
Random Forest 8	RF_8
Mutual Information 1	MI
Mutual Information 2	MI2
Point-Biserial Correlation (absolute value)	PBC_abs
Log Fold Change	LFC
Method 2	**Abbreviation**
Inherent Feature Importance 1 (unchanged from Method 1)	IFI_1
Inherent Feature Importance 2 (on top 250 Genes in Step 2)	IFI_2
SHapley Additive exPlanations	SHAP
Permutation Importance	PI

the KEGG database was conducted using the KEGGRESTpy python package. This query returned all pathways that the gene is listed as a part of. In addition, if KEGG listed any diseases associated with the gene, these were returned as well. In some cases the KEGG query returned no pathways at all, and for some genes the query failed completely to return any results, even after searching with gene name aliases. Among top features per cohort, the count of pathways these genes belonged to was collected.

3.9.3 Annotation Search

Among all the highest scoring genes, a different automated web search was conducted based on keywords to find if the genes had previously been associated with cancer. The MyGene.Info python package [7] was used to pull gene annotations from multiple online databases. Among all annotations collected this way, a search for any of the following keywords was conducted: *cancer, onco, carcino, tumor, metastatic, metastasis, melanoma, leukemia, lymphoma, myeloma, malignan, chemotherapy, chemo-therapy, sarcoma, tamoxifen, meningioma, meningeal tumor, mesothelioma, blastoma.* Any gene where at least one keyword was found among the annotations was considered to be a match for cancer keywords.

Chapter 4

RESULTS

Each step in the experiment generated different results for analysis. Classification metrics were averaged over 2 trials and 10 stratified folds for every model trained on each dataset and each subset of features, or gene-sets. In addition to classification metrics, feature importance scores were calculated in Step 1 on all 20,530 genes and Step 2 on only the top 250 genes from each Step 1 ranking system.

4.1 Step 1 Results

4.1.1 Baseline Performance Using All Genes

Results from Step 1 in this experiment include of the classification performance of the chosen models using all 20,530 genes, across datasets. There were 16 individual cohorts, and 5 combinations resulting in 21 total datasets. There were 9 model types trained and tested, creating 9 x 21 = 189 separate performance evaluations. Table 4.1 contains the Mathews Correlation Coefficient (MCC) as a classification metric for each cohort for each of the 9 models: 3 Logistic Regressions, 2 Linear Support Vector Classifiers, and 4 Random Forests. Model hyperparameters are provided in the methods section in Table 3.1

Plots for each cohort and for each method are available in the supplemental materials. One approach for plotting all results together is to average the performance of all models for each dataset, showing which datasets are most challenging, even with all genes available. Another way to analyze the results in aggregate is to compare the performance of each of the selected models averaged over all datasets.

Table 4.1: Step 1 average MCC results per dataset per model

Classifier	Logistic Regression			Linear Support Vector		Random Forest			
cohort	LR_0	LR_1	LR_2	LSVC_3	LSVC_4	RF_5	RF_6	RF_7	RF_8
BLCA	0.87	0.83	0.83	0.84	0.84	0.68	0.61	0.63	0.63
BRCA	0.98	0.98	0.98	0.98	0.98	0.95	0.95	0.95	0.94
COAD	1	0.99	0.99	1	1	1	1	1	1
ESCA	0.73	0.78	0.78	0.78	0.78	0.58	0.63	0.63	0.63
HNSC	0.95	0.96	0.96	0.95	0.95	0.91	0.92	0.92	0.93
KICH	1	1	1	1	1	1	1	1	1
KIRC	0.98	0.97	0.98	0.98	0.98	0.97	0.97	0.97	0.97
KIRP	1	1	1	1	1	0.98	0.98	0.97	1
LIHC	0.95	0.96	0.96	0.96	0.96	0.94	0.95	0.93	0.94
LUAD	0.98	0.98	0.97	0.99	0.99	0.94	0.95	0.96	0.96
LUNG	0.99	0.98	0.99	0.99	0.99	0.96	0.97	0.97	0.97
LUSC	0.99	0.99	0.99	0.98	0.96	0.98	0.98	0.97	0.97
PRAD	0.77	0.78	0.77	0.78	0.76	0.7	0.7	0.71	0.7
READ	1	1	1	1	1	0.9	0.9	0.9	0.9
STAD	0.95	0.97	0.97	0.97	0.97	0.92	0.92	0.9	0.89
THCA	0.95	0.95	0.95	0.94	0.94	0.92	0.93	0.93	0.93
UCEC	0.99	1	1	1	1	0.89	0.89	0.89	0.86
Combination [71]	0.97	0.97	0.97	0.97	0.97	0.91	0.93	0.93	0.93
Combination [28]	0.97	0.96	0.96	0.96	0.96	0.91	0.93	0.93	0.93
Combination [34]	0.97	0.97	0.97	0.97	0.97	0.9	0.93	0.93	0.94
PANCAN_selected	0.96	0.96	0.97	0.97	0.97	0.85	0.91	0.92	0.92

4.1.2 Cohort/Dataset Evaluation

High accuracy was achievable using all models and all cohorts. However, since all datasets were imbalanced, with a higher prevalence of Tumor labels, accuracy was not the best measure to distinguish between classification performance. F1 score and balanced accuracy are better measures for imbalanced class classification, but MCC is the most sensitive to misclassifications. Error bars indicate Standard Deviation of the averages over cohorts. Orange standard deviation bars indicate standard deviation over trial/folds. On average, models had higher standard deviation of trial results in Esophageal carcinoma (ESCA), meaning the random selection of 90 percent training data and 10 percent testing data had a large impact on classification performance. Figure 4.1.

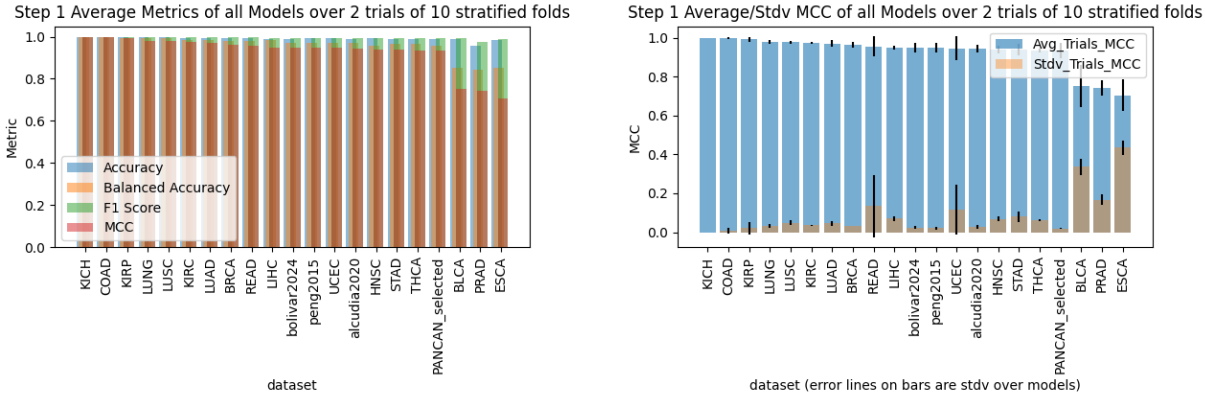

Figure 4.1: Step 1 performance results for each dataset, average of all 9 models. Left: There is little difference between Accuracy, or F1 score. Balanced accuracy and MCC are better metrics for slight differences in classification performance. Right: MCC average of 9 models, with standard deviation over trials in orange, error bars are standard deviation over the 9 models.

Kidney Chromophobe (KICH) had the highest average classification performance over all models, with 100% on all 9 models. Esophageal carcinoma (ESCA) had the lowest classification performance averaged over all models, and also the highest standard deviation

across different model's MCCs. Bladder Urothelial Carcinoma (BLCA), Prostate adenocarcinoma (PRAD), and ESCA were more difficult to predict than even PANCAN_selected, the combination of all single-cancer type cohorts together.

KICH average accuracy was 1.0, and ESCA average accuracy was 0.983, a very small difference. KICH balanced accuracy was 1.0 and ESCA was 0.854 a difference of .146. KICH had on average 100% classification accuracy, hence 1.0 f1 score, and 1.0 MCC. ESCA f1 score was .991, and ESCA MCC was .705.

4.1.3 Model Evaluation

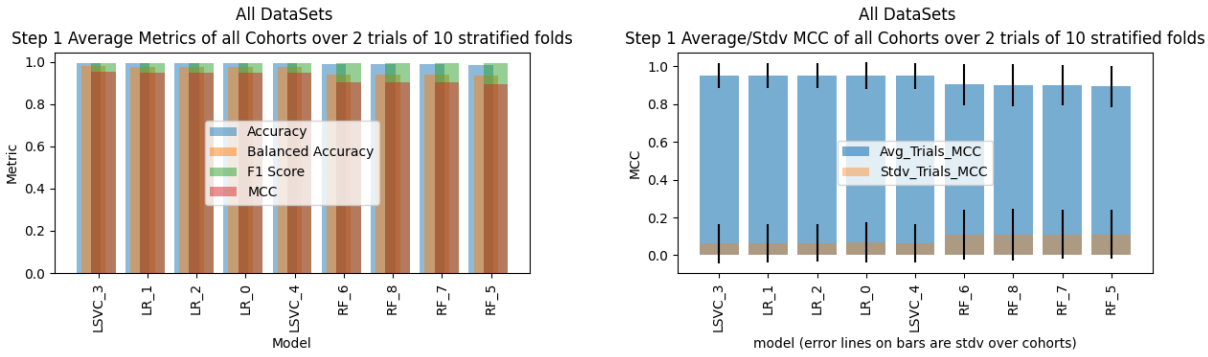

Figure 4.2: The average performance of each model over all datasets, including the combination datasets. Left: All metrics, Right: MCC with standard deviation over trials in orange, and standard deviation between datasets as error lines. Linear Support Vector Classifier (LSVC), Logistic Regression (LR), Random Forest (RF)

Average classification performance of all models over all datasets is generally very high using all 20,530 genes as seen in Figure 4.2. The differences in both Logistic Regression and Linear Support Vector Classifier models is negligible. Random Forests have very similar performance with each other and are slightly worse than the other classes of models. The main

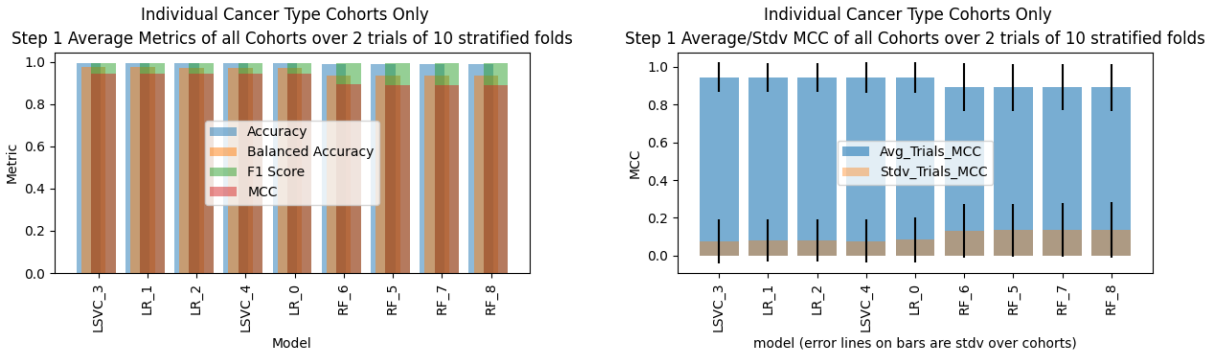

Figure 4.3: The average performance of each model over all individual cancer type cohorts. Left: All metrics, Right: MCC with standard deviation over trials in orange, and standard deviation between datasets as error lines. Linear Support Vector Classifier (LSVC), Logistic Regression (LR), Random Forest (RF)

purpose in including different models was to test different methods for feature importance ranking. Accuracy, and f1 scores are all very high due to class imbalance, so the MCC metric will be used exclusively for the rest of the experiment report.

Linear Support Vector Classifier 3 (LSVC_3) had the best performance averaged over all cohorts with .949 MCC, and averaged over single cancer type cohorts only with .952 MCC. Random Forest 5 (RF_5) had the lowest classification performance with .896 MCC over all cohorts, and .895 MCC over single-cancer-type cohorts only. This pattern in the average performance of all cohorts remained the same for individual cohorts.

A finding from Step 1 is that when all 20,530 features are available, Linear Support Vector Classifier and Logistic Regression are generally better classifiers for these datasets than Random Forest. Linear Support Vector Classifier and Logistic Regression are very similar in performance across different architectures, and with each other as shown in Figure 4.2. Random Forest models assigned a higher feature importance to relatively fewer features than did the Logistic Regression and Linear Support Vector Classifier models. Random

Forest Classifiers were constrained by the number of estimators (32 to 256), a maximum depth of 5 or 10, and a random subsection of features per tree of 143 (The square root of 20,530). So it is possible that not all features are used, possibly accounting for slightly poorer relative performance to the linear models. In contrast, the linear models of Linear Support Vector Classifier and Logistic Regression always assigned weights to all of the features/genes available.

The number of total samples in the cohort is not correlated with performance, but the number of the negative class, also the minority class, Solid Tissue Normal (STN) samples, in the cohort shows a slight correlation with performance in Figure 4.4. The proportion of normal tissue samples in the cohort, which is the degree to which the classes are imbalanced, is even more correlated with average model performance. Some cohorts are still especially challenging and there is still not a very clear strong linear relationship relating to samples, or class imbalance.

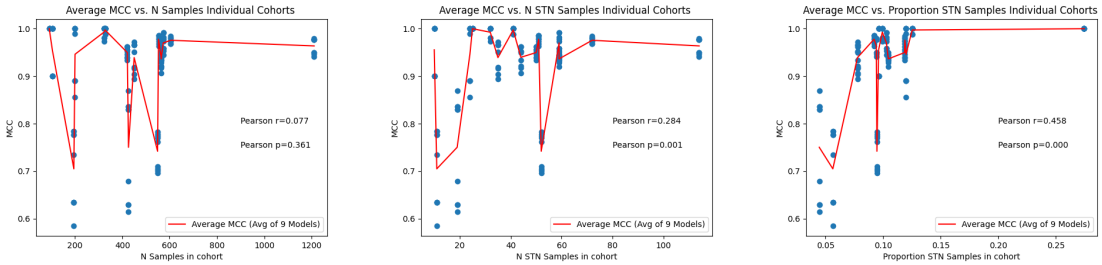

Figure 4.4: Individual cancer type cohorts only. Left: Number of samples in each cohort versus MCC, Center: Number of Solid Tissue Normal (STN) samples, the minority class, versus MCC. Right: Proportion of dataset Solid Tissue Normal, versus MCC.

4.1.4 Top 250 Genes

As explained in the methodology, each of the 9 models had model specific inherently interpretable attributes which allowed collecting feature importance. Along with statistics and

Mutual Information, for each of the 21 datasets, separate lists of feature rankings were created. Similarity plots on a per cohort basis are available in the supplementary materials. The pattern of model similarity was consistent across cohorts/datasets. The average number of overlapping genes, and the average difference in rank of features between ranking systems over all cohorts are displayed in Figure 4.5.

As was expected, Mutual Information 1 and 2 often selected mostly the same genes, as the only difference was a small change in the n_neighbors hyperparameter. Logistic Regression 1 and 2 usually chose mostly the same genes as well with, on average over cohorts, 227/250 overlapping. Between Logistic Regression 1 and 2 only the solver and regularization hyperparameters were different. Mutual Information 1 and 2 are less similar with each other as far as the ranks assigned to genes than Logistic Regression 1 and 2 were to each other. Linear Support Vector Classifier models, where the only differences were the class balance and regularization hyperparameters, shared on average 244/250, the most, overlapping genes. The average difference in ranks between Linear Support Vector Classifiers was 10, showing the ranks of features were also on average very similar as in Figure 4.6. Linear Support Vector Classifier's top 250 features shared between 153-193 overlapping genes with Logistic Regression models. The top 250 most important features chosen by random forests were in general very different from those chosen by the Linear Support Vector Classifiers and Logistic Regressions. Random Forest features were also less similar with each other across configurations. Interestingly, Random Forest models seemed to often select genes as important that were also selected by ranking genes individually with Mutual Information estimation.

The least amount of common genes was found between Log Fold Change (LFC), Absolute Value Point-Biserial Correlation (PBC or PBC_abs), and among all ML models' feature importance. Log Fold Change and Average Mean Distance, when selected only from significantly different genes, were identical across all cohorts. PBC_abs and LFC were moderately similar with around 73/250 on average. Mutual Information and PBC_abs/LFC had low overlap. Random Forest models were somewhat closely overlapped with each other, but

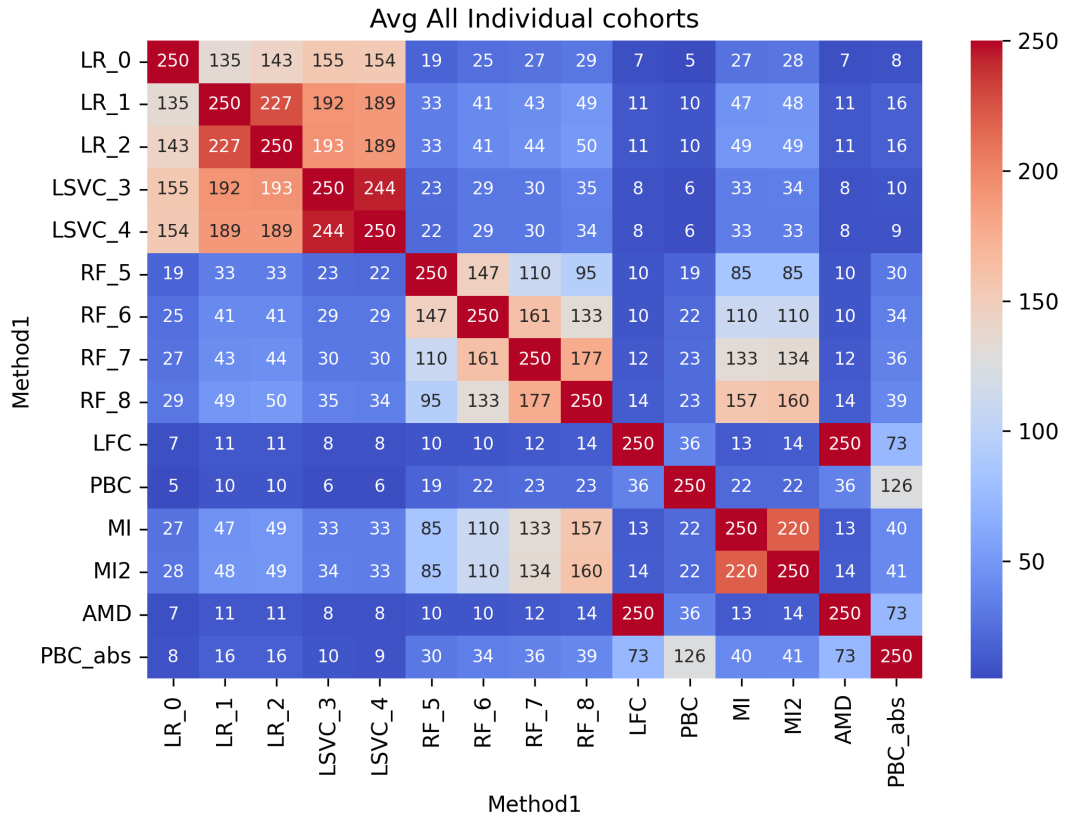

Figure 4.5: The average number of overlapping genes between lists of top 250 genes averaged among all datasets. Logistic Regression 1 (LR_1) and Logistic Regression 2 (LR_2) and Mutual Information 1 and 2 (MI, MI2) each shared 220/250 genes on average over all datasets. Linear Support Vector Classifier 3 (LSVC_3) and Linear Support Vector Classifier 4 (LSVC_4) had 244/250 genes overlapping on average over all datasets, the highest of any pair of non-identical lists. Random Forest (RF), Point-Biserial Correlation (PBC), Log Fold Change (LFC), Absolute Mean Difference (AMD)

much less so than Linear Support Vector Classifiers and Logistic Regressions were, which is not completely surprising given that the Random Forest models had more differences in hyperparameters and architecture.

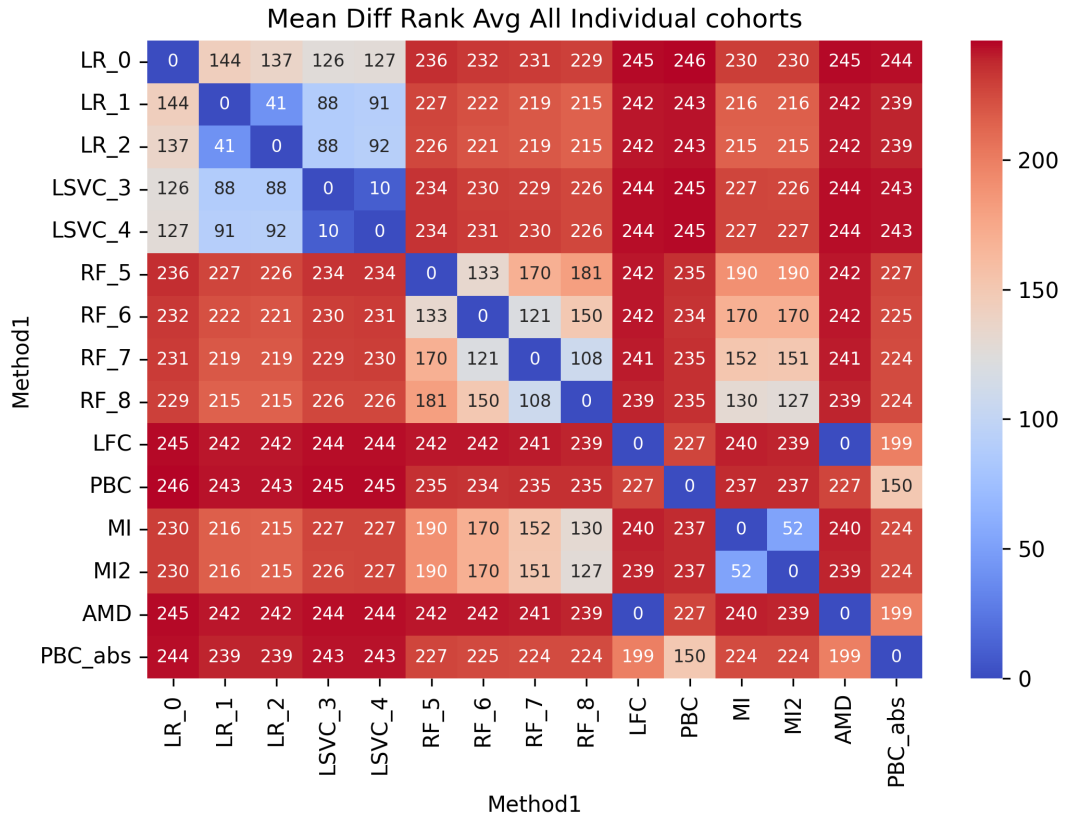

Figure 4.6: The average difference in rank among genes in each list is another way to measure similarity. Genes not in both sets were counted as having 250 rank difference. This plot shows the average difference in rank between lists, averaged over all cohorts/datasets. The average difference in rank was only 52 for Mutual Information 1 and 2 (MI, MI2) and 41 for Logistic Regression 1 and Logistic Regression 2 (LR_1, LR_2). Linear Support Vector Classifier 3 and Linear Support Vector Classifier 4 (LSVC_3, LSVC_4 had on average over datasets, average 10 rank difference per gene. Random Forest (RF), Point-Biserial Correlation (PBC), Log Fold Change (LFC), Absolute Mean Difference (AMD)

4.2 Step 2 results

4.2.1 Results, Performance with 250 Genes

In Step 2, models were trained using the top 250 highest ranked genes by feature importance ranking, statistical measures, or Mutual Information, from Step 1. This was done to accomplish three goals: to compare performance between using all features, and the selected top 250 features, secondly to compare performance between using OncoKB genes only, and all available The Cancer Genome Atlas genes. And third, to calculate SHAP, PI, and inherent feature importance again (IFI_2) as further methods for biomarker selection. Of the previous gene-sets, Mutual Information (MI), Log Fold Change (LFC), and Point-Biserial Correlation (PBC_abs) generated lists were used to train Logistic Regression 0 (LR_0). All other gene-sets were used to train the same type and configuration of model that had been used to select the list from all 20,530 genes. All feature importance or measures were based on absolute magnitude, considering a feature important regardless of if it helped the model to predict the positive or the negative class.

4.2.2 Cohort/Dataset Evaluation

For each cohort, the average and standard deviation of Mathew's Correlation Coefficient (MCC) over trial/folds of all methods was averaged. The top 250 gene-sets taken from OncoKB only are not included in the averages in Figure 4.7.

Between datasets there is varying classification performance achievable using lists of 250 top ranked features. As with using all genes in Step 1, Bladder Urothelial Carcinoma (BLCA), Prostate adenocarcinoma (PRAD), and Esophageal carcinoma (ESCA) are the most difficult and there is also more variation between tested feature-sets/models shown in Figure 4.7.

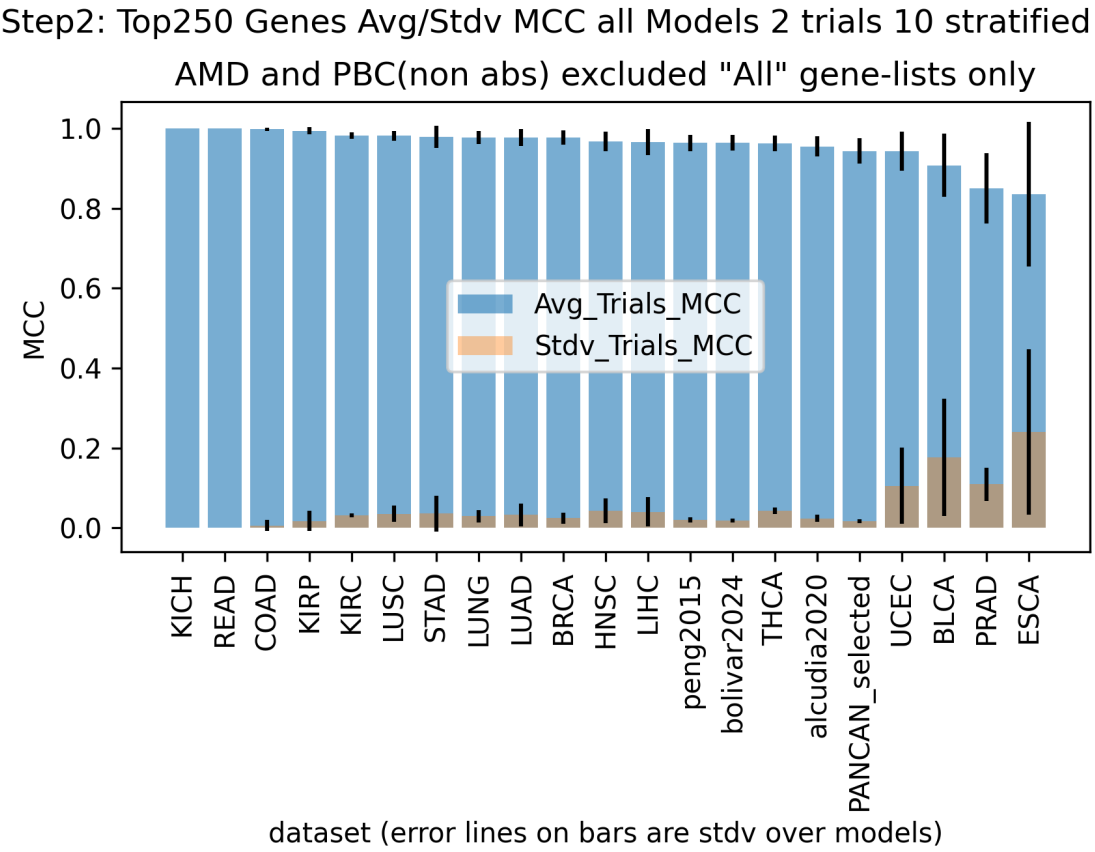

Figure 4.7: Step 2 classification metrics using lists of top 250 genes. For each dataset the MCC is averaged over all gene-sets; the order of cohorts by average MCC is similar to Step 1: BLCA, ESCA, and PRAD are still the bottom three. Error lines are standard deviation across methods.

4.2.3 Feature List and Model Evaluation

In Step 2, rather than simply comparing models which were using the same 20,530 features, as in Step 1, the comparison is between models now using *differing sets* of 250 features. Results are displayed in Figure 4.8. Methods for top-250 gene selection had different relative performance in different datasets. There were 21 datasets tested with 13 methods, the specific dataset results are available in the supplemental materials. The specific performance

of methods in the Breast invasive carcinoma (BRCA) and Esophageal carcinoma (ESCA) cohorts are shown for example in Appendix Figure A.1.

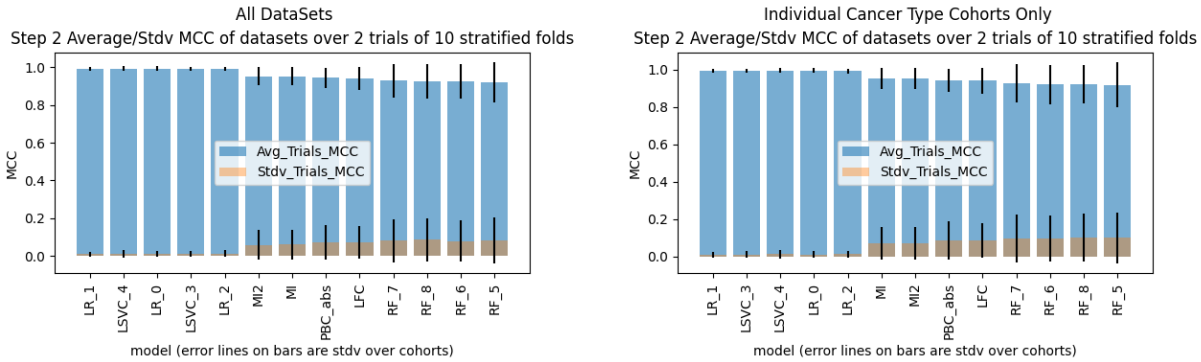

Figure 4.8: The average MCC of classification using the top 250 genes of datasets plotted for each gene-set creation method. Left: Average of all datasets, Right: average of individual cohorts only. There is not a great difference between all datasets and individual cohorts. As in Step 1, Logistic Regressions (LR) and Linear Support Vector Classifiers (LSVC) perform best. Mutual Information (MI), Log Fold Change (LFC), and Point-Biserial Correlation (PBC) were intermediary between the Logistic Regressions/Linear Support Vector Classifiers and the Random Forests (RF)

4.2.4 Comparisons

Two types of general comparisons between classification performance metrics are possible in Step 2. The first comparison is the classification performance between the same model trained on all 20,530 features (genes) and on only the top 250 features by that model's inherent feature importance rankings from Step 1. The second comparison is between the same model trained on the top 250 genes selected from all genes, and the top 250 selected from OncoKB genes.

All genes versus OncoKB genes

In the all-gene sets, any gene among the 20,530 original genes in the TCGA dataset could be used to create the 250 gene-set. In the OncoKB-gene sets, only genes that were in the 1195 genes in the OncoKB database listed as cancer genes were included. In the statistics ranking sets, given that genes had to be statistically significantly different in their expressions, some non-OncoKB genes were included to bring the list up to a length of 250. A comparison of set results is shown in Figure 4.9. Each dataset is different; for example, BRCA on average is classified better and the difference between OncoKB-gene sets and all-gene sets is small. BLCA, which was one of the more difficult datasets, was classified better by all-gene sets than OncoKB-gene sets by a larger amount, as in Figure A.2 in the Appendix.

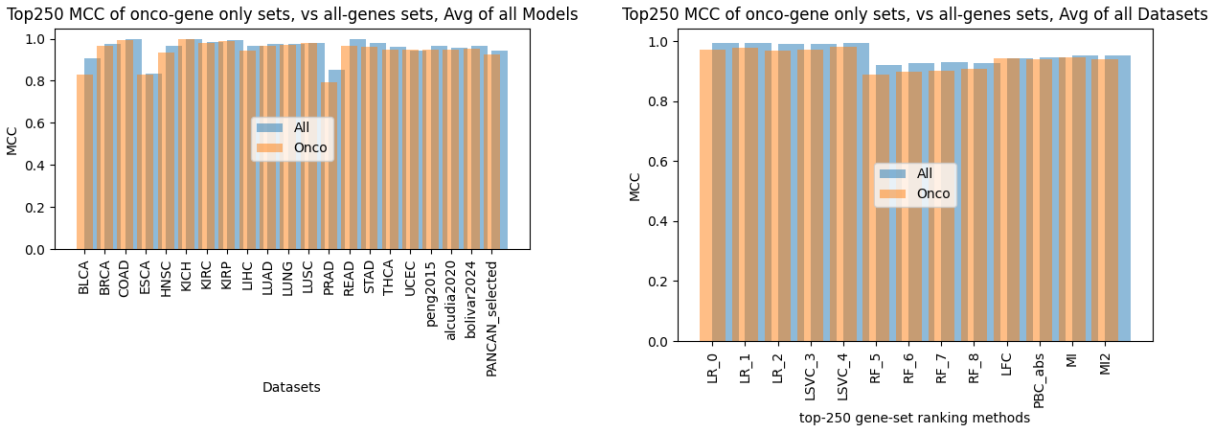

Figure 4.9: Average MCC of all-genes compared to OncoKB-gene only gene-sets. Left: The average of all methods per cohort. Right: The average of all datasets per method. In most datasets all-genes on average lead to better performance. In most methods, the all-gene version leads to better performance averaged over all datasets than the OncoKB-gene version. Logistic Regression (LR), Linear Support Vector Classifier (LSVC), Random Forest (RF), Point-Biserial Correlation (PBC_abs), Log Fold Change (LFC), Mutual Information (MI)

Results of Step 2 show that when using the top 250 genes to train and evaluate these classifiers, no improvement in classification performance is obtained by using only OncoKB-genes, but actually performance is statistically significantly better using the top 250 genes from among all genes as in Figure 4.9. Comparing results from all datasets and all feature lists/models, all-gene methods had on average slightly better performance as shown in Figure 4.10.

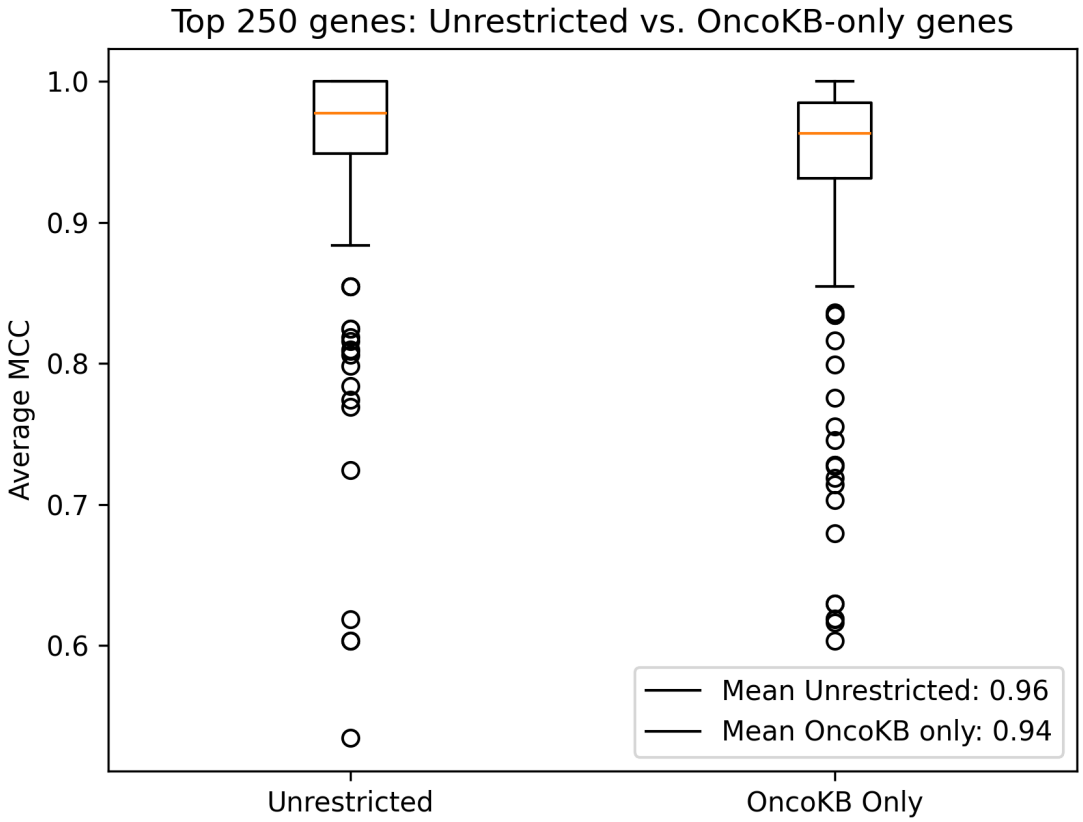

Figure 4.10: Among all models trained on 250 features in Step 2, on average there is no improvement in using only OncoKB-genes rather than all genes. In fact, the average of the unrestricted group is higher. The groups are not normally distributed, and had equal variance. A two-sided Mann-Whitney U test found the groups to be statistically significantly different with p=.0.000011

250 genes versus 20,530 genes

Another comparison was possible between model performance trained in Step 1 using all 20,530 genes, and the same models trained using the top 250 genes by feature importance. Using only 250 selected features leads to on average better performance than fitting models with all 20,530 available features. Comparative classification performance is shown in Figure 4.11. Box-plots and a statistical analysis show that all-gene sets lead to average better performance as shown in Figure 4.12.

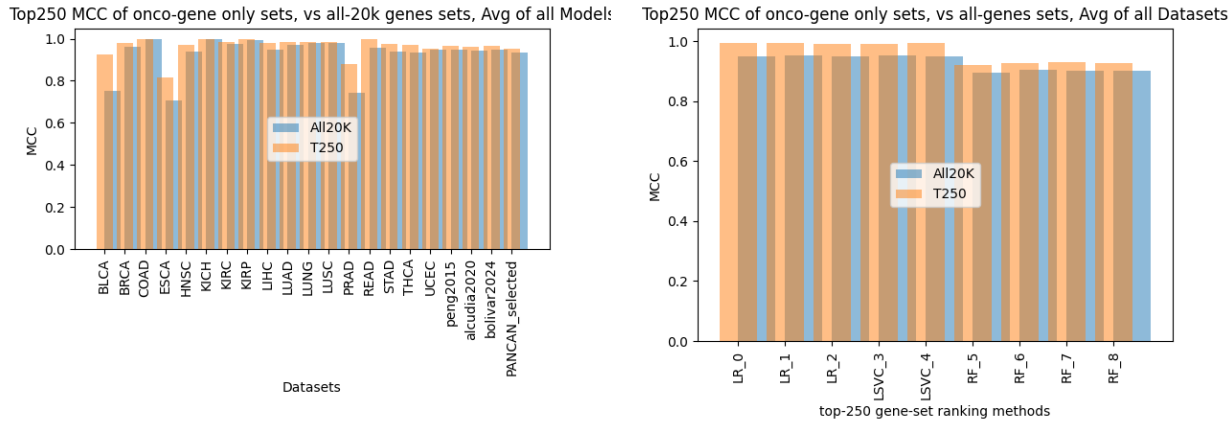

Figure 4.11: Average MCC of models trained on all 20,530 features and the top 250 features by inherent feature importance. Left: datasets plotted with the average of all models. Right: models plotted with the average of all datasets. Among all datasets and all models, on average the top 250 features lead to better classification performance. Logistic Regression (LR), Linear Support Vector Classifier (LSVC), Random Forest (RF), Point-Biserial Correlation (PBC_abs), Log Fold Change (LFC), Mutual Information (MI). T250 = Top 250 Ranked genes as features. All20K = All 20,530 genes used as features

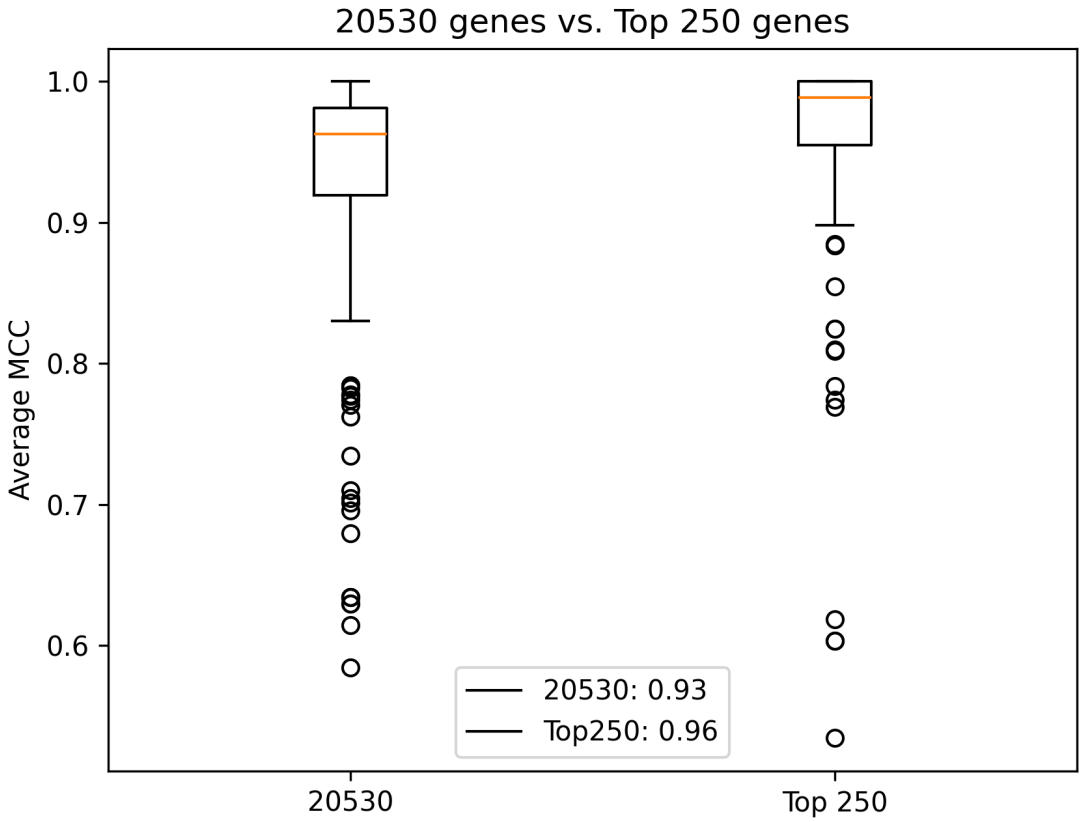

Figure 4.12: Among all models trained on 250 features in Step 2, on average there is an improvement over the same model trained on 20,530 genes. The groups are not normally distributed and had equal variance. A two-sided Mann-Whitney U test found the groups to be statistically significantly different with $p = 3 \times 10^{-9}$

4.2.5 Proportion OncoKB-genes selected

While selecting only OncoKB cancer genes for gene-sets did not on average improve performance, it is still of interest whether OncoKB-genes are selected more often than at random by all-gene rankings methods. This is one method of evaluating selection methods based on the rate of finding already known biomarkers. Out of 1195 OncoKB cancer genes which also were in this experiment's The Cancer Genome Atlas data, a further 463 are labeled as "is onco-gene," and 396 as "is tumor suppressor gene." In this paper, OncoKB-genes refer to any listed cancer gene in the OncoKB database, but OncoKB's designation of "onco-gene," refers to the specific nature of this cancer gene. The distribution of OncoKB gene counts in compiled sets and the top 20 gene-sets by OncoKB count are shown in Figure 4.13. The average OncoKB count per method averaged over individual cohorts is shown in Figure 4.14.

4.2.6 Method 2: 2nd Ranking of top 250 Features

For each original ranking from Step 1 (On all 20,530 genes), three additional re-rankings were created based on the models trained in Step 2. The same top 250 genes were present, but re-ordered based on a new feature importance score, either SHapley Additive exPlanations (SHAP), Permutation Importance (PI), or model-specific Inherent Feature Importance calculated again (IFI_2). This creates a total of four versions of the same list including the original order from Step 1, which was inherent feature importance when using 20,530 genes, statistical measures, or Mutual Information. To compare similarity between ranking lists, the average difference in rank between genes in the lists was calculated, as shown in Figure 4.15. On average, inherent feature importance for the 2nd time with 250 genes (IFI_2) is closer to SHAP with 250 genes than to the original inherent feature importance using all 20,530 genes (IFI_1). Permutation Importance is the most different of all. An example of the different top 20 genes by four re-ranking methods with 250 genes selected by Logistic Regression 0 (LR_0) initially for the BRCA cohort is shown in Figure 4.16.

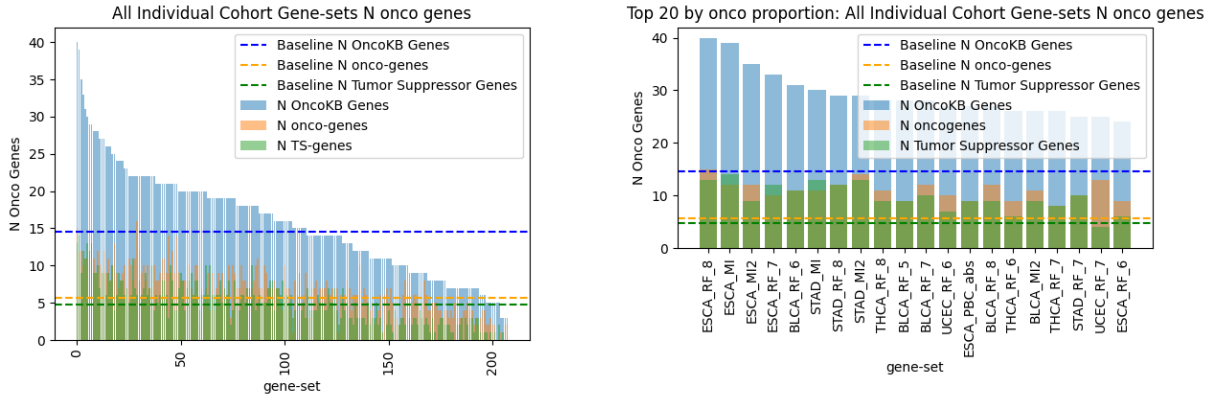

Figure 4.13: Left: For all gene-sets constructed for individual cancer type cohorts, the number of genes in the OncoKB database are plotted below unlabeled in descending order. Also plotted in orange are OncoKB designated "onco-genes," and in green designated as tumor suppressor genes. Baselines show the expected counts for 250 randomly selected genes from The Cancer Genome Atlas data. Right: the top 20 gene-sets by OncoKB gene count. Gene-sets for the Esophageal carcinoma (ESCA) cohort are the top four. In the right hand plot, gene-set abbreviations are cohort_Method1_model id. Logistic Regression (LR), Linear Support Vector Classifier (LSVC), Random Forest (RF), Mutual Information (MI), Point-Biserial Correlation (PBC_abs)

4.2.7 SHAP

SHapley Additive exPlanations (SHAP) is a model-agnostic interpretable Machine Learning (iML) method. SHAP and other model-agnostic iML techniques can be applied to a variety of different models, rather than inherent feature importance, which is model specific. SHAP values are calculated by estimating the impact on model prediction by the values of each feature, based on game-theoretical Shapley Values [80, 12]. Shapley Value was originally a game-theory concept, involving players playing a cooperative game. Applied to machine learning, features are treated as players cooperating in the game of making a correct prediction. Assigning a fair payout to a player is an analogy to describe how SHAP assigns importance to a feature. SHAP can account for feature interactions, and works by train-

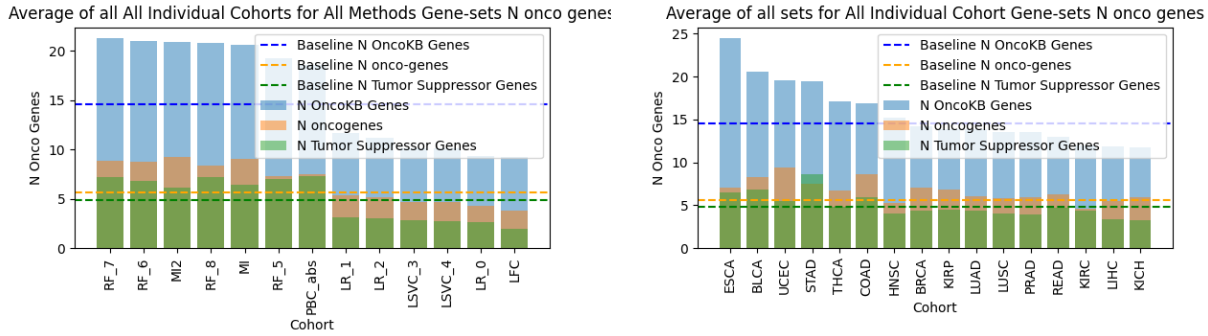

Figure 4.14: For all gene-sets constructed for individual cancer type cohorts, the number of genes in the OncoKB database are plotted. Also plotted in orange are OncoKB designated "onco-genes," and in green designated as tumor suppressor genes. Baselines show the expected counts for 250 randomly selected genes from The Cancer Genome Atlas data. Left: Average of all cohorts by method. Random Forest (RF), Mutual Information (MI), and Point-Biserial Correlation (PBC_abs) sets contain more OncoKB-genes on average than Logistic Regression (LR), Linear Support Vector Classifier (LSVC), and Log Fold Change (LFC). Right: Average of all methods per cohort, ESCA, BLCA, UCEC, and STAD have the most OncoKB-genes on average.

ing the ML model with all possible coalitions of features (or approximating this,) hence an exponential computation requirement increase with the number of features [66]. SHAP summary plots from three different models and gene-sets from the BRCA dataset are shown to illustrate SHAP values for feature importance, and to show feature selection overlap in Appendix Figure A.3.

4.2.8 Permutation Importance

Permutation Importance (PI) is another model-agnostic iML method. One feature at a time is randomly shuffled and the impact on the model's accuracy is measured over repeats, 20 in this experiment. Unlike other feature importance methods, PI can be negative, which

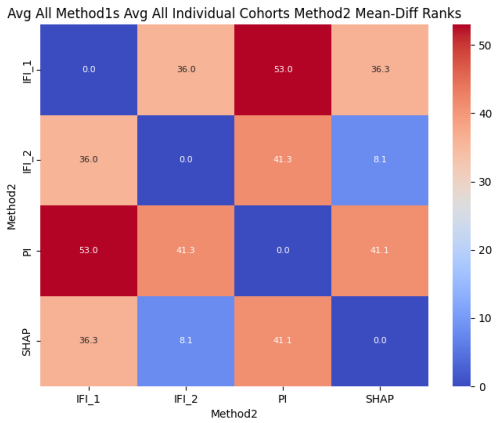

Figure 4.15: The top 250 genes from each Method 1 (inherent feature importance, statistics, or Mutual Information (IFI_1)) were re-ranked three times by Inherent Feature Importance again (IFI_2), SHapley Additive exPlanations (SHAP) and Permutation Importance (PI). An average over all individual cohorts of the average difference in rank of genes is shown above. Typically IFI_2 and SHAP are not very different from each other, and moderately different from the original ranking system (IFI_1). PI is the most different of all methods.

means effectively removing that feature actually improves the model's performance. In other feature importance methods a positive or negative value indicate whether the feature tends to influence prediction of the positive or negative class. In this experiment, the absolute value of all feature importance methods besides PI was used. For some models trained on some cohorts there was little or no PI at all, or only negative PI. PI usually resulted in feature ranks that diverged most from other scores. Permutation Importance results varied from dataset to dataset and between models, for a small example in the BRCA dataset, Logistic Regression (LR_0, and LR_1) and Random Forest (RF_8) are shown in Figure A.4 in the Appendix.

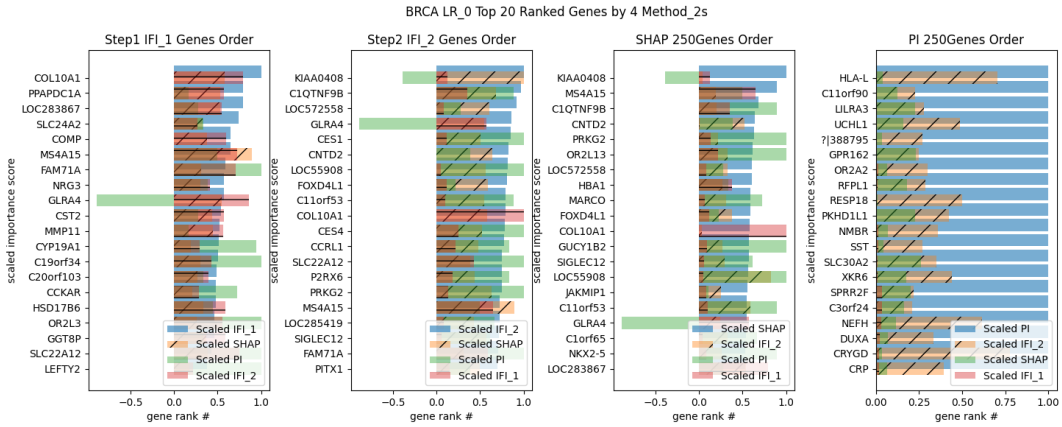

Figure 4.16: For the Breast invasive carcinoma (BRCA) cohort, Logistic Regression 0 (LR_0) was used to create a list of the top 250 genes by inherent feature importance (IFI_1). This plot shows the different scaled scores of each Method 2 of the top 20 genes. Left: original order, Inherent Feature Importance 1 (IFI_1), 2nd from Left: Inherent Feature Importance 2 (IFI_2), 2nd from Right: SHapley Additive exPlanations (SHAP), Right: Permutation Importance (PI.) This is one example how the re-ranking of Method-2s differ or are similar to one another.

4.2.9 Overlapping Unique Genes Between Datasets

For each dataset, 13 gene-sets of the top 250 genes were created (excluding OncoKB sets). Taking the unique genes of all 13 sets, it is of interest to see which datasets contained the most overlapping genes. Similarity between cohorts of adjacent or similar organ systems would support the idea that the selected genes have biological relevance. See Figure A.5 in the Appendix for a visualization of number of unique genes selected in all top 250 sets per cohort.

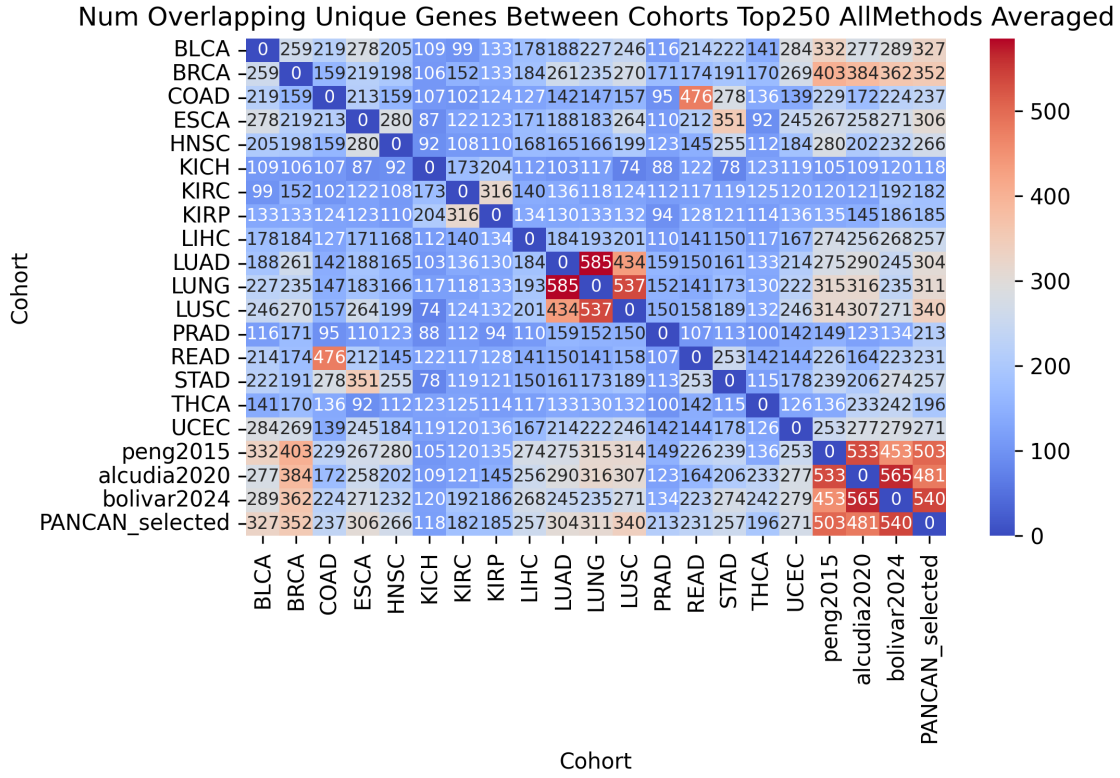

Figure 4.17: For each dataset, the number of overlapping unique genes in all top 250 lists. As is expected the combination datasets shared many genes with the individual cohorts used to create the combination. The combination sets themselves were similar. LUNG, the combination of Lung adenocarcinoma (LUAD) and Lung squamous cell carcinoma (LUSC) shared the most with LUAD and then LUSC individually. LUSC and LUAD themselves share many overlaps. Colon adenocarcinoma (COAD) and Rectum adenocarcinoma (READ) share many overlaps. Stomach adenocarcinoma (STAD) and Esophageal carcinoma (ESCA) are similar. Kidney renal clear cell carcinoma (KIRC) and Kidney renal papillary cell carcinoma (KIRP) are similar, but less so with Kidney Chromophobe (KICH).

4.3 Step 3 results

4.3.1 Results, Classification with smaller subsets

Using the top N genes from each ranking list, 3 new distinct models in addition to the original 9 were used as evaluation models. Each model was fitted with N features from each method, and then classification performance was measured on each dataset. N was between 26 and 2, to conform to the set lengths of external research comparison sets. There were 12 values of N tested: 26,25,22,15,14,12,10,9,6,5,4,2. Additionally, N of 24, 13, and 3 were used only for comparison with external research gene-sets where one or more genes were missing from the The Cancer Genome Atlas data. The purpose of the classification is to evaluate the methods used to create the feature lists. A list of biomarkers is intended to be small to allow for cheaper testing, and more focused research. Evaluating each list at multiple N is a method for evaluating the ranks of the features chosen, as at smaller N, only the highest ranked features will be used.

In Step 3, the evaluation step, unlike in Step 2, inherent feature importance, SHAP, or Permutation Importance were not collected. These calculations in Step 2 added computation time and limited the types of models that could be used. In Step 3, each list was evaluated using all 12 evaluation models, unlike in Step 2, where only the same model which originally created the list was used. Or, Logistic Regression 0 (LR_0), which for Log Fold Change, Point-Biserial Correction, and Mutual Information was used in Step 2. In addition to the original 9 models, a Support Vector Classifier with a polynomial kernel was included. Support Vector Classifiers with more complex than linear kernels were among the best models in earlier hyperparameter search tests. In addition, Peng et al. used a Support Vector Classifier for their final evaluation of the 14-gene signature that their methods yielded. A Decision Tree Classifier was also used to give a diversity of perspectives, and because with few amounts of features this simple model is more appropriate, it also contains its own inherently interpretable properties. Lastly, a Gradient Boosting Classifier (GBC) was included for further model diversity, and because it is a tree-based model of a different variant than

Random Forest.

4.3.2 Random Baseline

The 12 evaluation models were trained on 20 random genes of gene-set length N for each N tested between 26 and 2. The average MCCs of 20 random gene-sets used to train evaluation models are shown in Figure 4.18. Logistic Regression 2 (LR_2) is noticeably worse than other models. Logistic Regression with "newton-cg" (Newton Conjugate Gradient) as the solver hyperparameter, *tol*=1e-5, and regularization C=1, does not do well with a small number of randomly selected features.

From Gene-Set-Lengths 26 through 10, the Polynomial Support Vector Classifier 1 (Polynomial Kernel SVC) has the best average performance, but as N features drops below 10, the Polynomial Kernel SVC's relative performance drops. Gene-Set-Length N affects the relative performance of models: Decision Tree Classifier (DT_0) improves in comparison to other models as the number of features decreases, as does Gradient Boosting Classifier (GBC_3). Logistic Regression 0 (LR_0), Logistic Regression 1 (LR_1), and Linear Support Vector Classifier 4 (LSVC_4) do better than the Random Forests (RF) with higher number of features, but as number of features drop, RFs begin to do better. Linear Support Vector Classifier 3 (LSVC_3) has the best average performance on random genes with Gene-Set-Length N=2. LSVC_3 is with class balancing hyperparameter set to true, unlike LSVC_4. Linear Support Vector Classifier 3 (LSVC_3) regularization is stronger with *C=.1* as opposed to LSVC_4 with *C=10*. C is the scikit-learn hyperparameter which is inversely proportional to regularization (which is meant to avoid overfitting) strength. In addition to 20 randomly chosen gene-sets from all 20,530 The Cancer Genome Atlas genes, 20 randomly chosen gene-sets chosen from only OncoKB genes were also tested. On average, for each Evaluation Model trained on random sets of Gene-Set-Length N, the OncoKB sourced random sets lead to slightly better classification as shown in Figure 4.18 but without much clear consistent pattern across models and Gene-Set-Lengths as shown in Appendix Figure A.6.

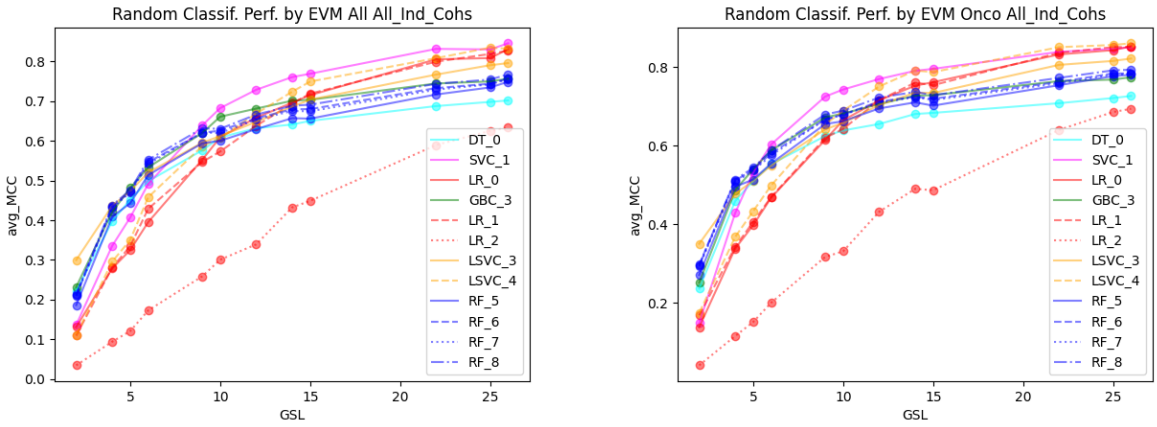

Figure 4.18: Random Gene-set from All genes (Left), and from OncoKB only (Right) classification performance by Evaluation Model. Average of all individual cohorts (All_Ind_Cohs). Each evaluation model's MCC is plotted separately against gene-set-length (GSL). As expected classification performance on average decreases with less features.

4.3.3 Evaluation Model performance on experiment sets

To see how different evaluation models performed with experimentally generated feature sets as Gene-Set-Length changes, the average of all experimental sets was taken. There were 52 experiment gene-sets per each of the 16 individual cohorts, the results of each genes set were averaged together over these 16 cohorts. In contrast to the random gene-sets, all of these gene-sets were assembled using Step 1 (Inherent Feature Importance with 20,530 genes, statistical measures, or Mutual Information) (IFI_1) and Step 2 ranking methods (On Step 1's top 250 genes: Inherent Feature Importance #2 (IFI_2), SHAP, or PI). Logistic Regression 2 (LR_2), which did poorly with random genes, does well until about 10 features or less, when performance drastically decreases. This model configuration apparently does well with enough useful features, but is worse than other models in using few or poorer quality features, as seen in Figure 4.19.

Gene-sets in this experiment were constructed by different methods with a theoretical

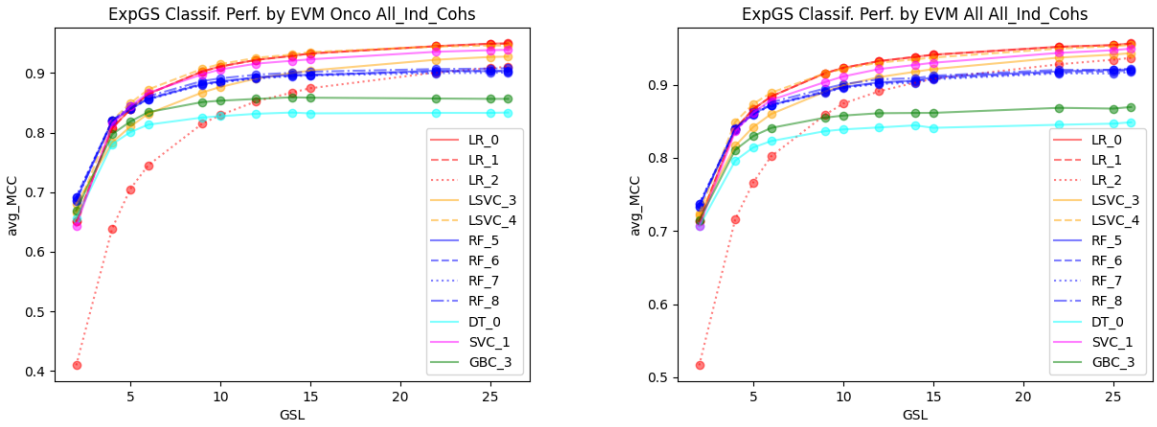

Figure 4.19: Average MCC of all individual cohorts over all experiment sets per Evaluation Model. Left: All-gene sets, Right: OncoKB-gene sets. Average of all individual cohorts (All_Ind_Cohs). Each evaluation model's MCC is plotted separately against gene-set-length (GSL.) In both plots, MCC declines with number of features, and the difference between models is greater at 26 features (GSL=26) than at 2.

basis for selecting more important features, and were expected to better than features selected at random. The random gene baseline shows that often the classification task can be accomplished decently even with randomly selected genes, indicating that there are possibly many genes which contain useful information. The gap between selected and random features widens as the number of features decreases. Plotting the average of all experiment gene-sets by Evaluation Model and against Gene-Set-Length: MCC-R shows classification performance above random baseline, and MCC All - OncoKB (MCC A-O) shows classification performance of all-gene sets versus OncoKB-only gene-sets. See Figure 4.20.

The all-gene random baseline MCC for random sets, across all individual cohorts, across all tested gene-set lengths, across all evaluation models is .569. When using random gene-sets of OncoKB genes only, the MCC baseline is .616. The majority of the 52 methods using all genes averaged over all individual cohorts and all tested gene-set lengths were above both baselines. For each cohort, and each Gene-Set-Length, MCC-R is the MCC of the same

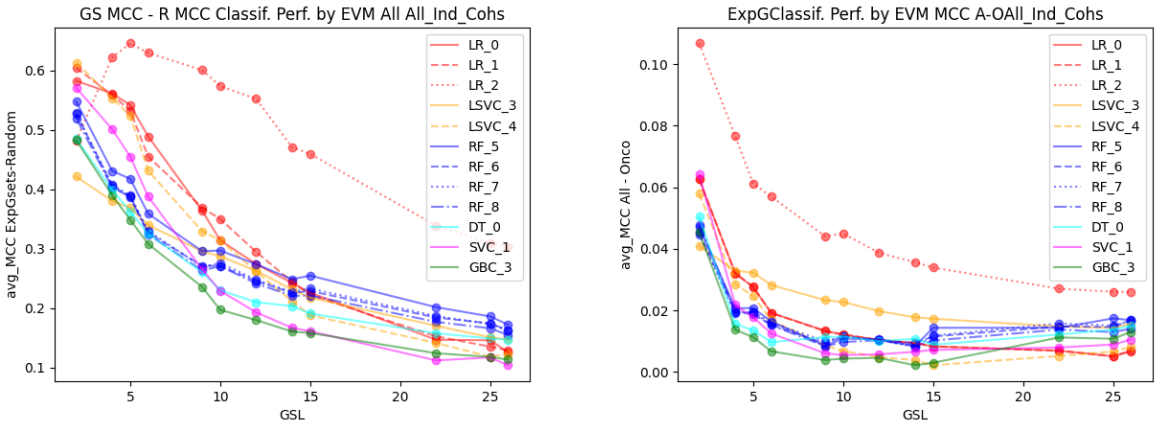

Figure 4.20: Left: Average of all experiment sets average of all individual cohorts MCC minus random set MCC (MCC-R). All-gene sets. As gene-set length (GSL) decreases, MCC-R increases. Right: For each Evaluation Model the average MCC of all all-gene sets minus the average MCC of all OncoKB-gene sets. The difference is small, but positive on average for all models, showing unrestricted gene-sets on average were slightly better for classification than OncoKB-only gene-sets

Evaluation Model trained on an experimental set minus the random MCC. Averaging MCC-R over all Gene-Set-Lengths, all Evaluation Models, and all individual cancer-type cohorts, the order of experimental method performance is shown in Figure 4.21.

Part of the motivation of using 12 different Evaluation Models was to evaluate the quality of the features selected, which were gene-sets of potential biomarkers. Since there was a diversity of model architectures, using the average of all Evaluation Models means the effects of any Evaluation Model and feature-set interaction will be smoothed. Averaged over all Evaluation Model, all cohorts and all Gene-Set-Lengths, the best over-all method was Mutual Information with inherent feature importance. Which was mutual information estimation to select the initial top 250 genes, and then Logistic Regression 0 (LR_0) inherent feature importance to re-order them. Mutual Information and Random Forest are common in the top 20, as well as Point-Biserial Correlation, and there is very slight difference among the top methods. Among the worst sets on average, Log Fold Change and Permutation

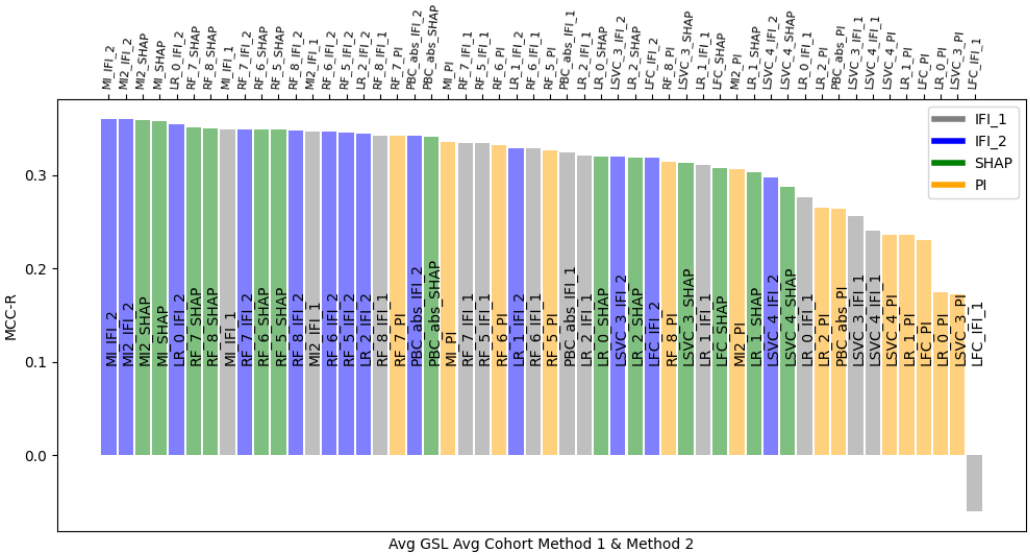

Figure 4.21: Average MCC-R over all individual cohorts (Avg Cohort) and all Gene-Set-Lengths (Avg GSL). All Methods but the Log Fold Change original ranking are above random baseline. Method abbreviation is in the form of Method1_Method2. Method-1s: Mutual Information (MI, MI2), Logistic Regression (LR_0, LR_1, LR_2), Linear Support Vector Classifier (LSVC_3, LSVC_4), Random Forest (RF_5, RF_6, RF_7, RF_8), Point-Biserial Correlation (PBC_abs), and Log Fold Change (LFC). Method-2s: Inherent Feature Importance 1 (IFI_1) (Step 1 original ranks), Inherent Feature Importance 2 (IFI_2) (Step 2 on 250 Genes), SHapley Additive exPlanations (SHAP), and Permutation Importance (PI).

Importance (PI) re-ordering are common, but the difference is still not very pronounced.

The Step 3 results MCC-R for 52 all-gene sets were also categorized and averaged by Method 1, the ranking for selecting the first 250 genes, and Method 2, the method for re-ranking those features. Mutual Information, Random Forest, Inherent Feature Importance 2nd round (IFI_2), and SHAP were the best ranking systems on average as shown in Figure 4.22. Averaging all 52 methods, all Gene-Set-Lengths, the order of datasets by average classification of all methods remained little changed from Step 1 (classification with 20,530

genes and 9 models) and Step 2 (Classification with 250 genes and 9 models), with Kidney Chromophobe (KICH) still the most successfully classified, and Prostate adenocarcinoma (PRAD), Esophageal carcinoma (ESCA), and Bladder Urothelial Carcinoma (BLCA) having the most errors as in Figure A.8.

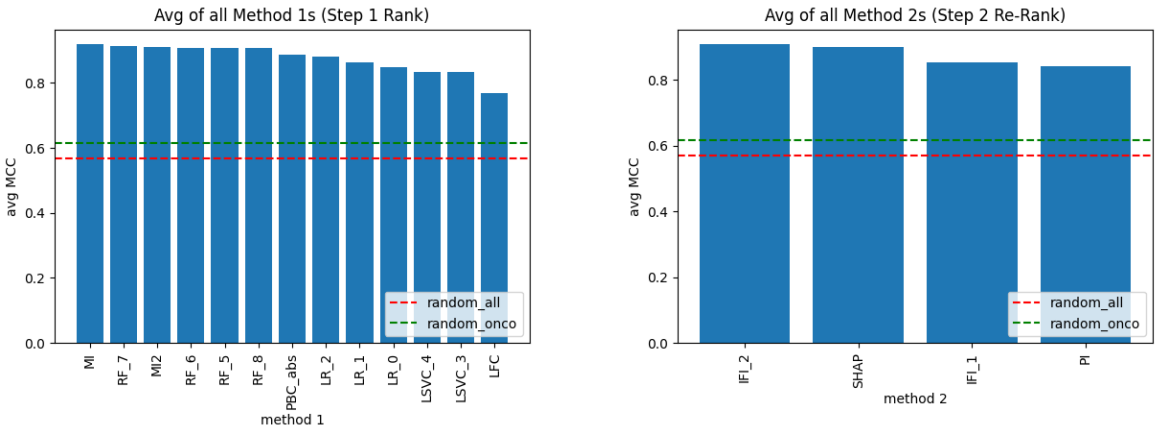

Figure 4.22: MCC performance averaged over all cohorts, and all Gene-Set-Lengths. Left: average by method 1, Right: average by method 2. Mutual Information (MI, MI2) and Rand Forest (RF) are top among Method-1s. Inherent feature importance #2 (IFI_2) was the best re-ranking Method-2 on average, closely followed by SHapley Additive exPlanations (SHAP). Logistic Regression (LR), Linear Support Vector Classifier (LSVC), Point-Biserial Correlation (PBC_abs), Log Fold Change (LFC). Inherent Feature Importance 1, Step 1 original order (IFI_1), Permutation Importance (PI)

4.3.4 Biomarkers from other Research Comparison

In the following figures, all 12 evaluation model (EVM) results were averaged together, the standard deviation between EVM is represented by error lines on the bars.

The previous analysis of best methods averaged all Gene-Set-Lengths and all cohorts. For each specific Gene-Set-Length tested, and each specific dataset, different best methods may emerge. In the following sections, each list of biomarkers from an external published paper was used as features for training the 12 evaluation models for the dataset specific to

that research paper. The results from the 12 models' MCC were averaged. The equivalent top-ranked number of genes from each of the 104 ranking systems in this experiment for that dataset were also used to train models separately. (There were 52 lists in total, and 52 OncoKB only versions of the same lists.)

The comparison is between identical ML models trained on different sets of features. Each set of features were the results of the selection and ranking of genes in Step 1 and Step 2. So, by comparing the performance of ML models using a small subset of features, the hope is to compare which methods were best for arriving at sets of features that lead to best classification performance.

Gene Set Length (GSL) 26, de la Guardia-Bolívar et al.

de la Guardia-Bolívar et al. provided the names of 26 genes of the 27 selected by paired differential expression analysis. The dataset used was the combination of 8 cohorts: COAD, BRCA, LUAD, KIRC, STAD, LIHC, THCA, and UCEC, which are the cohorts that the researchers used to assemble the list of biomarkers. Using the top 26 features selected in the experimental methods with this combination dataset, the top 5 best methods were all based on Mutual Information estimation (MI) with between .95 to .94 MCC. The MCC of the researcher's biomarkers evaluated by 12 Evaluation Models averaged was .907 which was 54th out of 105 methods, with a mean of .895 and standard deviation of .058. See Figure A.9 in the Appendix.

Gene Set Length (GSL) 24, "Decipher" Das et al.

Das et al. provided 25 genes that are a part of the "Decipher" gene-signature, which was not created by the researchers but included in their survey. Decipher is intended to a give a risk score for recurrence of Prostate cancer. One of these genes was not found by any other alias or ID in this The Cancer Genome Atlas data, so only 24 genes of the 25 were used, based on the Prostate adenocarcinoma (PRAD) cohort dataset. As these biomarkers were not intended for classification of normal or healthy tissue it is not surprising that they fared

poorly in comparison to the other sets of selected features in this experiment for this purpose. These features were also barely better than random baseline, which adds support to the idea that biomarkers are task-specific. With a Gene-Set-Length of 24 on the PRAD dataset, one of the more difficult datasets, the top 5 methods were not Mutual Information, but included Logistic Regression, Linear Support Vector Classifier, and Random Forest. There is a higher standard deviation among Evaluation Models.

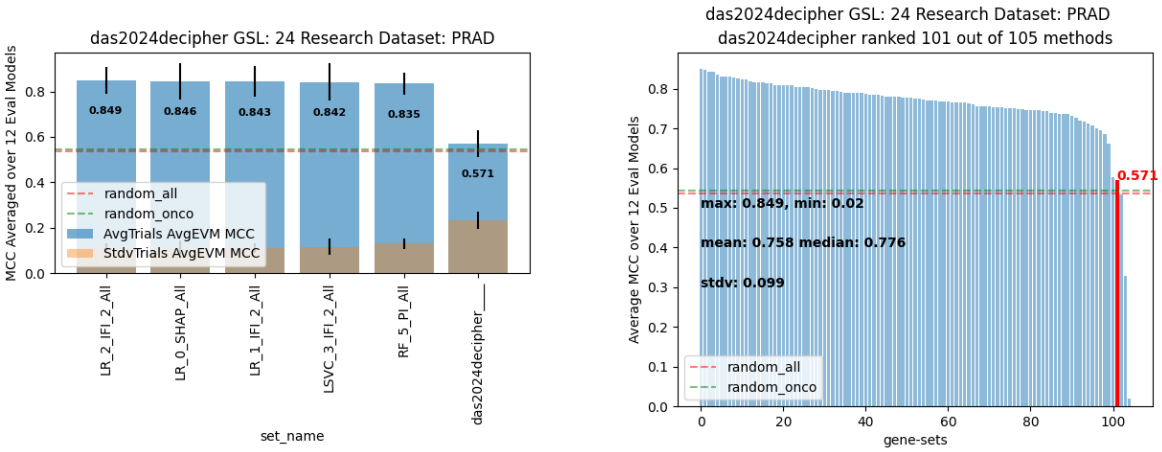

Figure 4.23: PRAD cohort: "Decipher" gene-signature for PRAD recurrence risk 24 biomarkers compared to experimental 24 feature sets. Left: Top 5 methods by MCC. Right: All methods' MCC's. Method 1: Logistic Regression 2, 0, 1 (LR_2, LR_0, LR_1) Linear Support Vector Classifier 3 (LSVC_3), Random Forest 5 (RF_5). Method 2: Inherent Feature Importance 2 (IFI_2), SHAP, Permutation Importance (PI). "_All" = unrestricted (not OncoKB only.)

Gene Set Length (GSL) 22, Kallah-Dagadu et al.

Kallah-Dagadu et al. used ML models and iML techniques with a wrapper method to select features for classification of the Breast invasive carcinoma (BRCA) dataset between tumor and normal samples, making the study very similar to this experiment. In general classification on this dataset with 22 genes was very achievable with the best of these methods

scoring an average MCC of .98, and the authors' biomarkers scoring .96 which is very close. See Appendix for Figure A.10.

Gene Set Length (GSL) 14, Peng et al.

Peng et al. selected 14 genes based on the analysis of multiple cancers, differential gene expression (DGE), and pathway analysis. The dataset used was the combination of BLCA, BRCA, COAD, HNSC, LIHC, LUAD, and LUSC. Using 14 genes on this dataset, the top 5 methods were Mutual Information or Random Forest based. The biomarkers from Peng et al. led to an average of .716 MCC, which was lower than 94 out of 104 other sets of features tested. See Figure A.11in the Appendix. These biomarkers were selected also in part based on biological considerations of the gene-pathways outside of any statistical or ML analysis.

Gene Set Length 12, Nikitina et al.

Nikitina et al. published 12 gene biomarkers for distinguishing Prostate adenocarcinoma (PRAD) from benign prostatic hyperplasia. The data used was from a different source than The Cancer Genome Atlas (TCGA), and additionally, the samples were formalin-fixed. When using these 12 biomarkers as features for predicting normal or tumor samples in the TCGA dataset, the performance was underwhelming. When using 12 features on the PRAD dataset, Point-Biserial Correlation ranked genes re-ranked by Logistic Regression 0 inherent Feature Importance, or Permutation Importance were two of the top five methods as in Appendix Figure A.12.

Gene Set Length 10, 9 Wan et al.

The genes provided by Wan et al. were selected by analyzing multiple cancer cohorts, and so to test them, each individual cohort was classified using these 10 or 9 gene-sets and the results averaged over all 16 individual cohorts. The 10 gene biomarker labeled as "NumTypes" was selected by ranking gene expressions by the number cancer cohorts they

were differentially expressed in. "AtLeastOne," is a set of 9 genes, that there was at least one patient where the gene was differentially expressed in every cancer cohort. For all 16 individual cohorts separately, on average over all cohorts, using 10 or 9 of the top ranked genes, Mutual Information based methods are 4/5 of the top five with re-ranking by SHAP or Inherent Feature Importance #2 (IFI_2). As is expected, the cohort-specific gene-sets on average are usually better than Wan et al.'s gene-sets, but these features are better than randomly chosen features. Surprisingly, the 10 or 9 biomarker genes provided by Wan et al. when applied to 16 separate cohorts on average has somewhat comparable performance to features chosen specifically for each cohort. See Figure A.13 and Figure A.14 in the Appendix section.

Gene Set Length 6 Zheng et al.

Zheng et al. published 6 novel biomarkers for Esophageal squamous cell carcinoma (ESCC) using common Differential Gene Expression between datasets, and hub-gene analysis. The Cancer Genome Atlas dataset of Esophageal carcinoma (ESCA) was used for training and testing the evaluation models using these six genes, or six genes chosen by the other methods in this experiment.

ESCA was among the most difficult cohorts to classify on average. When using only 6 genes, the standard deviation over trials and between Evaluation Models is higher. Two of the top five methods are OncoKB-only gene-sets. Logistic Regression and Linear Support Vector Classifier are also present among Method-1s. There is a greater difference between selected genes and random genes for classification. The poorer performance of Zheng et al.'s set may be on account of the difference of intended cancer type ESCA versus ESCC. Though Zheng et al.'s set is still significantly better than random as seen in the Appendix Figure A.15.

Gene Set Length 5,3,2 Coleto-Alcudia et al.

Coleto-Alcudia et al. employed a "multi-objective optimization, using an Artificial Bee Colony based on Dominance (ABCD) algorithm" to select biomarkers. The objectives were minimizing the number of features and maximizing classification accuracy using an implementation of Support Vector Classifier. From this study biomarkers for the following cohorts were tested: LUNG, the combination of Lung adenocarcinoma (LUAD) and Lung squamous cell carcinoma (LUSC): 5 genes. Breast invasive carcinoma (BRCA) with 4 genes, but one was missing from this The Cancer Genome Atlas data, as Coleto-Alcudia et al. started with a BRCA dataset containing 60,000 gene expressions. Prostate adenocarcinoma (PRAD), with 2 genes. The ABCD algorithm was also used with a combination dataset of BRCA, LIHC, ESCA, TGCT, THCA, LUAD, and LUSC, resulting in 5 genes.

When using 5 genes to predict the LUNG dataset, a combination of LUAD and LUSC, the top 5 methods included Random Forest ids 6 and 7, Mutual Information, with two Method-2 Permutation Importance re-rankings, shown in Appendix Figure A.16.

When using only 4 genes to classify the dataset of 7 combined cohorts, Random Forest is most common as 4/5 of the top five, with one Mutual Information based Method 1. High classification performance was still possible. The researcher's set does poorly relatively in this experimental context, which may point to differences in data sources being crucial, but still better than random as shown in Appendix Figure A.17.

Using only 3 genes to classify the Breast invasive carcinoma (BRCA) dataset, very high classification performance is still possible. Coleto-Alcudia et al. provided a list of 4 genes, one of which was not present in this TCGA data. Despite this, surprisingly, owing to potential lost feature interaction in ML models, the 3 out of 4 genes provided by the researchers were still good predictors. Interestingly, the first 3 features of Kallah-Dagadu et al.'s 22 iML ranked gene-set were among the top 5 gene-sets, the other 4 were Random Forest derived shown in Figure 4.24.

Using only 2 genes to predict Prostate adenocarcinoma (PRAD,) one of the more difficult

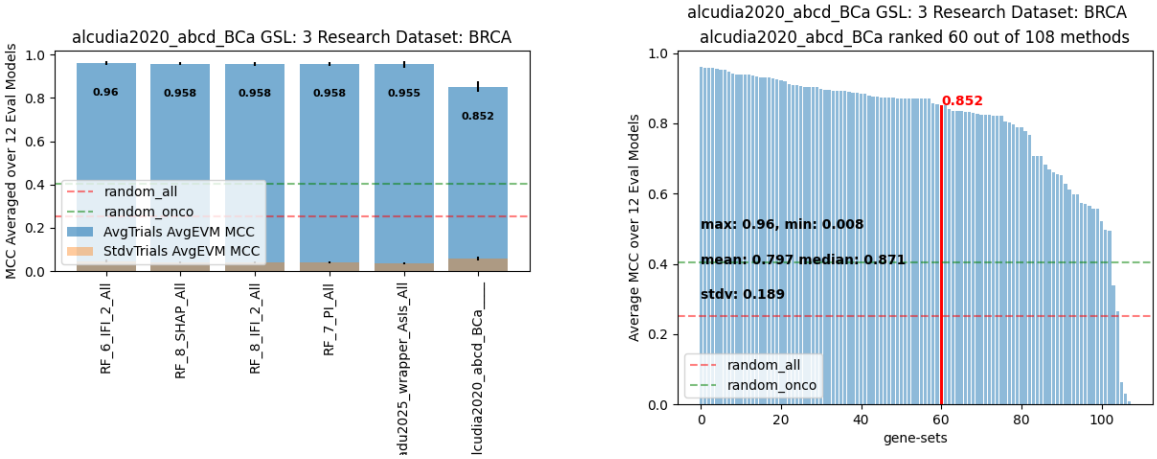

Figure 4.24: BRCA: Coleto-Alcudia et al. ABCD 3 biomarkers versus experimental 3 feature sets. Left: Top five methods by MCC. Right: All methods' MCCs. Method 1: Random Forest 6, 7, 8 (RF_6, RF_8, RF_7) Kallah-Dagadu et al. 22-gene iML wrapper biomarkers. Method 2: Inherent Feature Importance 2 (IFI_2), SHapley Additive exPlanations (SHAP), Permutation Importance (PI). "_All" = unrestricted (not OncoKB only).

datasets was the most challenging configuration. There is a high standard deviation between Evaluation Models. The top 2 genes by Point-Biserial Correlation without any re-ranking was the 3rd best model. The top models performed much higher than random, but the two genes, *C14orf72* and *CSRP2*, arrived at by Coleto-Alcudia et al.'s ABCD algorithm for the PRAD dataset were almost on par with randomly selected genes in this experiment. Figure 4.25 shows the relative performance. *DLX2* and *APOBEC3C* were the best two genes on average for classification, which were the result of Random Forest 7 (RF_7) initial feature selection, and then SHAP re-ranking. With Permutation Importance re-ranking, *MED21* is the 2nd gene, and *APOBEC3C* came third. *EPHA10* and *DLX1* were the top two by Point-Biserial Correlation without any iML ranking. Random Forest 8 (RF_8) re-ordered by SHAP and Random Forest Mean Decrease Impurity again both arrived at *DLX2*, *EPHA10*, *MED21*, and *APOBEC3C* as the top 4. *DLX2* has been noted as significant in multiple types of cancer [67, 97, 42]. *APOBEC3C* was associated with Pancreatic Cancer, and other

cancer types, in published studies [74, 90, 64, 25].

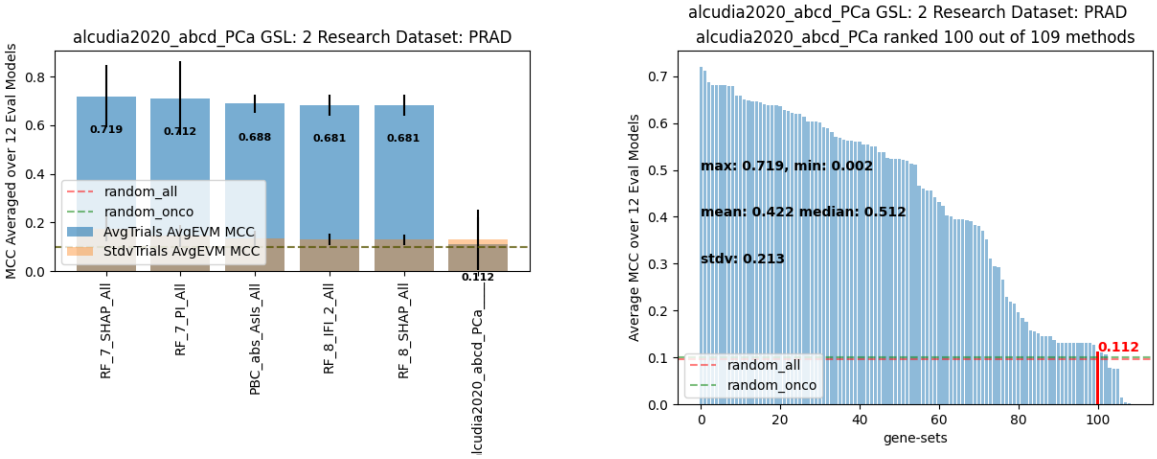

Figure 4.25: PRAD: Coleto-Alcudia et al. ABCD 2 biomarkers versus experimental 2 feature sets. Left: Top five methods by MCC. Right: All methods' MCCs. Method 1: Random Forest 7, 8 (RF_7, RF_8), Point-Biserial Correlation (PBC_abs). Method 2: SHapley Additive exPlanations (SHAP), Permutation Importance (PI), "AsIs" (IFI_1 original order), Inherent Feature Importance 2 (IFI_2). "_All" = unrestricted (not OncoKB only).

4.4 Step 4: Potential Biomarker Evaluation

Using only OncoKB genes to select feature sets was not found to be useful, but randomly selected OncoKB genes were slightly better than unrestricted random genes for classification with small Gene-Set-Lengths (Between 26 and 2 tested). It is possible that some methods more commonly select these known cancer genes and in combination with other features obtain higher performance. For Method-1 (selecting 250 genes), Random Forest based methods, Mutual Information, and Point-Biserial Correlation selected more than the expected random baseline, while the other methods did not. This equally applies to OncoKB "onco-genes," and tumor suppressor genes, see Appendix Figure A.7.

It is also of interest if OncoKB-genes ranked highly, and Method-2 only re-ranked the 250 genes selected by Method-1 and did not select different genes. Using the top 12 ranked genes by each method, as 12 is midway between 26 and 2, the pattern is similar but less than with the 250 gene-set relative to the expected random baseline. See Figure 4.26.

When averaging all Method-1s together and examining the OncoKB-gene count by Method-2 re-ranking for the top 12, the original order of Method 1 ranks (IFI_1) contains the most OncoKB genes. Where as re-ranking slightly drops the OncoKB content, with Permutation Importance having the largest drop. There is only a slight correlation between performance of a method and the number of OncoKB genes in the top 12 as in Figure 4.27.

An alternative method for evaluating lists of biomarkers involved using the Kyoto Encyclopedia of Genes and Genomes (KEGG) for additional gene annotations. The KEGGRESTpy package [43] was used to query the KEGG database for all 20,530 genes, using Entrez ID or gene aliases when necessary. Besides pathways the gene belongs to, and diseases the gene is noted as being associated with, a keyword search of all the results was used to check for any mention of relationship with cancer. Out of 20,530 TCGA data genes, 18,922 genes were successfully queried from KEGG. Out of these, 2285 gene full annotations contained at least one cancer related key word. There was no significant correlation between method performance and genes with at least one key word in their annotations in the top

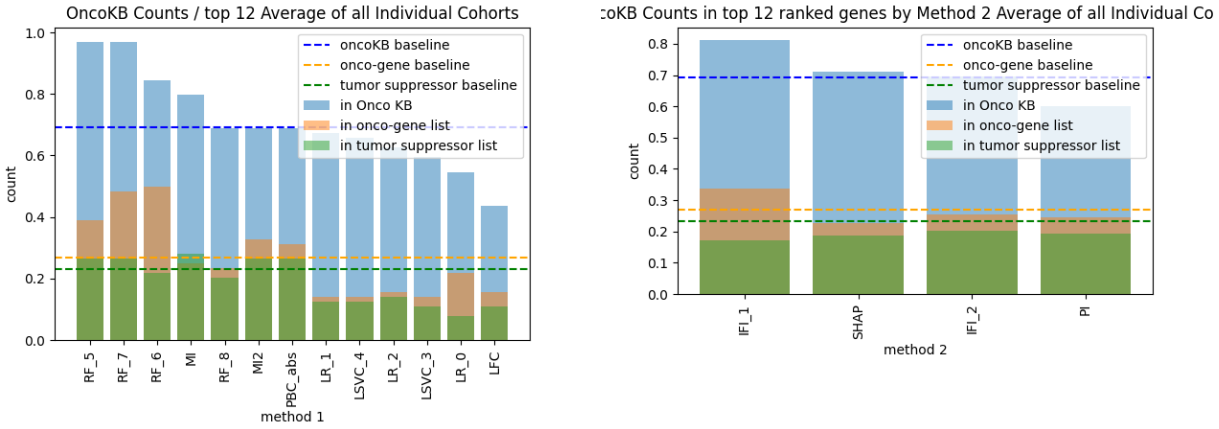

Figure 4.26: On average over all individual cohorts: Left: Number of OncoKB genes in the top 12 ranked features, organized by Method 1 and average of Method-2s. Right: Organized by Method 2 and average of Method-1s. Random Forest (RF), Mutual Information (MI), Point-Biserial Correlation (PBC_abs), Logistic Regression (LR), Linear Support Vector Classifier (LSVC), Log Fold Change (LFC). Inherent Feature Importance 1 original order (IFI_1), SHapley Additive exPlanations (SHAP), Inherent Feature Importance 2 (IFI_2), Permutation Importance (PI).

12. See Figure 4.27.

For each cohort, 52 gene rankings of all TCGA genes were created (13 Method-1s and 4 Method-2s, OncoKB-only gene-sets are excluded.) Each gene was scored by counting the number of the 52 ranking systems where the gene was in the top 12 ranked genes, modified by the average Evaluation Method classification performance with Gene-Set-Length between 26 and 2. The top 10 genes for each cohort by this scoring system are displayed in Figure A.3. KEGG annotations contain known pathways and associated diseases for particular genes. Some KEGG annotations are sparse with no known pathways or diseases provided, and a certain number of TCGA genes could not be queried from KEGG at all. For the top 10 scored genes for each cohort the number of times a pathway or disease was in each gene's annotations was counted. The top 10 pathways are displayed by count, and when counts are tied, the gene's scoring rank is used to determine which pathways to display in Table

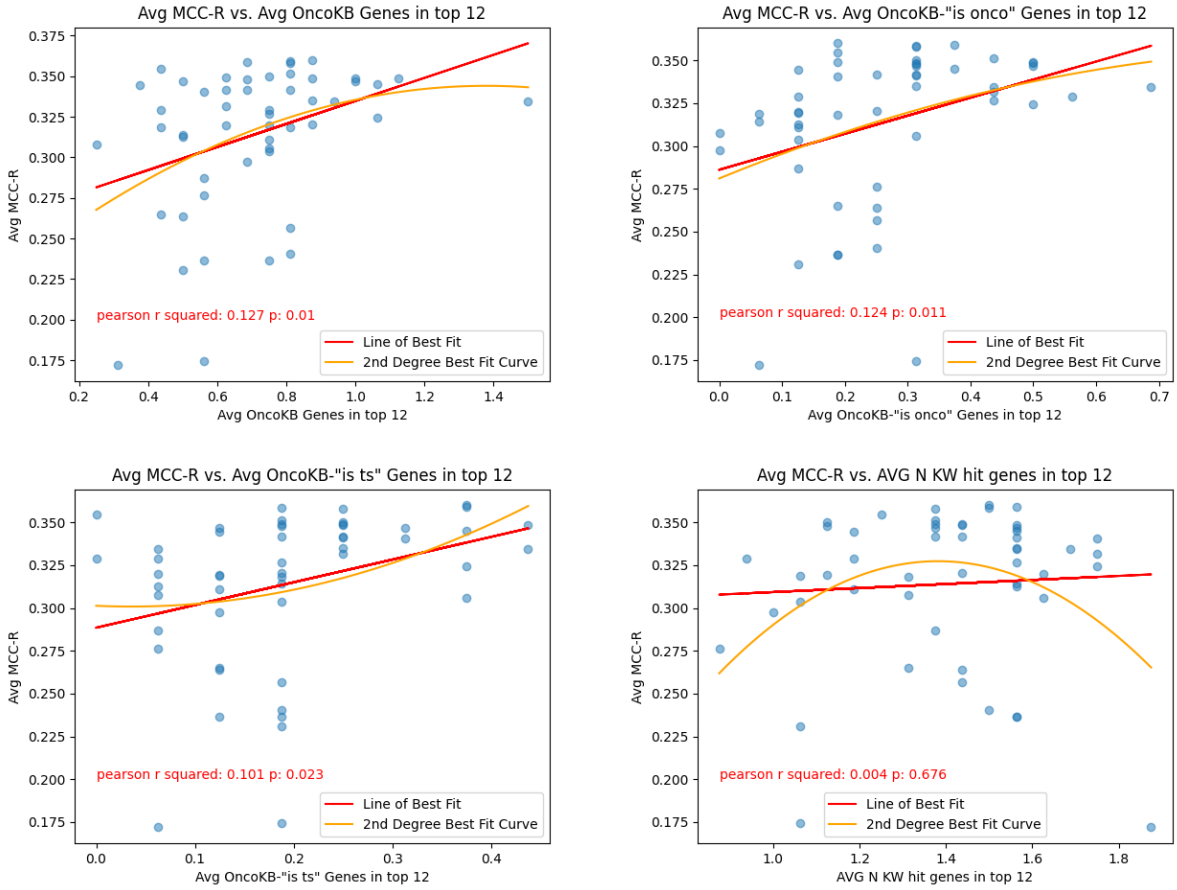

Figure 4.27: Average of individual cohorts and Gene-set-lengths 26 through 2 MCC-R. Top Left: N OncoKB genes in top 12, Top Right N OncoKB "onco-genes" in Top 12, Bottom Left: N OncoKB tumor suppressor genes, Bottom Right: N genes with at least one cancer keyword in KEGG query annotations. OncoKB genes in the top 12 ranked genes are slightly positively correlated with better MCC-R, but key word hit genes are not.

A.3. The same system was used to display the top 5 associated diseases. Often times the diseases or pathways seem unrelated to cancer or to the region or organ of the cancer type. A single gene can belong to multiple pathways, and the annotations are not limited to gene expression only, but also include other types of genomic analysis. The search for cancer-related keywords and membership in the OncoKB database were not significantly correlated

with the scoring and selection of the top 10 genes per cohort.

4.5 Discussion

iML techniques for feature importance and feature selection are effective in choosing a small list of features from TCGA data. The features are gene expressions leading to possible biomarkers in cancer data. It has been shown that for classification, a small list of features are effective for classification of tumor or normal tissue. Many of these techniques had similar performance, though performance varies across different datasets and with different gene-set lengths. The lists of biomarkers generated by these methods were often superior to biomarkers from other sources for this classification task. With small Gene-Set-Length, Mutual Information estimates (MI) were among the best selection methods for biomarkers initially. Using iML techniques to further refine lists of length 250 was useful, but the improvement in performance is slight.

The selected datasets are often challenging for feature selection because many possible combinations of features lead to similar high performance. However, most proposed methods did consistently better in classification performance than random selection of genes. Re-ranking features with SHAP did not improve the performance of small sets as much as expected. Permutation importance sometimes returned negligible or even zero re-ranking information. iML methods did outperform statistical methods, except Mutual Information, as was expected.

What was surprising was that the Mutual Information score was one of the best selection/ranking systems, as far as evaluated by small gene sets for classification. Mutual Information, here estimated by scikit-learn's mutual_inf_classif [11], has been noted as a feature selection method in other studies [70, 77, 84]. This estimation of Mutual Information was done one gene at a time, so does not account for feature interactions. Part of the purpose of this experiment with iML techniques was that ML models use feature interactions for predictions, while statistical measures like Log Fold Change (LFC) for differential Gene Expression (DGE) do not. For this data and use case, the Mutual Information estimate

was usually superior in general to both Log Fold Change, or Point-Biserial Correlation for selecting the best genes as features for classification. Since Mutual Information measures potentially non-linear dependence, it may be better for choosing ML features than other statistical methods. Log Fold Change constrained by a t-test may also omit features which would be useful for classification, but do not meet statistical significance tests. Furthermore, Log Fold Change ranks genes by the magnitude of the ratio of mean gene expression across sample groups, when this may not be the single most informative criteria for a feature in ML classification.

Despite Mutual Information being on average the best of Method-1s, after the initial selection of 250 genes based on Mutual Information estimate, the best top features between 26 and 2 top features were those further narrowed down with iML techniques. This shows, that as expected, accounting for feature interaction is important for assigning feature importance in ML models, and offers an improvement in feature selection. What was further interesting was that Mutual Information estimates and Random Forest Mean Decrease Impurity (MDI) feature importance selected often the same features. Furthermore, after Mutual Information, Random Forests MDI were often among the next best Method-1 250 gene selecting algorithms. (When like-wise evaluated using only the top 26 through 2 ranked features.) This points perhaps to how the underlying Random Forest algorithm relates to the information content of the features.

Comparison of means across groups using the absolute difference or Log Fold Change, controlled by statistical significance thresholds does not often select features that lead to good relative performance. Feature importance from iML methods often provided very different rankings of feature importance depending on the methods employed. However, the overlap between methods gives multiple perspectives and support for a given feature importance. It is difficult to determine one best approach, because the results vary based on classification model, dataset, and number of genes. However, Mutual Information with iML re-ranking rated best of all tested methods on average of experiment parameters.

Implications

In the search to understand the mechanisms of cancer, and to develop new treatments, gene expression biomarkers are an established and potentially useful tool. Given that there are many possible specific genes and their expressions to study in the case of specific cancers and aspects of those cancers, narrowing down the list could aid research. The most basic approach is to study differential gene expression through a statistical lens, and measure the fold change. While this approach has scientific validity, and has already been used to good effect, it is possible that some potentially useful biomarkers could be overlooked by these methods. Researchers are experimenting with a variety of other gene expression biomarker identification methods, including the use of Machine Learning, iML, and incorporating biological knowledge. A way to test and compare methods could help to determine which are most useful, and which resulting biomarkers are most worth investing the effort of more detailed study into.

The best features for ML models may not translate into the most important genes to study, but neither do statistically differently expressed genes necessarily. See Appendix "BRCA Top Ranked Genes" for an example analysis of biomarkers identified in this experiment.

Researching all top genes for all datasets is beyond the scope of this paper, but the number of publications concerning gene biomarkers highlights the need for a more systematized approach toward already published putative biomarkers. Only *MMP11*, and *FIGF* had specific cancer annotations in KEGG. Arriving at the same gene expression biomarkers across various methods offers better evidence for considering the expression of this gene significant. If the highest ranked genes returned by a method are already confirmed biomarkers, then the search can be expanded to lower ranked and possibly as of yet less studied genes.

iML techniques lead to selecting better features for classification on average in these tested datasets of 16 TCGA cancer cohorts individually, and 5 combinations of cohorts. Additionally, Mutual Information presented itself as a valuable initial feature selection technique.

The use of these features in ML classification and their relation to biological significance has been offered some support, but not adequately enough yet to make stronger conclusions.

Challenges

With these datasets there is often small difference in classification performance between different sets of features, which means there are a multitude of features that are useful for classification, and no clear single best set of features, either by statistical or iML methods. One benefit of using multiple ranking systems is to see which features or genes do reappear in the gene sets generated by different methods, that share a similar high classification performance. Since gene sets are assembled by very different criteria and algorithms, if a gene appears in diverse highly performing sets, this indicates a high likelihood of being a good biomarker. If a method selects known biomarkers, it is possible the method can be used to find novel biomarkers which share similar criteria with the selected genes.

Interpreting and presenting the results of these experiments was challenging. There were many dimensions to the experiment with 21 datasets, 13 initial ranking methods, 4 secondary ranking methods, 2 feature-pools (unrestricted or OncoKB), and multiple 12 small Gene-Set-Length tests. Analyzing all of the results is difficult across all dimensions. By narrowing the scope of the experiment, more specific and interesting findings might be possible. Summaries and averages were necessary to analyze and present the results, which may cover up some interactions or patterns.

The TCGA data obtained was imbalanced, with significantly more Primary Tumor samples, than Normal Tissue Samples. Additional Normal Tissue samples could help better evaluate classification performance. Additionally, lack of Solid Tissue Normal samples limited the amount of cohorts which could be used, and also make further separating the data into validation and hold-out test sets less feasible. The data was also usually very separable between the two classes, as see in dimensionality reduction visualizations, and generally very high classification accuracy. Attempting a more specific type of classification than merely normal tissue versus tumor, such as patient response to treatment, or survival analysis could

be beneficial for finding more specific biomarkers to study.

Future Work

From this experiment, several avenues for future work are apparent. With more computational resources, the gene-set selection techniques could incorporate model agnostic methods before an initial feature selection. This means more complex models without inherent feature importance methods could also be used in the initial ranking/selection steps, as Step 1 using 20,530 genes was in the experiment. Training each of the 9 models in Step 2 on every list of 250 genes, and including random lists at this step, would greatly increase computation needs and complexity, but would provide a better evaluation of feature selection methods. A version of the experiment could be repeated incorporating non-TCGA sourced data, and with enough data, further splitting the data into validation and hold-out test sets. The hold-out test sets could be used for final evaluation and give a better, less biased, appraisal of the feature sets for ML classification.

An analysis of PANCAN_selected dataset results was omitted, but findings from combination of multiple cancer type datasets may be useful for finding biomarkers that apply to cancer in general. A more focused analysis of PANCAN or any one dataset would be useful for further testing biomarker identification using these approaches. The drawbacks of using one dataset are that results cannot necessarily be generalized to other datasets. The advantages of using only one dataset means less variables in the experiment and more feasible in-depth analysis. It is not feasible to present detailed analysis of top genes from many datasets in depth in one presentation. A better automated approach for detecting already known or theorized RNA biomarkers in cancer is needed. As some KEGGRESTpy queries failed, and the cancer keyword search method is not as precise or meaningful as could be hoped. Furthermore, a manual search in peer-reviewed articles returned many publications relating to a gene's relation to cancer, even if the automated queries did not return results with a cancer keyword.

Chapter 5

CONCLUSION

In this experiment we aimed to compare methods for identifying gene expression biomarkers for classifying Tumor versus Normal tissue samples for multiple cancer types. We first trained nine inherently interpretable machine learning models to classify Tumor or Normal Tissue using all 20,530 genes on 16 TCGA cohorts. Per each cohort we also ranked all genes based on Mutual Information estimation and statistical tests, Point-Biserial Correlation, and Log Fold Change. Then we retrained the same models using only the top 250 out of 20,530 genes, using each ranking system. We then collected SHAP, Permutation Importance, and inherent feature importance scores again. These scores were used to re-rank the lists of top 250 genes. We tested all final ranking methods by refining the lists to between 26 and 2 of the highest ranked genes and training new evaluation models. We compared the average Evaluation Model classification performance between selected feature sets. Finally, we used the OncoKB and KEGG databases to assess what relationship the genes chosen by these methods had with known biomarkers, and applied pathway analysis to map the gene's biological significance.

Some of the tested methods lead to better Evaluation Model classification than the biomarkers from literature. Log Fold Change, constrained by Welch's t-test, representing statistical analysis for biomarker selection, was among the least effective selection methods tested. Mutual Information estimation was among the best methods, as was Random Forest Mean Decrease Impurity (MDI) feature importance. Mutual Information and Random Forest MDI often chose similar features, more so than Mutual Information and Logistic Regression and Linear Support Vector Classifier, but not as similar as other more closely related methods. Adjusting the scikit-learn Mutual Information estimation algorithm hyperparameter

of *n_neighbors* made very little difference on average, but with the default 3 being slightly superior to 5. Using SHAP or another round of inherent model specific feature importance improved refinement of small sets of biomarkers for classification. Permutation Importance was less often helpful. Restricting gene-set selection to known cancer-genes from OncoKB did not help performance, but hurt it when using 250 features. Using the top 250 genes instead of all 20,530 leads to better performance with the tested inherently interpretable models.

On average, Mutual Information, or Mutual Information in combination with iML techniques led to better classification results than many iML alone methods. Point-Biserial Correlation was in the middle of tested methods. Logistic Regression and Linear Support Vector Classifier coefficients were the best methods for selecting a larger number of features (250) for those same models to use as features for classification. However, using those ranking systems with smaller gene-set lists was generally worse than Mutual Information, Random Forest, and Point-Biserial Correlation. Methods were ranked based on average classification performance over all datasets and gene-set-lengths tested, but the best methods varied by specific dataset and number of features used. Evaluation with biomedical annotations was less conclusive, showing only a slight correlation with classification power and genes in the OncoKB database.

The objective of selecting biomarkers is to narrow down a list of genes for further research that may be causally related to cancer via their expression, or may help identify therapeutic targets. Genes that are selected as important features by various methods were commonly overlapping, from similar pathways, and contained known biomarkers. This supports the idea that these methods may be helpful in identifying novel biomarkers. iML techniques as well as iML in combination with Mutual Information and Point-Biserial Correlation are more effective than statistical differential expression methods in selecting possible gene expression biomarkers for classification from TCGA RNA Seq cancer datasets. The future work suggested is experimenting with further iML techniques, ML models, and a more finely tuned evaluation process. Specific in-depth analysis of individual cancer type datasets and

selected genes is necessary. Then, a comparison between results from different datasets can be conducted.

BIBLIOGRAPHY

[1] 1.4. Support Vector Machines -scikit-learn. Available online at: https://scikit-learn.org/stable/modules/svm.html, last accessed on 12.07.2025.

[2] bioinfokit documentation - Renesh Bedre. Available online at: https://reneshbedre.github.io/blog/howtoinstall.html, last accessed on 08.08.2025.

[3] Genome.gov ribonucleic Acid (RNA). Available online at: https://www.genome.gov/genetics-glossary/Ribonucleic-Acid-RN, last accessed on 08.06.2025.

[4] Home - GEO - NCBI. Available online at: https://www.ncbi.nlm.nih.gov/geo, last accessed on 08.08.2025.

[5] LinearSVC. Available online at: https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC, last accessed on 12.07.2025.

[6] LogisticRegression. Available online at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html, last accessed on 12.11.2025.

[7] mygene: Python Client for MyGene.Info services. Available online at: https://github.com/biothings/mygene.py, last accessed on 08.05.2025.

[8] NCBI NIH Gene. Available online at: https://www.ncbi.nlm.nih.gov/datasets/gene/, last accessed on 08.08.2025.

[9] OncoKB™ - MSK's Precision Oncology Knowledge Base. Available online at: https://www.oncokb.org/, last accessed on 08.03.2025.

[10] PC datasources. Available online at: https://www.pathwaycommons.org/pc2/datasources, last accessed on 12.09.2025.

[11] scikit-learn mutual_info_classif. Available online at: https://scikit-learn/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html, last accessed on 08.03.2025.

[12] SHAP documentation — SHAP latest documentation. Available online at: `https://shap.readthedocs.io/en/latest/`, last accessed on 08.06.2025.

[13] TCGA Study Abbreviations | NCI Genomic Data Commons.

[14] UCSC Xena. Available online at: `https://xenabrowser.net/`, last accessed on 08.03.2025.

[15] Definition of RNA - NCI Dictionary of Cancer Terms - NCI, February 2011. Archive Location: nciglobal,ncienterprise.

[16] The Cancer Genome Atlas Program (TCGA) - NCI, May 2022. Archive Location: nciglobal,ncienterprise.

[17] Abhineet Agarwal, Ana M. Kenney, Yan Shuo Tan, Tiffany M. Tang, and Bin Yu. Integrating Random Forests and Generalized Linear Models for Improved Accuracy and Interpretability, May 2025. arXiv:2307.01932 [stat].

[18] Qasem Al-Tashi, Maliazurina B. Saad, Amgad Muneer, Rizwan Qureshi, Seyedali Mirjalili, Ajay Sheshadri, Xiuning Le, Natalie I. Vokes, Jianjun Zhang, Jia Wu, Qasem Al-Tashi, Maliazurina B. Saad, Amgad Muneer, Rizwan Qureshi, Seyedali Mirjalili, Ajay Sheshadri, Xiuning Le, Natalie I. Vokes, Jianjun Zhang, and Jia Wu. Machine Learning Models for the Identification of Prognostic and Predictive Cancer Biomarkers: A Systematic Review. *International Journal of Molecular Sciences*, 24(9), April 2023.

[19] Fadi Alharbi and Aleksandar Vakanski. Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review. *Bioengineering*, 10(2):173, February 2023. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[20] Dominik Buschmann, Anna Haberberger, Benedikt Kirchner, Melanie Spornraft, Irmgard Riedmaier, Gustav Schelling, and Michael W. Pfaffl. Toward reliable biomarker signatures in the age of liquid biopsies - how to standardize the small RNA-Seq workflow. *Nucleic Acids Research*, 44(13):5995–6018, July 2016.

[21] Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E. Rudolph, Rona Yaeger, Tara Soumerai, Moriah H. Nissan, Matthew T. Chang, Sarat Chandarlapaty, Tiffany A. Traina, Paul K. Paik, Alan L. Ho, Feras M. Hantash, Andrew Grupe, Shrujal S. Baxi, Margaret K. Callahan, Alexandra Snyder, Ping Chi, Daniel C. Danila, Mrinal Gounder, James J. Harding, Matthew D. Hellmann, Gopa Iyer, Yelena Y. Janjigian, Thomas Kaley, Douglas A. Levine, Maeve Lowery, Antonio Omuro, Michael A. Postow, Dana Rathkopf, Alexander N. Shoushtari,

Neerav Shukla, Martin H. Voss, Ederlinda Paraiso, Ahmet Zehir, Michael F. Berger, Barry S. Taylor, Leonard B. Saltz, Gregory J. Riely, Marc Ladanyi, David M. Hyman, José Baselga, Paul Sabbatini, David B. Solit, and Nikolaus Schultz. OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, (1):1–16, May 2017. Publisher: Wolters Kluwer.

[22] Yin-Wen Chang and Chih-Jen Lin. Feature Ranking Using Linear SVM. In *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*, pages 53–64. PMLR, December 2008. ISSN: 1938-7228.

[23] H Charu Meena, R Sagaya Jansi, S Aishwarya, Shanmugaraj Balamurugan, and Panthagani Praveen Kumar. Exploring Breast Cancer-Associated Genes: A Comprehensive Analysis and Competitive Endogenous RNA Network Construction. *Archives of Razi Institute*, 80(2):347–360, April 2025.

[24] Joe W. Chen and Joseph Dhahbi. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *Scientific Reports*, 11(1):13323, June 2021. Publisher: Nature Publishing Group.

[25] Zhishan Chen, Wanqing Wen, Jiandong Bao, Krystle L. Kuhs, Qiuyin Cai, Jirong Long, Xiao-ou Shu, Wei Zheng, and Xingyi Guo. Integrative genomic analyses of APOBEC-mutational signature, expression and germline deletion of APOBEC3 genes, and immunogenicity in multiple cancer types. *BMC Medical Genomics*, 12(1):131, September 2019.

[26] Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6, January 2020.

[27] Davide Chicco and Giuseppe Jurman. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, 16(1):4, February 2023.

[28] Veredas Coleto-Alcudia and Miguel A. Vega-Rodríguez. A multi-objective optimization approach for the identification of cancer biomarkers from RNA-seq data. *Expert Systems with Applications*, 193:116480, May 2022.

[29] David Crosby, Sangeeta Bhatia, Kevin M. Brindle, Lisa M. Coussens, Caroline Dive, Mark Emberton, Sadik Esener, Rebecca C. Fitzgerald, Sanjiv S. Gambhir, Peter Kuhn, Timothy R. Rebbeck, and Shankar Balasubramanian. Early detection of cancer. *Science*, 375(6586):eaay9040, March 2022. Publisher: American Association for the Advancement of Science.

[30] Paweł Czyż, Frederic Grabowski, Julia E. Vogt, Niko Beerenwinkel, and Alexander Marx. Beyond Normal: On the Evaluation of Mutual Information Estimators, October 2023. arXiv:2306.11078 [stat].

[31] Huijuan Dai, Wenting Xu, Lulu Wang, Xiao Li, Xiaonan Sheng, Lei Zhu, Ye Li, Xinrui Dong, Weihang Zhou, Chenyu Han, Yan Mao, and Linli Yao. Loss of SPRY2 contributes to cancer-associated fibroblasts activation and promotes breast cancer development. *Breast Cancer Research*, 25(1):90, July 2023.

[32] Pijush Das, Anirban Roychowdhury, Subhadeep Das, Susanta Roychoudhury, and Sucheta Tripathy. sigFeature: Novel Significant Feature Selection Method for Classification of Gene Expression Data Using Support Vector Machine and t Statistic. *Frontiers in Genetics*, 11, April 2020. Publisher: Frontiers.

[33] Sreyashi Das, Mohan Kumar Dey, Ram Devireddy, and Manas Ranjan Gartia. Biomarkers in Cancer Detection, Diagnosis, and Prognosis. *Sensors*, 24(1):37, January 2024. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[34] Elisa Díaz de la Guardia-Bolívar, Juan Emilio Martínez Manjón, David Pérez-Filgueiras, Igor Zwir, and Coral del Val. Explainable Machine Learning Models Using Robust Cancer Biomarkers Identification from Paired Differential Gene Expression. *International Journal of Molecular Sciences*, 25(22):12419, January 2024. Number: 22 Publisher: Multidisciplinary Digital Publishing Institute.

[35] Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British Journal of Cancer*, 124(4):686–696, February 2021.

[36] Dina Mohamed Ahmed Samir Elkahwagy, Caroline Joseph Kiriacos, and Manar Mansour. Logistic regression and other statistical tools in diagnostic biomarker studies. *Clinical and Translational Oncology*, 26(9):2172–2180, September 2024.

[37] Marina Elkommos-Zakhary, Neeraja Rajesh, and Vladimir Beljanski. Exosome RNA Sequencing as a Tool in the Search for Cancer Biomarkers. *Non-Coding RNA*, 8(6):75, December 2022. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[38] Fiona Katharina Ewald, Ludwig Bothmann, Marvin N. Wright, Bernd Bischl, Giuseppe Casalicchio, and Gunnar König. A Guide to Feature Importance Methods for Scientific Inference. volume 2154, pages 440–464. 2024. arXiv:2404.12862 [stat].

[39] Zhiyu Fan, Yingli Chen, Dongsheng Yan, and Qianzhong Li. Effects of Differentially Methylated CpG Sites in Enhancer and Promoter Regions on the Chromatin Structures of Target LncRNAs in Breast Cancer. *International Journal of Molecular Sciences*, 25(20):11048, January 2024. Publisher: Multidisciplinary Digital Publishing Institute.

[40] Miguel A. García-Campos, Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. Pathway Analysis: State of the Art. *Frontiers in Physiology*, 6, December 2015. Publisher: Frontiers.

[41] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science (New York, N.Y.)*, 286(5439):531–537, October 1999.

[42] William JF Green, Graham Ball, Geoffrey Hulman, Catherine Johnson, Gerry Van Schalwyk, Hari L. Ratan, Daniel Soria, Jonathan M. Garibaldi, Richard Parkinson, Joshua Hulman, Robert Rees, and Desmond G. Powe. KI67 and DLX2 predict increased risk of metastasis formation in prostate cancer–a targeted molecular approach. *British Journal of Cancer*, 115(2):236–242, July 2016. Publisher: Nature Publishing Group.

[43] Kai Guo. guokai8/KEGGRESTpy, October 2025. original-date: 2024-11-11T02:32:00Z.

[44] Yansong Han and Yuexia Li. Comprehensive Exploration of M2 Macrophages and Its Related Genes for Predicting Clinical Outcomes and Drug Sensitivity in Lung Squamous Cell Carcinoma. *Journal of Oncology*, 2022(1):1163924, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/1163924.

[45] Erik Hilborn, Olle Stål, and Agneta Jansson. Estrogen and androgen-converting enzymes 17-hydroxysteroid dehydrogenase and their involvement in cancer: with a special focus on 17-hydroxysteroid dehydrogenase type 1, 2, and breast cancer. *Oncotarget*, 8(18):30552–30562, February 2017.

[46] Zeenia Jagga and Dinesh Gupta. Machine Learning for Biomarker Identification in Cancer Research – Developments Toward its Clinical Application. *Personalized Medicine*, 12(4):371–387, August 2015.

[47] Neil K. Jairath, Alan Dal Pra, Randy Vince, Robert T. Dess, William C. Jackson, Jeffrey J. Tosoian, Sean M. McBride, Shuang G. Zhao, Alejandro Berlin, Brandon A. Mahal, Amar U. Kishan, Robert B. Den, Stephen J. Freedland, Simpa S. Salami, Samuel D. Kaffenberger, Alan Pollack, Phuoc Tran, Rohit Mehra, Todd M. Morgan,

Adam B. Weiner, Osama Mohamad, Peter R. Carroll, Matthew R. Cooperberg, R. Jeffrey Karnes, Paul L. Nguyen, Jeff M. Michalski, Jonathan D. Tward, Felix Y. Feng, Edward M. Schaeffer, and Daniel E. Spratt. A Systematic Review of the Evidence for the Decipher Genomic Classifier in Prostate Cancer. *European Urology*, 79(3):374–383, March 2021.

[48] Mohd Haris Jamal, Pratyush Porel, and Khadga Raj Aran. Emerging biomarkers for pancreatic cancer: from early detection to personalized therapy. *Clinical and Translational Oncology*, May 2025.

[49] Marta Jardanowska-Kotuniak, Michał Dramiński, Michal Wlasnowolski, Marcin Łapiński, Kaustav Sengupta, Abhishek Agarwal, Adam Filip, Nimisha Ghosh, Vera Pancaldi, Marcin Grynberg, Indrajit Saha, Dariusz Plewczynski, and Michał J. Dabrowski. Unveiling Epigenetic Regulatory Elements Associated with Breast Cancer Development. *International Journal of Molecular Sciences*, 26(14):6558, January 2025. Publisher: Multidisciplinary Digital Publishing Institute.

[50] Gabriel Kallah-Dagadu, Mohanad Mohammed, Justine B. Nasejje, Nobuhle Nokubonga Mchunu, Halima S. Twabi, Jesca Mercy Batidzirai, Geoffrey Chiyuzga Singini, Portia Nevhungoni, and Innocent Maposa. Breast cancer prediction based on gene expression data using interpretable machine learning techniques. *Scientific Reports*, 15(1):7594, March 2025. Publisher: Nature Publishing Group.

[51] Hala Fawzy Mohamed Kamel and Hiba Saeed A. Bagader Al-Amodi. Exploitation of Gene Expression and Cancer Biomarkers in Paving the Path to Era of Personalized Medicine. *Genomics, Proteomics & Bioinformatics*, 15(4):220–235, August 2017.

[52] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Yuriko Matsuura, and Mari Ishiguro-Watanabe. KEGG: biological systems database as a model of the real world. *Nucleic Acids Research*, 53(D1):D672–D677, January 2025.

[53] Shamika Ketkar, Lindsay C. Burrage, and Brendan Lee. RNA Sequencing as a Diagnostic Tool. *JAMA*, 329(1):85–86, January 2023.

[54] Javed Khan, Jun S. Wei, Markus Ringnér, Lao H. Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R. Antonescu, Carsten Peterson, and Paul S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6):673–679, June 2001.

[55] Clarissa M. Koch, Stephen F. Chiu, Mahzad Akbarpour, Ankit Bharat, Karen M. Ridge, Elizabeth T. Bartom, and Deborah R. Winter. A Beginner's Guide to Analysis of RNA Sequencing Data. *American Journal of Respiratory Cell and Molecular Biology*, 59(2):145–157, August 2018.

[56] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, June 2004.

[57] S. Kumar, A. Mohan, and R. Guleria. Biomarkers in cancer screening, research and detection: present and future: a review. *Biomarkers*, 11(5):385–405, 2006. PMID: 16966157.

[58] Erik G Learned-Miller. Entropy and Mutual Information.

[59] Wentian Li. Mutual information functions versus correlation functions. *Journal of Statistical Physics*, 60(5):823–837, September 1990.

[60] Yixuan Li. Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction. *Applied and Computational Mathematics*, 7(4):212, 2018.

[61] Yang-Hsiang Lin, Chau-Ting Yeh, Cheng-Yi Chen, and Kwang-Huei Lin. Pseudogene: Relevant or Irrelevant? *Biomedical Journal*, 48(3):100790, June 2025.

[62] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[63] Lei Lv, Yujia Zhao, Qinqin Wei, Ye Zhao, and Qiyi Yi. Downexpression of HSD17B6 correlates with clinical prognosis and tumor immune infiltrates in hepatocellular carcinoma. *Cancer Cell International*, 20(1):210, June 2020.

[64] Candace D. Middlebrooks, A. Rouf Banday, Konichi Matsuda, Krizia-Ivana Udquim, Olusegun O. Onabajo, Ashley Paquin, Jonine D. Figueroa, Bin Zhu, Stella Koutros, Michiaki Kubo, Taro Shuin, Neal D. Freedman, Manolis Kogevinas, Nuria Malats, Stephen J. Chanock, Montserrat Garcia-Closas, Debra T. Silverman, Nathaniel Rothman, and Ludmila Prokunina-Olsson. Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nature Genetics*, 48(11):1330–1338, November 2016. Publisher: Nature Publishing Group.

[65] Thomas P Minka. A comparison of numerical optimizers for logistic regression.

[66] Christoph Molnar. *Interpretable Machine Learning*. Lulu.com, 2020. Google-Books-ID: jBm3DwAAQBAJ.

[67] Monica Morini, Simonetta Astigiano, Yorick Gitton, Laura Emionite, Valentina Mirisola, Giovanni Levi, and Ottavia Barbieri. Mutually exclusive expression of DLX2 and DLX5/6 is associated with the metastatic potential of the human breast cancer cell line MDA-MB-231. *BMC Cancer*, 10(1):649, November 2010.

[68] Anastasia S. Nikitina, Elena I. Sharova, Svetlana A. Danilenko, Tatiana B. Butusova, Alexandr O. Vasiliev, Alexandr V. Govorov, Elena A. Prilepskaya, Dmitry Y. Pushkar, and Elena S. Kostryukova. Novel RNA biomarkers of prostate cancer revealed by RNA-seq analysis of formalin-fixed samples obtained from Russian patients. *Oncotarget*, 8(20):32990–33001, March 2017.

[69] Yasunobu Nohara, Koutarou Matsumoto, Hidehisa Soejima, and Naoki Nakashima. Explanation of Machine Learning Models Using Improved Shapley Additive Explanation. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 546–546, Niagara Falls NY USA, September 2019. ACM.

[70] Anjan Kumar Payra and Anupam Ghosh. Mutual Information –The Biomarker of Essential Gene Predictions in Gene-Gene-Interaction of Lung Cancer. In Jyotsna Kumar Mandal, Somnath Mukhopadhyay, Paramartha Dutta, and Kousik Dasgupta, editors, *Computational Intelligence, Communications, and Business Analytics*, pages 232–244, Singapore, 2019. Springer.

[71] Li Peng, Xiu Wu Bian, Di Kang Li, Chuan Xu, Guang Ming Wang, Qing You Xia, and Qing Xiong. Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types. *Scientific Reports*, 5(1):13413, August 2015. Publisher: Nature Publishing Group.

[72] Josef Perktold, Skipper Seabold, Kevin Sheppard, ChadFulton, Kerby Shedden, jbrockmendel, j grana6, Peter Quackenbush, Vincent Arel-Bundock, Wes McKinney, Ian Langmore, Bart Baker, Ralf Gommers, yogabonito, s scherrer, Yauhen Zhurko, Matthew Brett, Enrico Giampieri, yl565, Jarrod Millman, Paul Hobson, Vincent, Pamphile Roy, Tom Augspurger, tvanzyl, alexbrc, Tyler Hartley, Fernando Perez, Yuji Tamiya, and Yaroslav Halchenko. statsmodels/statsmodels: Release 0.14.2, April 2024.

[73] João Pessoa, Marta Martins, Sandra Casimiro, Carlos Pérez-Plasencia, and Varda Shoshan-Barmatz. Editorial: Altered Expression of Proteins in Cancer: Function and Potential Therapeutic Targets. *Frontiers in Oncology*, 12:949139, June 2022.

[74] Yunzhen Qian, Yitao Gong, Xuan Zou, Yu Liu, Yusheng Chen, Ruijie Wang, Zhengjie Dai, Yesiboli Tasiheng, Xuan Lin, Xu Wang, Guopei Luo, Xianjun Yu, He Cheng, and Chen Liu. Aberrant APOBEC3C expression induces characteristic genomic instability in pancreatic ductal adenocarcinoma. *Oncogenesis*, 11(1):35, June 2022. Publisher: Nature Publishing Group.

[75] Malihe Ram, Ali Najafi, and Mohammad Taghi Shakeri. Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest. *Iranian Journal of Pathology*, 12(4):339–347, 2017.

[76] Julia D. Ransohoff, Yuning Wei, and Paul A. Khavari. The functions and unique features of long intergenic non-coding RNA. *Nature Reviews Molecular Cell Biology*, 19(3):143–157, March 2018. Publisher: Nature Publishing Group.

[77] Kimberly Roche, F. Alex Feltus, Jang Pyo Park, Marie-May Coissieux, Chenyan Chang, Vera B. S. Chan, Mohamed Bentires-Alj, and Brian W. Booth. Cancer cell redirection biomarker discovery using a mutual information approach. *PLOS ONE*, 12(6):e0179265, June 2017. Publisher: Public Library of Science.

[78] Brian C. Ross. Mutual Information between Discrete and Continuous Data Sets. *PLOS ONE*, 9(2):e87357, February 2014.

[79] Anne E Sarver, Aaron L Sarver, Venugopal Thayanithy, and Subbaya Subramanian. Identification, by systematic RNA sequencing, of novel candidate biomarkers and therapeutic targets in human soft tissue tumors. *Laboratory Investigation*, 95(9):1077–1088, September 2015.

[80] Lloyd S. Shapley. Notes on the N-Person Game — II: The Value of an N-Person Game. Technical report, August 1951.

[81] Garima Shukla, Sofia Singh, Chetan Dhule, Rahul Agrawal, Shipra Saraswat, Amal Al-Rasheed, Mohammed S. Alqahtani, and Ben Othman Soufiene. Point biserial correlation symbiotic organism search nanoengineering based drug delivery for tumor diagnosis. *Scientific Reports*, 14(1):6530, March 2024. Publisher: Nature Publishing Group.

[82] Kenong Su, Qi Yu, Ronglai Shen, Shi-Yong Sun, Carlos S. Moreno, Xiaoxian Li, and Zhaohui S. Qin. Pan-cancer analysis of pathway-based gene expression pattern at the individual level reveals biomarkers of clinical prognosis. *Cell Reports Methods*, 1(4), August 2021. Publisher: Elsevier.

[83] Sarah P. Suehnholz, Moriah H. Nissan, Hongxin Zhang, Ritika Kundra, Subhiksha Nandakumar, Calvin Lu, Stephanie Carrero, Amanda Dhaneshwar, Nicole Fernandez, Benjamin W. Xu, Maria E. Arcila, Ahmet Zehir, Aijazuddin Syed, A. Rose Brannon, Julia E. Rudolph, Eder Paraiso, Paul J. Sabbatini, Ross L. Levine, Ahmet Dogan, Jianjiong Gao, Marc Ladanyi, Alexander Drilon, Michael F. Berger, David B. Solit, Nikolaus Schultz, and Debyani Chakravarty. Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer. *Cancer Discovery*, 14(1):49–65, January 2024.

[84] Muhammad Aliyu Sulaiman and Jane Labadin. Feature selection based on mutual information. In *2015 9th International Conference on IT in Asia (CITA)*, pages 1–6, August 2015.

[85] Boyu Sun, Ziyu Xun, Nan Zhang, Kai Liu, Xiangqi Chen, and Haitao Zhao. Single-cell RNA sequencing in cancer research: discovering novel biomarkers and therapeutic targets for immune checkpoint blockade. *Cancer Cell International*, 23(1):313, December 2023.

[86] Ashraf Abou Tabl, Abedalrhman Alkhateeb, Waguih ElMaraghy, Luis Rueda, and Alioune Ngom. A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. *Frontiers in Genetics*, 10, March 2019.

[87] Paramjit S. Tappia and Bram Ramjiawan. Biomarkers for Early Detection of Cancer: Molecular Aspects. *International Journal of Molecular Sciences*, 24(6):5272, January 2023. Publisher: Multidisciplinary Digital Publishing Institute.

[88] Tian Tian, Fu Hong, Zhiwen Wang, Jiaru Hu, Ni Chen, Lei Lv, and Qiyi Yi. HSD17B6 downregulation predicts poor prognosis and drives tumor progression via activating Akt signaling pathway in lung adenocarcinoma. *Cell Death Discovery*, 7(1):341, November 2021. Publisher: Nature Publishing Group.

[89] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121, April 2001.

[90] Tajinder Ubhi, Olga Zaslaver, Andrew T. Quaile, Dennis Plenker, Pinjiang Cao, Nhu-An Pham, Angéla Békési, Gun-Ho Jang, Grainne M. O'Kane, Faiyaz Notta, Jason Moffat, Julie M. Wilson, Steven Gallinger, Beáta G. Vértessy, David A. Tuveson, Hannes L. Röst, and Grant W. Brown. Cytidine deaminases APOBEC3C and APOBEC3D promote DNA replication stress resistance in pancreatic cancer cells. *Nature Cancer*, 5(6):895–915, June 2024. Publisher: Nature Publishing Group.

[91] Quan Wan, Hayley Dingerdissen, Yu Fan, Naila Gulzar, Yang Pan, Tsung-Jung Wu, Cheng Yan, Haichen Zhang, and Raja Mazumder. BioXpress: an integrated RNA-seq-derived gene expression database for pan-cancer analysis. *Database*, 2015:bav019, January 2015.

[92] Xiao-Fei Wang, Bo Liang, Cheng Chen, Da-Xiong Zeng, Yu-Xiu Zhao, Nan Su, Wei-Wei Ning, Wen Yang, Jian-An Huang, Ning Gu, and Ye-Han Zhu. Long Intergenic Non-protein Coding RNA 511 in Cancers. *Frontiers in Genetics*, 11, July 2020. Publisher: Frontiers.

[93] Iris H. Wei, Yang Shi, Hui Jiang, Chandan Kumar-Sinha, and Arul M. Chinnaiyan. RNA-Seq Accurately Identifies Cancer Biomarker Signatures to Distinguish Tissue of Origin. *Neoplasia*, 16(11):918–927, November 2014.

[94] Eyal Winter. Chapter 53 The shapley value. In *Handbook of Game Theory with Economic Applications*, volume 3, pages 2025–2054. Elsevier, January 2002.

[95] Xiaochen Xi, Tianxiao Li, Yiming Huang, Jiahui Sun, Yumin Zhu, Yang Yang, and Zhi John Lu. RNA Biomarkers: Frontier of Precision Medicine for Cancer. *Non-Coding RNA*, 3(1):9, February 2017.

[96] Chuyu Xiao, Fuyang Hong, Guanzi Chen, Wenli Xu, and Yusheng Jie. MASP1 in stomach adenocarcinoma: linking diagnosis, prognosis, and tumor immunity. *Discover Oncology*, 16(1):1085, June 2025.

[97] Z.-H. Yan, Z.-S. Bao, W. Yan, Y.-W. Liu, C.-B. Zhang, H.-J. Wang, Y. Feng, Y.-Z. Wang, W. Zhang, G. You, Q.-G. Zhang, and T. Jiang. Upregulation of DLX2 Confers a Poor Prognosis in Glioblastoma Patients by Inducing a Proliferative Phenotype. *Current Molecular Medicine*, 13(3):438–445, March 2013. Publisher: Bentham Science Publishers.

[98] Hongjun Yu, Chaoqun Wang, Shanjia Ke, Yanan Xu, Shounan Lu, Zhigang Feng, Miaoyu Bai, Baolin Qian, Yue Xu, Zihao Li, Bing Yin, Xinglong Li, Yongliang Hua, Menghua Zhou, Zhongyu Li, Yao Fu, and Yong Ma. An integrative pan-cancer analysis of MASP1 and the potential clinical implications for the tumor immune microenvironment. *International Journal of Biological Macromolecules*, 280:135834, November 2024.

[99] Liuhai Zheng, Linzhi Li, Jun Xie, Hai Jin, and Naishuo Zhu. Six Novel Biomarkers for Diagnosis and Prognosis of Esophageal squamous cell carcinoma: validated by scRNA-seq and qPCR. *Journal of Cancer*, 12(3):899–911, January 2021.

[100] Xiaobo Zhou, Kuang-Yu Liu, and Stephen T. C. Wong. Cancer classification and prediction using logistic regression with Bayesian gene selection. *Journal of Biomedical Informatics*, 37(4):249–259, August 2004.

Appendix A

APPENDIX A

A.1 Additional Results Figures

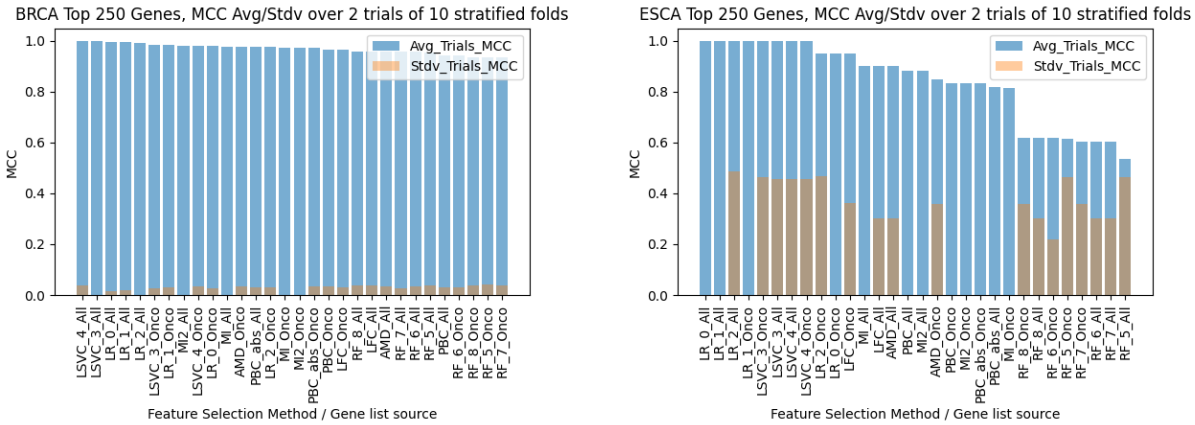

Figure A.1: Cohorts Breast invasive carcinoma (BRCA) Left, and Esophageal carcinoma (ESCA) Right. Step 2 MCC using top 250 genes by feature ranking method, OncoKB-gene sets included. With BRCA all sets and models achieved higher than .9 MCC. In ESCA, Random Forest (RF) models and sets performed significantly worse than other models. Method abbreviations are Logistic Regression (LR), Linear Support Vector Classifier (LSVC), model id, and suffix *_all* for 250 genes selected from all 20,530 TCGA genes, and suffix *_Onco* for restricting genes as well to only those in the OncoKB database.

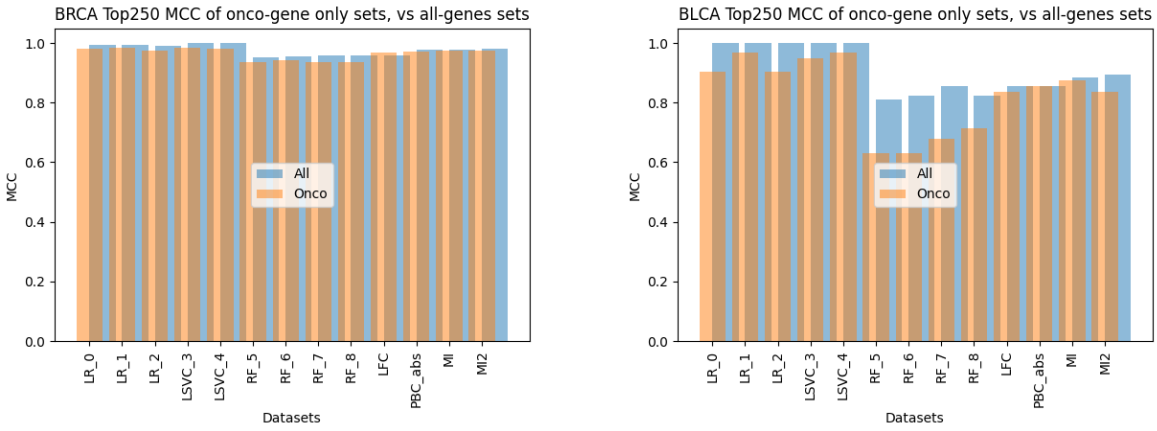

Figure A.2: Left: Breast invasive carcinoma (BRCA), Right: (Bladder Urothelial Carcinoma) BLCA. Using top 250 genes: MCC comparing all-gene sets versus OncoKB-gene only sets. In the BRCA dataset, there is slight difference, with usually all-gene sets being better. In BLCA the difference is more pronounced. Logistic Regression (LR), Linear Support Vector Classifier (LSVC), Random Forest (RF), Point-Biserial Correlation (PBC_abs), Log Fold Change (LFC), Mutual Information (MI)

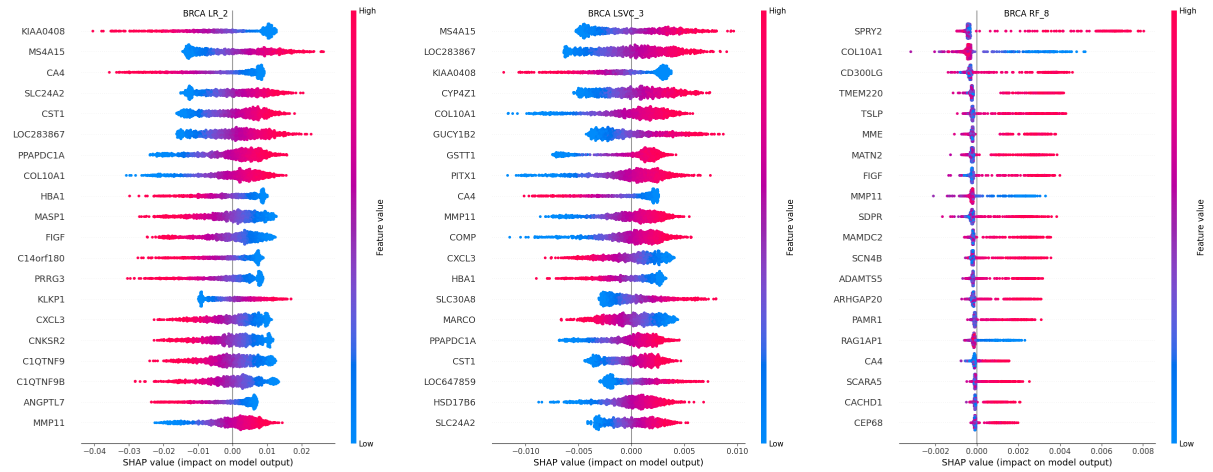

Figure A.3: For example: SHAP summary plots show how the feature value of each sample affected model with positive values on the X axis tending toward predicting the positive class (Primary Tumor). The color of the dots indicate a smaller value of the feature (blue) or higher (red.) The absolute average impact on model output of all samples was used as a feature ranking method. These three SHAP plots from the BRCA dataset, showing Left: Logistic Regression id 2 (LR_2), Center: Linear Support Vector Classifier 3 (LSVC_3), Right: Random Forest 8 (RF_8) show that the genes *COL10A1* and *MMP11* appeared in the top 20 features of three different architectures, out of 250 genes also selected by different inherent importance methods

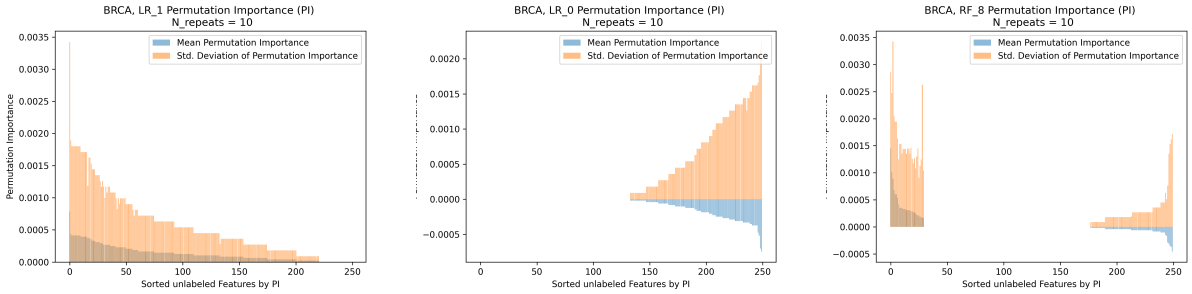

Figure A.4: Left: Logistic Regression 1 (LR_1), Center Logistic Regression 0 (LR_0), Right: Random Forest 8 (RF_8) On the Breast invasive carcinoma (BRCA) dataset, the top 250 features (by previous inherent feature importance ranking in Step 1 using all 20,530 features) were used to train the model, and Permutation Importance (PI) was calculated. PI can be negative indicating effectively removing the feature actually helps classification accuracy.

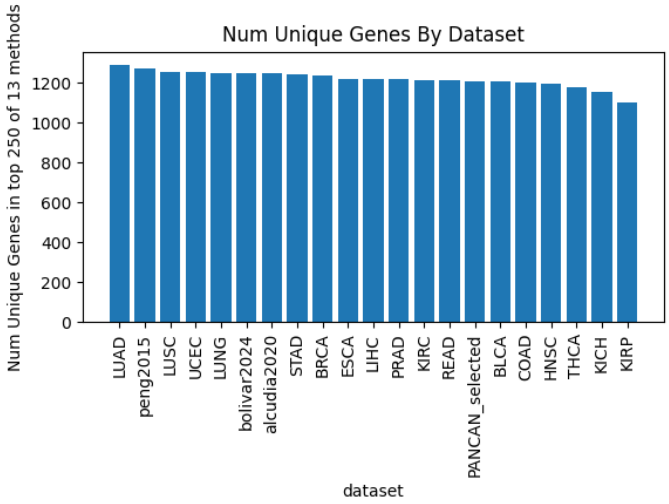

Figure A.5: For each dataset the number of unique genes in all 13 "all-gene" 250 sets is plotted. Among 3,250 total selected genes, there are between 1,000 and 1,200 unique genes. Lung adenocarcinoma (LUAD) has the highest number of unique genes of any dataset, showing less similarity in the top 250 lists. Kidney renal papillary cell carcinoma (KIRP) had the least number of unique genes and hence highest agreement between lists.

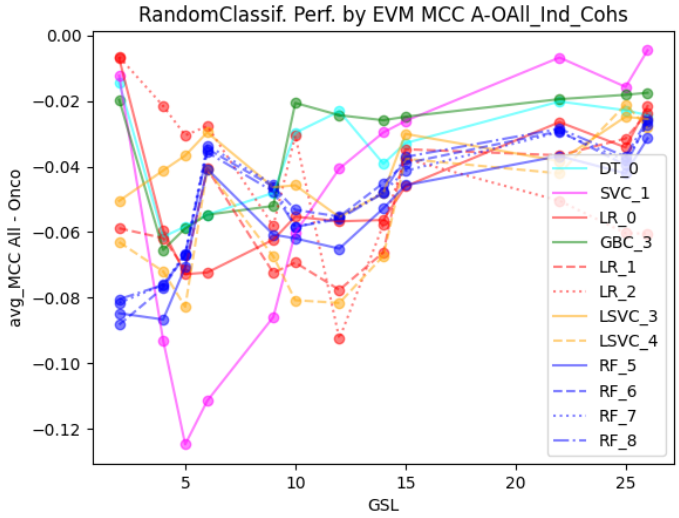

Figure A.6: Average MCC of random all-gene sets minus average MCC of random OncoKB gene sets. Random OncoKB gene-sets on average are better features for all Evaluation Models indicated by all plot lines below 0, but with no very clear pattern. With Gene-Set-Length (GSL) 5 the polynomial kernel Support Vector Classifier had on average .12 better MCC on random OncoKB genes than random genes at large. But when Gene-Set-Length (GSL) is 26 or 2, the difference is negligible.

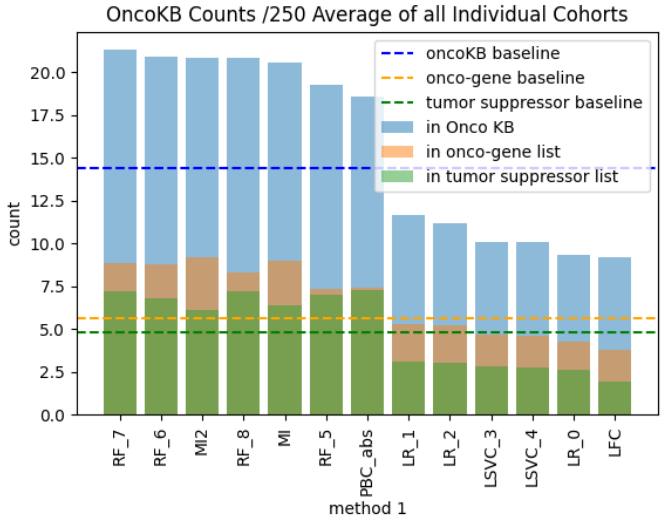

Figure A.7: The number of OncoKB genes in the top 250 genes by each Method-1 on average over all individual cohorts. Random Forest (RF), Mutual Information (MI), Point-Biserial Correlation (PBC_abs), Logistic Regression (LR), Linear Support Vector Classifier (LSVC), Log Fold Change (LFC).

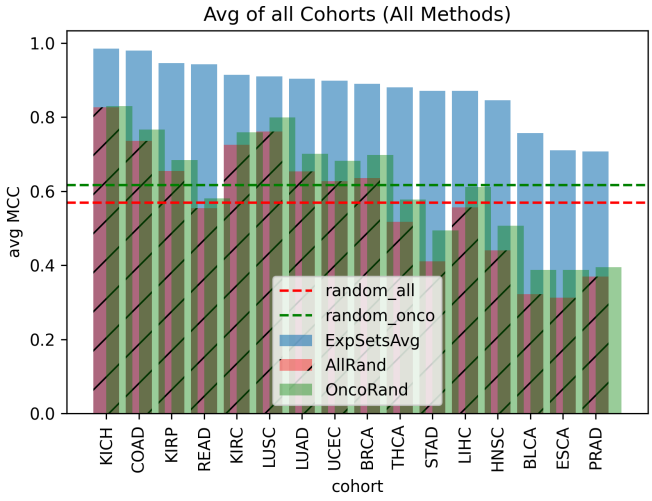

Figure A.8: Average MCC per cohort dataset, average of all methods, and all tested Gene-Set-Lengths from 26 to 2. Random baseline sets shown for comparison. KICH was most easily classified, both with experimentally generated feature sets, and random feature sets.

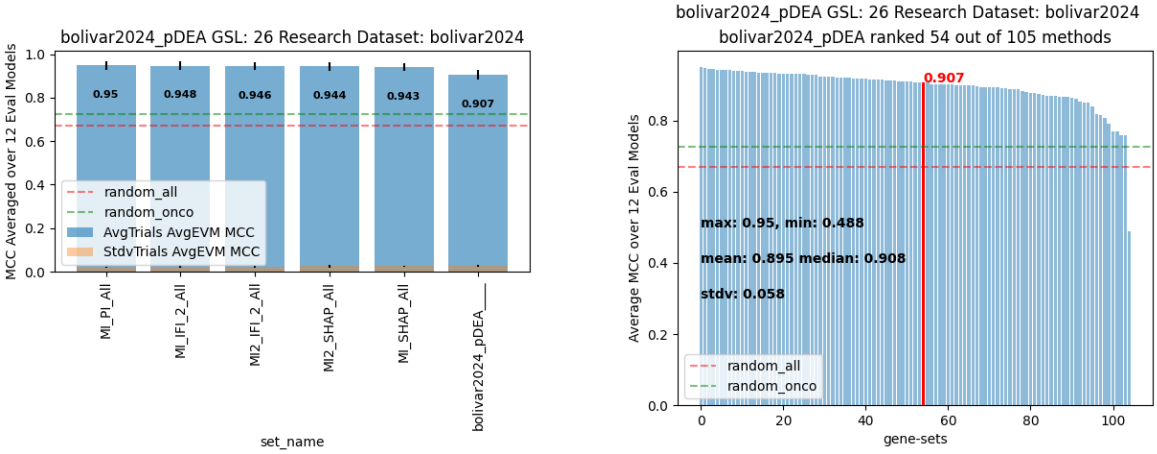

Figure A.9: 8 cohort combination dataset: de la Guardia-Bolívar et al. 26 paired Differential Expression biomarkers compared to experimental 26 feature sets. Left: Top 5 methods by MCC. Right: All methods' MCCs. Method 1: Mutual Information 1 or 2 (MI, MI2), Method 2: Inherent Feature Importance 2 (IFI_2), SHapley Additive exPlanations (SHAP). "_All" = unrestricted (not OncoKB only.)

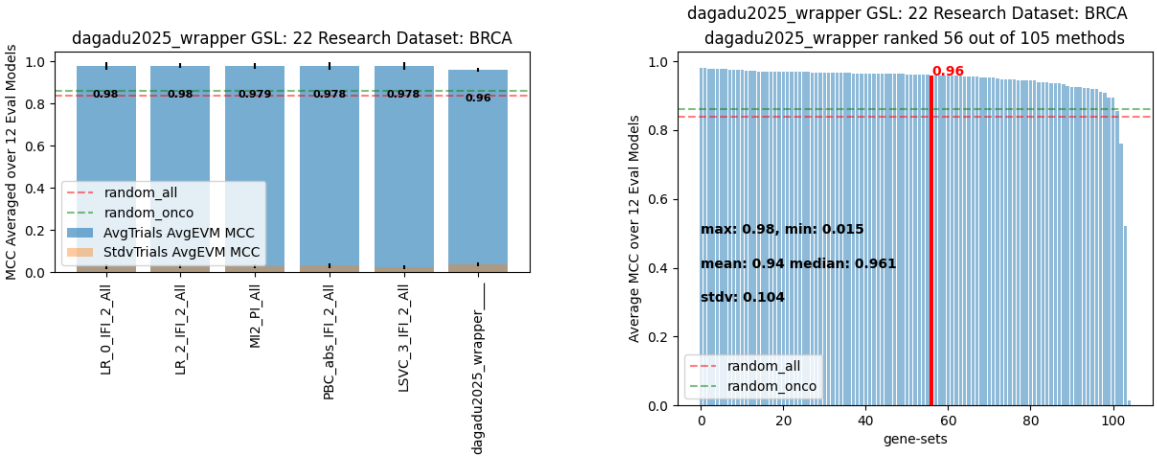

Figure A.10: BRCA cohort: Kallah-Dagadu et al. iML wrapper method based 22 biomarkers compared to experimental 22 feature sets. Left: Top 5 methods by MCC. Right: All methods' MCCs. Method 1: Logistic Regression 1, 2 (LR_0, LR_2), Mutual Information 2 (MI2), Point-Biserial Correlation (PBC_abs), Linear Support Vector Classifier 3 (LSVC_3). Method 2: Inherent Feature Importance 1 (IFI_1), Inherent Feature Importance 2 (IFI_2), Permutation Importance (PI). "_All" = unrestricted (not OncoKB only).

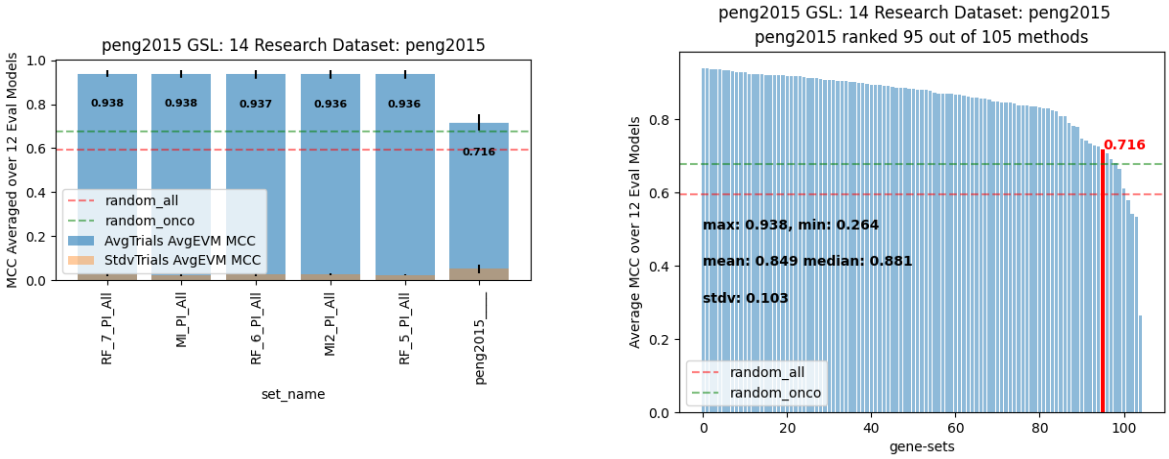

Figure A.11: Seven cohort combination: Peng et al. Differential Gene Expression and pathway analysis based 12 biomarkers compared to experimental 12 feature sets. Left: Top 5 methods by MCC. Right: All methods' MCCs. Method 1: Random Forest 7, 6, 5 (RF_7, RF_6, RF_5), Mutual Information 1, 2 (MI, MI2). Method 2: Permutation Importance (PI). "_All" = unrestricted (not OncoKB only).

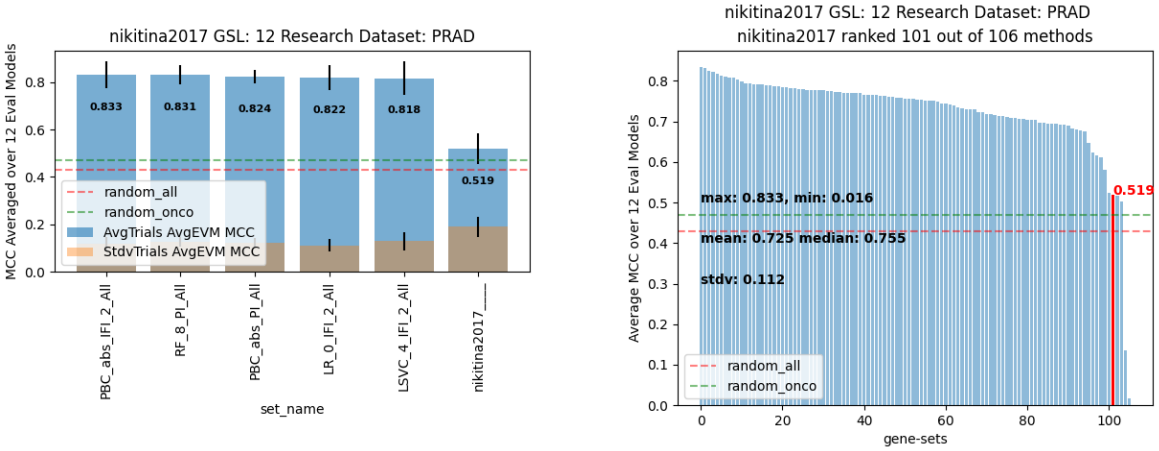

Figure A.12: PRAD cohort: Nikitina et al. Differential Gene Expression based 12 biomarkers compared to experimental 12 feature sets. Left: Top five methods by MCC. Right: All methods' MCCs. Method 1: Point-Biserial Correlation (PBC_abs), Random Forest (RF_8), Logistic Regression 0 (LR_0), Linear Support Vector Classifier (LSVC_4). Method 2: Inherent Feature Importance 2 (IFI_2), Permutation Importance (PI). "_All" = unrestricted (not OncoKB only).

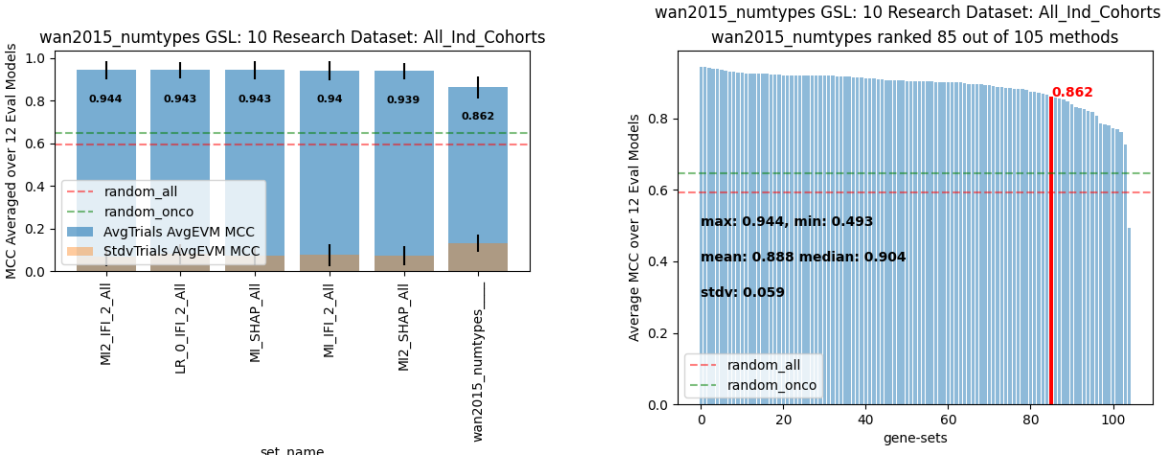

Figure A.13: Average of all 16 cohorts individually classified: Wan et al. 10 cohort-general Differential Gene Expression biomarkers "NumTypes" versus experimental 10 feature sets. Left: Top 5 methods by MCC. Right: All methods' MCCs. Method 1: Mutual Information 2, 1 (MI2, MI), Logistic Regression 0 (LR_0). Method 2: Inherent Feature Importance 2 (IFI_2), SHapley Additive exPlanations (SHAP). "_All" = unrestricted (not OncoKB only).

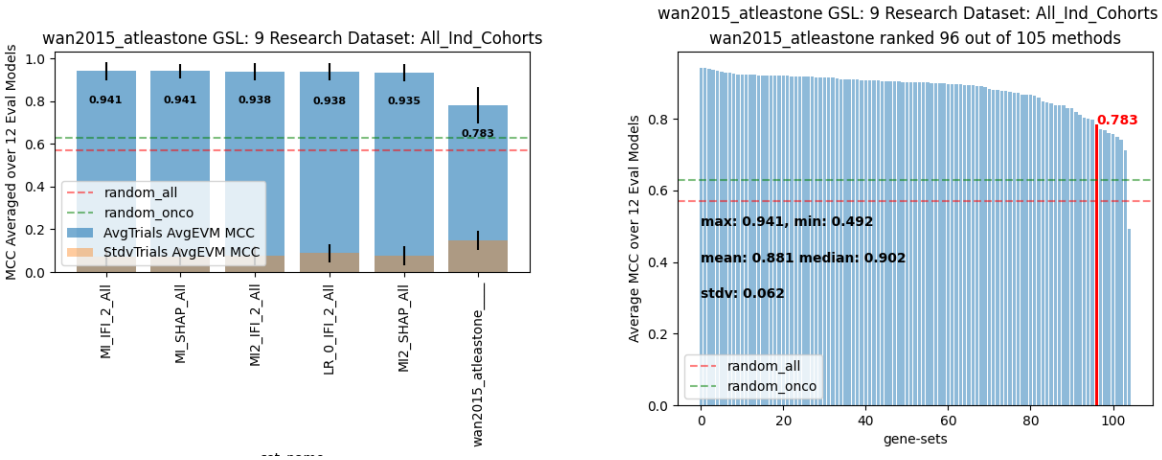

Figure A.14: Average of all 16 cohorts individually classified: Wan et al. 9 cohort-general biomarkers "AtLeastOne" versus experimental 9 feature sets. Left: Top 5 methods by MCC. Right: All methods' MCCs. Method 1: Mutual Information 1, 2 (MI, MI2), Logistic Regression 0 (LR_0). Method 2: Inherent Feature Importance 2 (IFI_2), SHapley Additive exPlanations (SHAP). "_All" = unrestricted (not OncoKB only).

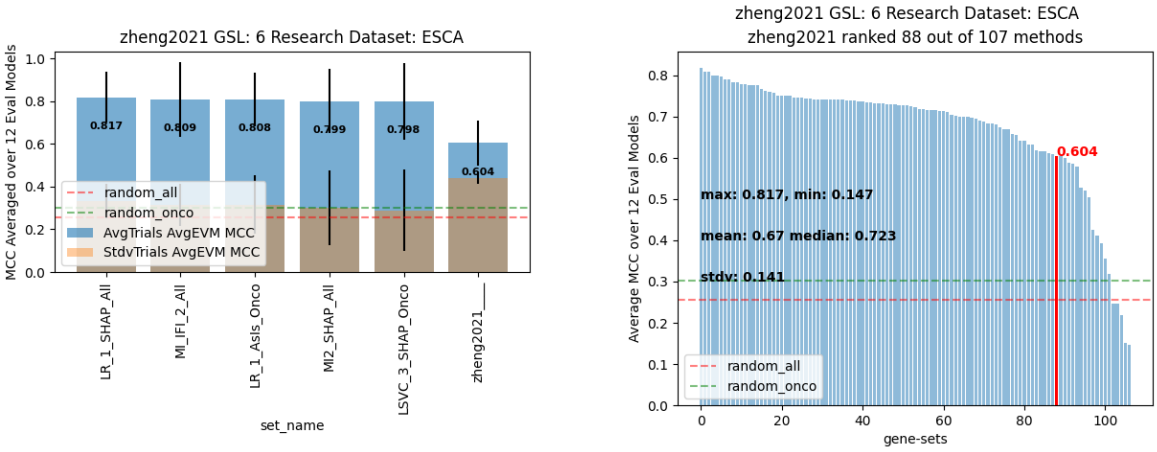

Figure A.15: ESCA: Zheng et al. 6 overlapping Differential Gene Expression (DGE) hub-gene biomarkers versus experimental 9 feature sets. Left: Top 5 methods by MCC. Right: All methods' MCCs. Method 1: Logistic Regression 1 (LR_1), Mutual Information 1, 2 (MI, MI2), Linear Support Vector Classifier 3 (LSVC_3). Method 2: SHapley Additive exPlanations (SHAP), Inherent Feature Importance 2 (IFI_2), "AsIs" (IFI_1 unchanged). "_All" = unrestricted (not OncoKB only). "_Onco" = OncoKB only genes.

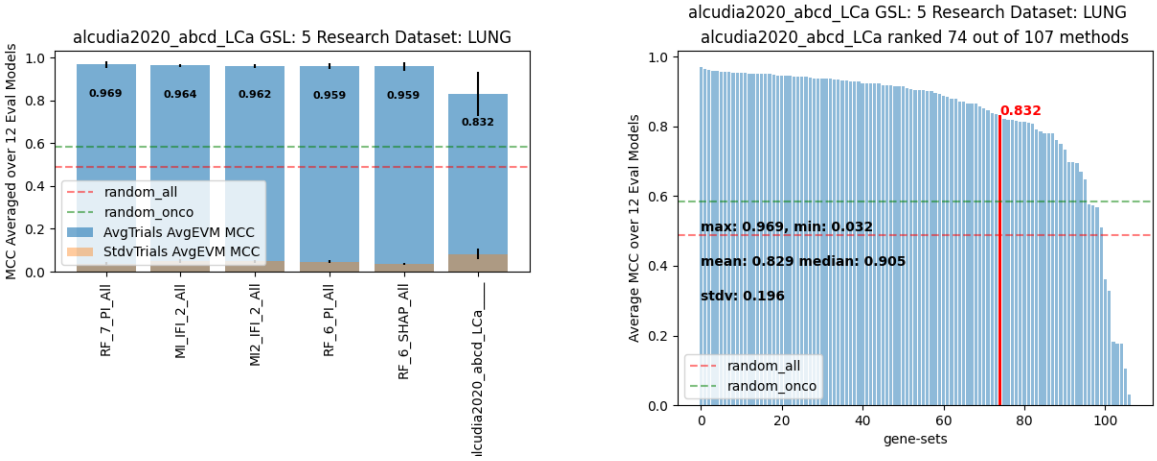

Figure A.16: LUAD + LUSC dataset: Coleto-Alcudia et al. Artifical Bee Colony based on Dominance (ABCD) 5 biomarkers versus experimental 5 feature sets. Left: Top 5 methods by MCC. Right: All methods' MCCs. Method 1: Random Forest 7, 6 (RF_7, RF_6) Mutual Information 1, 2 (MI, MI2). Method 2: Permutation Importance (PI), Inherent Feature Importance 2 (IFI_2), SHapley Additive exPlanations (SHAP). "_All" = unrestricted (not OncoKB only).

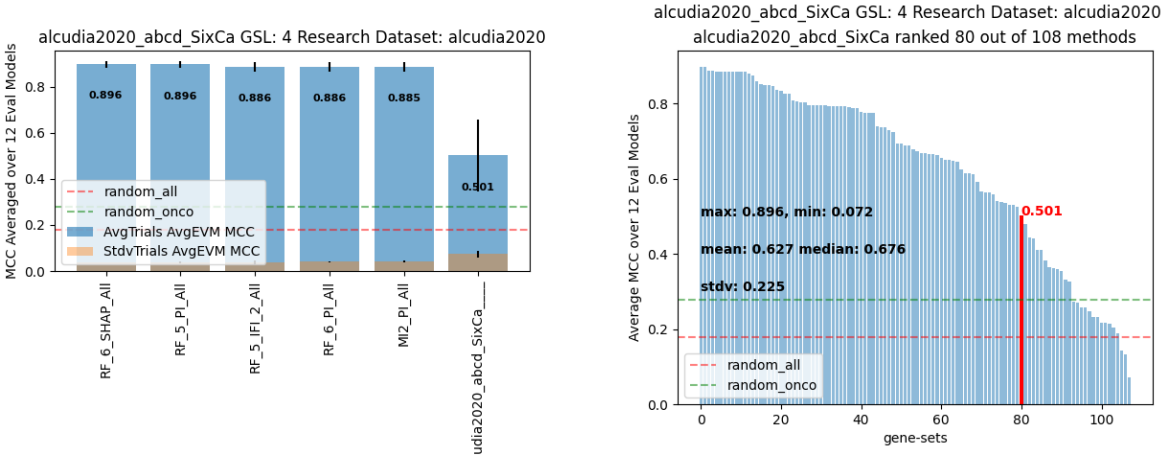

Figure A.17: 7 cohort combination dataset: Coleto-Alcudia et al. ABCD 4 biomarkers versus experimental 4 feature sets. Left: Top 5 methods by MCC. Right: All methods' MCCs. Method 1: Random Forest 6, 5 (RF_6, RF_5), Mutual Information 2 (MI2). Method 2: SHapley Additive exPlanations (SHAP), Permutation Importance (PI), Inherent Feature Importance 2 (IFI_2). "_All" = unrestricted (not OncoKB only).

A.2 TCGA Cohort Study Abbreviations

Table A.1: Abbreviations for cancer type cohorts in TCGA [13]

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

A.3 Research Biomarker lists

Table A.2: A table of the gene-sets from previous literature which were used to train classifiers during the evaluation state. * These gene-sets contained 1 gene name or more **gene, names that was not in the TCGA data features, so the list is one gene shorter than intended. Names in parentheses are alias gene names used in the original paper.

Reference: Cohorts	genes (number of genes)
[34]: Cohorts: COAD, BRCA, LUAD, KIRC, STAD, LIHC, THCA, UCEC	26 genes: ANGPTL1, GSTM5, IQGAP3, UHRF1, CCBE1, MYBL2, CHRDL1, PKMYT1, ABCA8, CTHRC1, PKNOX2, UBE2C, DES, RELN, ADH1B, MT1M, CDT1, FAM111B, SFRP1, C7, GHR, LYVE1, IGSF10, MFAP4, RNF150, HBB
[33] : Cohorts: PRAD	25 genes: ACTB, ANLN, AURKA, AURKB, BIRC5, CCNB1, CDCA3, CDCA8, CDC6, CDC7, CDK1, CHEK1, CHEK2, HMMR, KIF20A, KIF2C, MKI67, NEK2, NUSAP1, PTTG1, RRM2, TOP2A, TPX2, UBE2C
[50] : Cohorts: BRCA	22 genes: COL10A1, NPR1, SDPR, RPLP0P2, IBSP, UBE2T, AURKA, BMP8A, C16orf59, CD300LG, CENPL, CLEC3B, DONSON, ECT2, FIGF, FXYD1, GABRD, GINS1, LIMS2, MME, MMP11, TGFBR2
[71] : Cohorts: BLCA, BRCA, COAD, HNSC, LIHC, LUAD, LUSC	14 genes: KIF4A, NUSAP1, HJURP, NEK2, FANCI, UHRF1, FEN1, DTL, IQGAP3, KIF20A, TRIM59, CENPL, C16orf59, UBE2C

[68] : Cohorts: PRAD	12 genes: ANKRD34B, NEK5, KCNG3, PTPRT, FOXJ2, GATA6, NFE2L1, NFIL3, PRRX2, TEF, EBF2, ZNF238
[91]: Cohorts: All Cohorts	10 genes: COL10A1, COL11A1, MMP11, TMPRSS4, MMP1, ADH1B, MT1H, MT1G, CHRDL1, CA4
[91] : Cohorts: All Cohorts	9 genes: CCL21, GGT6, UBD, MMP7, NCAM1, CHRDL1, WFDC2, LCN2, KRT80
[99] : Cohorts: ESCA	6 genes: PBK, KIF2C, NUF2, KIF20A, RAD51AP1, DEPDC1
[28] : Cohorts: LUAD, LUSC	5 genes: FTSJ1, PLEKHA8, DCN, ABCC3, PRCD
[28] : Cohorts: BRCA, LIHC, ESCA, TGCT, THCA, LUAD, LUSC	4* genes: MYBL2, PRKCI, NRBP1, C18orf45 (TMEM241),*RNY4P7
[28] : Cohorts: BRCA	3* genes: RIMS3, C9orf172 (AJM1), KIFC1, * LINC01614
[28] : Cohorts: PRAD	2 genes: C14orf72, CSRP2

A.4 Pathway Analysis

Table A.3: For each cohort, the top 10 genes by number of occurrences in top 12 ranks of all gene-sets modified by the set's average evaluation model performance across all Gene-Set-Lengths from 26 to 2. A KEGG query returned pathway and disease annotations for each of the 10 genes, the top 10 pathways by count of the top 10 genes or gene rank are displayed.

cohort	top10	kwhs	onco_count	Top10Pathways	Top5Diseases

BLCA	ESM1 PLP1 PMP2 CMTM5 F10 CLEC3B C16orf89 CFD SNORA39 COL10A1	0	0	hsa04610 Complement and coagulation cascades (2) , hsa05150 Staphylococcus aureus infection (1) , hsa05171 Coronavirus disease - COVID-19 (1) , hsa04974 Protein digestion and absorption (1)	H00266 Hereditary spastic paraplegia (1) , H00679 Hypomyelinating leukodystrophy (1) , H00264 Charcot-Marie-Tooth disease (1) , H02257 Factor X deficiency (1) , H01770 Macular dystrophy (1)
BRCA	COL10A1 MMP11 CA4 KIAA0408 CD300LG TMEM220 HSD17B6 SPRY2 MASP1 FIGF	2	0	hsa01100 Metabolic pathways (2) , hsa04974 Protein digestion and absorption (1) , hsa00910 Nitrogen metabolism (1) , hsa04964 Proximal tubule bicarbonate reclamation (1) , hsa00140 Steroid hormone biosynthesis (1) , hsa00830 Retinol metabolism (1) , hsa01240 Biosynthesis of cofactors (1) , hsa05206 MicroRNAs in cancer (1) , hsa04610 Complement and coagulation cascades (1) , hsa05150 Staphylococcus aureus infection (1)	H00479 Metaphyseal dysplasias (1) , H00527 Retinitis pigmentosa (1) , H01887 3MC syndrome (1)
COAD	KRT80 CDH3 ESM1 OTOP2 BEST4 PYY OSBPL3 GLP2R CA7 LOC100190940	1	0	hsa04080 Neuroactive ligand-receptor interaction (2) , hsa04382 Cornified envelope formation (1) , hsa04514 Cell adhesion molecules (1) , hsa00910 Nitrogen metabolism (1) , hsa01100 Metabolic pathways (1)	H00639 Ectodermal dysplasia (1) , ectrodactyly (1) , and macular dystrophy (1) , H00785 Congenital hypotrichosis with juvenile macular dystrophy (1)
ESCA	HOXC11 UBE2T MMP12 SECISBP2L PSMB2 TMEM132C NCRNA00152 RD3 HOXC8 GABRD	0	2	hsa03460 Fanconi anemia pathway (1) , hsa03050 Proteasome (1) , hsa05010 Alzheimer disease (1) , hsa05012 Parkinson disease (1) , hsa05014 Amyotrophic lateral sclerosis (1) , hsa05016 Huntington disease (1) , hsa05017 Spinocerebellar ataxia (1) , hsa05020 Prion disease (1) , hsa05022 Pathways of neurodegeneration - multiple diseases (1) , hsa04080 Neuroactive ligand-receptor interaction (1)	H00238 Fanconi anemia (1) , H00837 Leber congenital amaurosis (1) , H00808 Idiopathic generalized epilepsies (1) , H02217 Juvenile myoclonic epilepsy (1) , H02564 Generalized epilepsy with febrile seizures plus (1)
HNSC	CA9 ADIPOQ NRG2 SH3BGRL2 ADH4 IGF2BP1 UBL3 GJC1 MGC12982 GRIN2D	1	0	hsa01100 Metabolic pathways (2) , hsa04936 Alcoholic liver disease (2) , hsa05014 Amyotrophic lateral sclerosis (2) , hsa00910 Nitrogen metabolism (1) , hsa03320 PPAR signaling pathway (1) , hsa04081 Hormone signaling (1) , hsa04152 AMPK signaling pathway (1) , hsa04211 Longevity regulating pathway (1) , hsa04920 Adipocytokine signaling pathway (1) , hsa04930 Type II diabetes mellitus (1)	H00967 Adiponectin deficiency (1) , H00606 Early infantile epileptic encephalopathy (1)
KICH	UGT3A1 IRX1 LOC723809 MAPK15 CAPSL PTGER1 KCTD16 RALYL C5orf38 EDAR	2	0	hsa04657 IL-17 signaling pathway (1) , hsa04020 Calcium signaling pathway (1) , hsa04080 Neuroactive ligand-receptor interaction (1) , hsa05163 Human cytomegalovirus infection (1) , hsa05200 Pathways in cancer (1) , hsa04060 Cytokine-cytokine receptor interaction (1) , hsa04064 NF-kappa B signaling pathway (1)	H00649 Ectodermal dysplasia (1) , hair-nail type (1) , H00651 Hypohidrotic ectodermal dysplasia (1)

KIRC	GPC5 OVCH2 AQP2 GABRD BIRC7 AIF1L HSPC072 PCDHB1 SLC9A4 TRPV6	1	0	hsa04962 Vasopressin-regulated water reabsorption (1) , hsa04080 Neuroactive ligand-receptor interaction (1) , hsa04082 Neuroactive ligand signaling (1) , hsa04723 Retrograde endocannabinoid signaling (1) , hsa04727 GABAergic synapse (1) , hsa05032 Morphine addiction (1) , hsa05033 Nicotine addiction (1) , hsa04120 Ubiquitin mediated proteolysis (1) , hsa04215 Apoptosis - multiple species (1) , hsa05145 Toxoplasmosis (1)	H00252 Congenital nephrogenic diabetes insipidus (1) , H00808 Idiopathic generalized epilepsies (1) , H02217 Juvenile myoclonic epilepsy (1) , H02564 Generalized epilepsy with febrile seizures plus (1) , H02030 Neonatal hyperparathyroidism (1)
KIRP	FRMD7 ATP12A UMOD TMEM207 TFAP2B GP2 RASL11B MUC15 HRG TRPV5	0	0	hsa00190 Oxidative phosphorylation (1) , hsa01100 Metabolic pathways (1) , hsa04928 Parathyroid hormone synthesis (1) , secretion and action (1) , hsa04961 Endocrine and other factor-regulated calcium reabsorption (1)	H00776 Congenital motor nystagmus (CMN) (1) , H00541 Autosomal dominant tubulointerstitial kidney disease (1) , H02011 Familial juvenile hyperuricemic nephropathy (1) , H02012 Medullary cystic kidney disease (1) , H00555 Char syndrome (1)
LIHC	NOL4 GABRD PTH1R CXorf36 ARSF COLEC10 BMPER CDH13 CLEC1B FAM83F	0	0	hsa04080 Neuroactive ligand-receptor interaction (2) , hsa04082 Neuroactive ligand signaling (1) , hsa04723 Retrograde endocannabinoid signaling (1) , hsa04727 GABAergic synapse (1) , hsa05032 Morphine addiction (1) , hsa05033 Nicotine addiction (1) , hsa04081 Hormone signaling (1) , hsa04928 Parathyroid hormone synthesis (1) , secretion and action (1) , hsa04961 Endocrine and other factor-regulated calcium reabsorption (1)	H00808 Idiopathic generalized epilepsies (1) , H02217 Juvenile myoclonic epilepsy (1) , H02564 Generalized epilepsy with febrile seizures plus (1) , H00479 Metaphyseal dysplasias (1) , H00495 Eiken dysplasia (1)
LUAD	CD5L GBP7 PYCR1 EMP2 STX11 AQP4 LOC84740 B3GNT6 AGER GLB1L3	0	0	hsa01100 Metabolic pathways (2) , hsa04621 NOD-like receptor signaling pathway (1) , hsa00330 Arginine and proline metabolism (1) , hsa01230 Biosynthesis of amino acids (1) , hsa04510 Focal adhesion (1) , hsa04130 SNARE interactions in vesicular transport (1) , hsa04962 Vasopressin-regulated water reabsorption (1) , hsa04976 Bile secretion (1) , hsa00515 Mannose type O-glycan biosynthesis (1) , hsa04148 Efferocytosis (1)	H00557 Cutis laxa (1) , H00558 Geroderma osteodysplasticum (1) , H01657 Nephrotic syndrome (1) , H00109 Familial hemophagocytic lymphohistiocytosis (1) , H00875 Megaloencephalic leukoencephalopathy with subcortical cysts (1)
LUSC	CACNA2D2 STX11 B4GALNT4 GKN2 GAL RS1 INMT NCAPH STON1-GTF2A1L SFTPC	0	0	hsa01100 Metabolic pathways (2) , hsa04010 MAPK signaling pathway (1) , hsa04260 Cardiac muscle contraction (1) , hsa04261 Adrenergic signaling in cardiomyocytes (1) , hsa04921 Oxytocin signaling pathway (1) , hsa05410 Hypertrophic cardiomyopathy (1) , hsa05412 Arrhythmogenic right ventricular cardiomyopathy (1) , hsa05414 Dilated cardiomyopathy (1) , hsa04130 SNARE interactions in vesicular transport (1) , hsa00513 Various types of N-glycan biosynthesis (1)	H02645 Cerebellar atrophy with seizures and variable developmental delay (1) , H00109 Familial hemophagocytic lymphohistiocytosis (1) , H00809 Familial epilepsy temporal lobe (ETL) (1) , H02475 Retinoschisis (1) , H00269 Primary microcephaly (1)

PRAD	DLX1 NKX2-3 MED21 APOBEC3C LOC644936 H3F3C DLX2 EFNB1 PCA3 PAK3	2	2	hsa03250 Viral life cycle - HIV-1 (1) , hsa05170 Human immunodeficiency virus 1 infection (1) , hsa04613 Neutrophil extracellular trap formation (1) , hsa05034 Alcoholism (1) , hsa05131 Shigellosis (1) , hsa05202 Transcriptional misregulation in cancer (1) , hsa05322 Systemic lupus erythematosus (1) , hsa04360 Axon guidance (1) , hsa04810 Regulation of actin cytoskeleton (1) , hsa05135 Yersinia infection (1)	H00458 Syndromic craniosynostoses (1) , H01992 Craniofrontonasal syndrome (1)
READ	SALL4 OTOP3 TMEFF2 CELSR3 OTOP2 KRT24 KRT80 GRIN2D FABP6 SLC17A8	2	1	hsa04382 Cornified envelope formation (2) , hsa04082 Neuroactive ligand signaling (2) , hsa04724 Glutamatergic synapse (2) , hsa05033 Nicotine addiction (2) , hsa04915 Estrogen signaling pathway (1) , hsa05150 Staphylococcus aureus infection (1) , hsa04020 Calcium signaling pathway (1) , hsa04024 cAMP signaling pathway (1) , hsa04080 Neuroactive ligand-receptor interaction (1) , hsa04713 Circadian entrainment (1)	H00634 Duane-radial ray syndrome (1) , H02283 IVIC syndrome (1) , H00606 Early infantile epileptic encephalopathy (1) , H00604 Deafness (1) , autosomal dominant (1)
STAD	COL10A1 ESM1 HOXC11 MTHFD1L HOTAIR HOXC10 HOXC9 RPLP0P2 CST1 GABRD	0	1	hsa04974 Protein digestion and absorption (1) , hsa00670 One carbon pool by folate (1) , hsa01100 Metabolic pathways (1) , hsa01240 Biosynthesis of cofactors (1) , hsa04970 Salivary secretion (1) , hsa04080 Neuroactive ligand-receptor interaction (1) , hsa04082 Neuroactive ligand signaling (1) , hsa04723 Retrograde endocannabinoid signaling (1) , hsa04727 GABAergic synapse (1) , hsa05032 Morphine addiction (1)	H00479 Metaphyseal dysplasias (1) , H00808 Idiopathic generalized epilepsies (1) , H02217 Juvenile myoclonic epilepsy (1) , H02564 Generalized epilepsy with febrile seizures plus (1)
THCA	GABRB2 HAPLN1 DPT NRXN1 RELN FAM84A GABRD CDKN2B C6orf138 SERINC2	1	2	hsa04080 Neuroactive ligand-receptor interaction (2) , hsa04082 Neuroactive ligand signaling (2) , hsa04723 Retrograde endocannabinoid signaling (2) , hsa04727 GABAergic synapse (2) , hsa05032 Morphine addiction (2) , hsa05033 Nicotine addiction (2) , hsa04726 Serotonergic synapse (1) , hsa04514 Cell adhesion molecules (1) , hsa04151 PI3K-Akt signaling pathway (1) , hsa04510 Focal adhesion (1)	H00606 Early infantile epileptic encephalopathy (1) , H02150 Infantile or early childhood epileptic encephalopathy (1) , H00756 Pitt-Hopkins syndrome (1) , H01649 Schizophrenia (1) , H00268 Lissencephaly (1)
UCEC	OTX1 GNGT1 SLC24A3 LRRN4CL ALDH1A2 LOC134466 LMOD1 KLHDC1 TUBA1C POC1A	2	0	hsa04550 Signaling pathways regulating pluripotency of stem cells (1) , hsa04014 Ras signaling pathway (1) , hsa04062 Chemokine signaling pathway (1) , hsa04081 Hormone signaling (1) , hsa04082 Neuroactive ligand signaling (1) , hsa04151 PI3K-Akt signaling pathway (1) , hsa04371 Apelin signaling pathway (1) , hsa04713 Circadian entrainment (1) , hsa04723 Retrograde endocannabinoid signaling (1) , hsa04724 Glutamatergic synapse (1)	H01241 Congenital diaphragmatic hernia (1) , H01869 Megacystis microcolon intestinal hypoperistalsis syndrome (1) , H01897 Oocyte/zygote/embryo maturation arrest (1) , H02481 Syndromic disorder with short stature (1)

A.5 BRCA Top Ranked Genes

From the Breast invasive carcinoma dataset (BRCA) the top scored genes based on number of occurrences in top 12 genes, and average set performance between 26 and 2 features used were *COL10A1*, *MMP11*, *CA4*, *KIAA0408*, *CD300LG*, *TMEM220*, *HSD17B6*, *SPRY2*, *MASP1*, and *FIGF*. The 10 most common pathways among these top 10 genes were often (cellular) metabolism related, see Table A.3.

COL10A1, *MMP11*, *CD300LG*, and *FIGF* were also identified in Kallah-Dagadu et al.'s iML wrapper method among 22 biomarker genes. *COL10A1*, *CA4*, and *MMP11* were selected by Wan et al. in the top 10 or 9 genes, with their differential gene expression based methods on multiple cancer cohorts. *COL10A1* was also selected as a top gene in this experiment in the Bladder Urothelial Carcinoma (BLCA) dataset. From the Kyoto Encyclopedia of Genes and Genomes (KEGG) *COL10A1* is listed as belonging to the Digestive system, and Protein digestion and absorption gene pathways. Metaphyseal dysplasias, a rare skeletal genetic disorder is listed in the KEGG with 7 associated genes, including *COL10A1*.

Among the five remaining genes not present in any other set in this experiment, a cursory search with the gene name among peer reviewed articles revealed in the first few results:

KIAA0408 was found to be among six identified gene biomarkers for predicting outcomes in Lung squamous cell carcinoma (LUSC) [44]. Additionally, *KIAA0408* was listed in another paper among 359 down-regulated genes in tumor samples of BRCA data [23].

HSD17B6 was named in the title of published studies on Lung adenocarcinoma (LUAD) [88], hepatocellular carcinoma (HCC) [63], and mentioned among a family of genes in a publication about Breast Cancer [45].

TMEM220 (along with again *COL10A1, MMP11*, and *CD300LG*) was identified with an ML-based biomarker discovery study in BRCA data [49]. Another published study on BRCA relating to DNA-methylation lists *TMEM220* among 11 long noncoding RNAs related to the survival of Breast Cancer patients [39].

SPRY2 [31] is the subject of a published study on Breast Cancer, and *MASP1* has noted

in publications on pan-cancer, and stomach cancer [98, 96].

A.6 Misc.

PTPN11_x, ERBB2_x in LUAD cohort changed to PTPN11 and ERBB2

TCGA gene names: There are 66 genes in the TCGA data labeled as "LOC" referring to the gene's position of the gene in the chromosome, and without an official symbol.

A.6.1 Scikit-Learn Logistic Regression Hyperparameters:

"newton-cg" is a Logistic Regression *solver* option. "Newton-CG" is Newton Conjugate Gradient. "Newton-Cholesky" apparently constructs full Hessian Matrix, Newton-CG uses an approximation instead for speed. [65, 6]

C in both Logistic Regression and Linear Support Vector is the adjustable regularization parameter, which is inversely proportional to regularization. The smaller C, the higher regularization. Too little regularization can lead to overfitting, too much regularization can lead to under-fitting. *Tol* is tolerance or stopping parameter, the smaller the more precise and more iterations.