Spring 2019

Machine learning with the cancer genome atlas head and neck squamous cell carcinoma dataset: improving usability by addressing inconsistency, sparsity, and high-dimensionality

Michael Rendleman
University of Iowa

Recommended Citation

Rendleman, Michael. "Machine learning with the cancer genome atlas head and neck squamous cell carcinoma dataset: improving usability by addressing inconsistency, sparsity, and high-dimensionality." MS (Master of Science) thesis, University of Iowa, 2019.
https://doi.org/10.17077/etd.37hw-o12m

Follow this and additional works at: https://ir.uiowa.edu/etd
Part of the Electrical and Computer Engineering Commons

MACHINE LEARNING WITH THE CANCER GENOME ATLAS HEAD AND NECK SQUAMOUS CELL CARCINOMA DATASET: IMPROVING USABILITY BY ADDRESSING INCONSISTENCY, SPARSITY, AND HIGH-DIMENSIONALITY.

by

Michael Rendleman

A thesis submitted in partial fulfillment
of the requirements for the
Master of Science degree in Electrical and Computer Engineering in the
Graduate College of
The University of Iowa

May 2019

Thesis Supervisor: Professor Thomas L. Casavant

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. John M. Buatti for contributions of domain knowledge, Dr. Brian J. Smith for statistical insights and experimental design guidance, Chibuzo Nwakama for assistance in executing experiments, Dr. Terry A. Braun for technical assistance and feedback on experimental design, Dr. Reinhard R. Beichel for analyses for dataset selection, the University of Iowa High Performance Computing group for computational resources, and Dr. Thomas L. Casavant for providing research guidance, mentorship, and other professional support.

ABSTRACT

In the era of precision oncology and publicly available datasets, the amount of information available for each patient case has dramatically increased. From clinical variables and PET-CT radiomics measures to DNA-variant and RNA expression profiles, such a wide variety of data presents a multitude of challenges. Large clinical datasets are subject to sparsely and/or inconsistently populated fields. Corresponding sequencing profiles can suffer from the problem of high-dimensionality, where making useful inferences can be difficult without correspondingly large numbers of instances. In this thesis we report a novel deployment of machine learning techniques to handle data sparsity and evaluate biomarkers in the form of unsupervised transformations of RNA data. Additionally, we evaluate the output of MutSig2CV from the Broad Firehose pipeline as a set of potential biomarkers to supplement our clinical predictive models.

We apply preprocessing, MICE imputation, and sparse principal component analysis (SPCA) to improve the usability of more than 500 patient cases from The Cancer Genome Atlas Head and Neck Squamous Cell Carcinoma dataset (TCGA-HNSC) for enhancing oncological decision support for Head and Neck Squamous Cell Carcinoma (HNSCC). Imputation was shown to improve prognostic ability of sparse clinical treatment variables. Dimensionality reduction of RNA expression profiles via SPCA improved computation cost and model training/evaluation time without affecting classifier performance while simultaneously providing a convenient avenue for consideration of biological context via gene ontology enrichment analysis. Statistical comparison of mutation significance features with similar but meaningless data revealed that while the MutSig2CV data have predictive value for the problem of predicting

two-year recurrence-free survival, this value yielded no significant performance increases to simpler, clinical feature-based models.

PUBLIC ABSTRACT

In recent years, more data is becoming available for historical oncology case analysis. A large dataset that describes over 500 patient cases of Head and Neck Squamous Cell Carcinoma is a potential goldmine for finding ways to improve oncological decision support. Unfortunately, the best approaches for finding useful inferences are unknown. With so much information, from DNA and RNA sequencing to clinical records, we must use computational learning to find associations and biomarkers.

The available data has sparsity, inconsistencies, and is very large for some datatypes. We processed clinical records with an expert oncologist and used complex modeling methods to substitute (impute) data for cases missing treatment information. We used machine learning algorithms to see if imputed data is useful for predicting patient survival. We saw no difference in ability to predict patient survival with the imputed data, though imputed treatment variables were more important to survival models.

To deal with the large number of features in RNA expression data, we used two approaches: using all the data with High Performance Computers, and transforming the data into a smaller set of features (sparse principal components, or SPCs). We compared the performance of survival models with both datasets and saw no differences. However, the SPC models trained more quickly while also allowing us to pinpoint the biological processes each SPC is involved in to inform future biomarker discovery.

We also examined ten processed molecular features for survival prediction ability and found some predictive power, though not enough to be clinically useful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Data generated for standard clinical oncology care has expanded exponentially. In addition to well-known clinical variables like symptoms, stage and histology, tumor specimens are now routinely sequenced for a range of mutations that may be more or less well characterized. These molecular profiles may suggest sensitivity to a range of molecularly targeted agents. Furthermore, high resolution, functional and molecular imaging methods (such as positron emission tomography-computer tomography (PET-CT) and magnetic resonance imaging (MRI)) create quantitative metrics described through radiomics features. These also suggest profiles that can guide intervention and response.

To facilitate the development of novel clinical decision support tools for oncologists, we have used publicly available data characterizing head and neck squamous cell carcinoma (HNSCC). These profiles present a large data analysis problem necessitating the use of machine learning, dimensionality reduction, and biological pathway analysis techniques. We utilize machine learning classifiers to predict patient two-year recurrence-free survival and evaluate a variety of feature sets to discover potential useful clinical biomarkers. Feature sets include combinations of patient clinical and molecular data. To improve utility of this dataset for oncological decision support, imputation and dimensionality reduction methods are used to transform feature sets to more usable and interpretable forms.

In this thesis, we investigate the prognostic ability of clinical variables before and after imputation procedures, RNA expression data, transformations of RNA expression data, and representations of exomic tumor variation. Ultimately, models trained on clinical data performed best, with imputed clinical data performing similarly to non-imputed data. Dimensionality reduction of RNA expression variables resulted in no significant changes in classifier performance, though was extremely helpful in reducing the necessary computation for training and evaluating models. A comparative analysis of clinical data and exomic tumor variation in the form of mutation significance data revealed that the while the mutation significance data does contain information predictive of patient survival, the clinical data is much more informative overall.

CHAPTER 2: BACKGROUND

Machine Learning

Machine learning (ML) is a computational field that uses algorithms to learn from existing data to discover relationships and build predictive models. Classification is a specific type of machine learning problem in which the algorithm is tasked with classifying data points based on how previous data were classified. Essentially, individual data points consist of a set of features and a class label. A classifier is trained on the training data, building a predictive model to classify new data points.

To evaluate the effectiveness of a predictive model, it must be tested on data that it has not “seen” before. One standard, systematic way of doing this is k-fold cross validation (CV), where the available data is split into k partitions, or folds. Models are trained k times using k-1 of the folds as training data, while the remaining fold serves as the testing data. The performance of the classifier is then evaluated based on how it classified points in the testing data during the validation runs.

Some classification algorithms utilize hyperparameters, parameters that are set prior to training that affect the training and classification procedures. In order to choose these values in an unbiased fashion, elaborations on k-fold CV can be applied. In this thesis, a nested CV procedure is employed to tune model hyperparameters. As in standard ten-fold CV, the data is split into ten folds (or partitions) for the outer CV. In each iteration, a single fold acts as the testing data and the remaining nine folds act as the training data. However, within each fold of the outer CV, a repeated grid search CV procedure (Krstajic et al. 2014) is carried out on the training data to estimate the best hyperparameter(s). Then, a model is trained on all of the training data with the best hyperparameter set, and its generalization performance is estimated using the testing data for that fold. The classifier's ability to accurately predict the class labels of test data points is then estimated from performance within the ten folds, using a classifier performance metric.

The area under the receiver operating characteristic curve (commonly denoted AUC) is the metric used to compare classifiers in this work. AUC is a very popular and commonly used classifier metric in the literature with an intuitive probabilistic interpretation: AUC is the probability that the classifier will score positive observations higher than negative observations. Mathematically, an AUC of 0.5 is equivalent to random guesses and is a standard baseline for this metric.

Missing Value Imputation

A common problem in machine learning is missing values: the absence of certain features for some cases. These gaps in knowledge can be problematic for many algorithms. One way to handle these data is to exclude entries with missing values. However, this means that fewer data points are available for classifier training and evaluation, which can negatively affect classifier generalization performance. To work around this issue, imputation methods are often employed to fill in or infer missing values (Ambler, Omar and Royston 2007). Simple imputation methods such as mean/median imputation, mode imputation, and value replacement can be uninformative or biased.

A more complex method, Multivariate Imputation by Chained Equations (MICE) (Groothuis-Oudshoorn 2011) minimizes bias by taking uncertainty into account during the imputation process, building multiple predictive models for each variable to be imputed. In the final step, the resulting multiple imputations are pooled together to produce a final imputation.

These procedures allow MICE to outperform single imputation methods (Zhang 2016, Ambler et al. 2007).

Sparse Principal Component Analysis

Principal component analysis (PCA) is a widely-used dimensionality reduction technique that calculates orthogonal linear combinations of the original features in an attempt to capture the maximum variance in the resulting variables. This often results in principal components that are linear combinations of all original variables, reducing the interpretability of individual components.

Sparse principal component analysis (SPCA) follows a similar procedure, though places a sparsity constraint on the resulting components (Zou, Hastie and Tibshirani 2006). This limits the number of original variables that can contribute to each sparse principal component (SPC). In the case of gene expression analysis, this allows biological interpretation of SPCs by performing gene ontology enrichment analysis on the list of genes that contribute to each computed feature.

Gene Ontology Enrichment Analysis

For the better part of two decades, the Gene Ontology project (The Gene Ontology Consortium et al. 2000, The Gene Ontology Consortium 2017) has been compiling a comprehensive resource of computable knowledge known as the Gene Ontology (GO). For a multitude of organisms, they have organized gene annotations regarding biological processes and biochemical activities that the gene or its gene products contribute to as well as the cellular components where the gene products are active.

Gene Ontology Enrichment Analysis (GOEA) is a technique where for a set of genes, their collective set of annotations are examined for terms that show up more often than expected by random chance, as determined by statistical tests. The resulting enriched GO terms allow interpretation of biological and biochemical context for the set of genes.

The PANTHER annotation database (Mi et al. 2017) stores annotations for protein-coding genes from the completely-sequenced genomes of 104 species. PANTHER supports GOEA for human genes and provides the API used by the Gene Ontology Consortium (http://geneontology.org).

Existing HNSCC Literature

Current HNSCC literature often focuses on association of regulation of specific genes with prognosis (Wang et al. 2017, Liu et al. 2018). Other groups, however, acknowledge the need for large-scale integrative analysis to capture potential novel biomarkers (Huang et al. 2016, Krempel et al. 2018, Hu et al. 2017). In other cancers, unsupervised transformations of molecular data (e.g. RNA sequencing, DNA methylation, miRNA sequencing) are known to be useful in machine learning-based survival prediction (Kim et al. 2018, Chaudhary et al. 2018). As of this writing, little work has been done with HNSCC in this manner. Literature on application of TCGA-HNSC Broad Firehose data and machine learning imputation of sparse clinical data is similarly unavailable.

TCGA-HNSC Dataset

The analyses presented are in part based on data generated by the Cancer Genome Atlas Research Network (TCGA): https://www.cancer.gov/tcga. TCGA is a coordinated effort to gather, share, and analyze next generation molecular sequencing data to improve our understanding of cancer mechanisms on a molecular level (Grossman et al. 2016). Data utilized in our analysis were obtained from the NCI Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/) and contained 528 TCGA-HNSC cases, including genotyping, solid-tumor RNA expression, whole exome sequencing, methylation data, and clinical information. In this work, only RNA expression variables and clinical information are considered. Clinical data includes tumor grading information, patient demographic data, smoking/alcohol histories, and several features related to disease progression such as lymphovascular invasion and margin status. HPV status (based on ISH and P16 testing) was also included, as HPV status has strong implications for prognosis and tumor development (Bratman et al. 2016, Chakravarthy et al. 2016). These data have been contributed from a number of studies from varying institutions, utilizing multiple platforms and assays that span significant time intervals. The work presented here addresses the challenges presented by this common form of dataset in oncological research.

Large, multi-institutional datasets present a variety of challenges to the development of methods and tools for clinical decision support. Namely, several clinical data fields in TCGA-HNSC offered issues of sparsity and inconsistency. Out of fifteen identified clinical characteristics relevant to treatment regimen, none were populated for every patient. More specifically, the number of cases (from a total possible 528 cases in TCGA) with missing or unavailable data for these fields ranged from 88 to 504, with a mean of 349.5 and median of 342 cases lacking data for each field. This is illustrated in Figure 1, where a vast majority of the treatment features had values available for less than half of the patient cases.

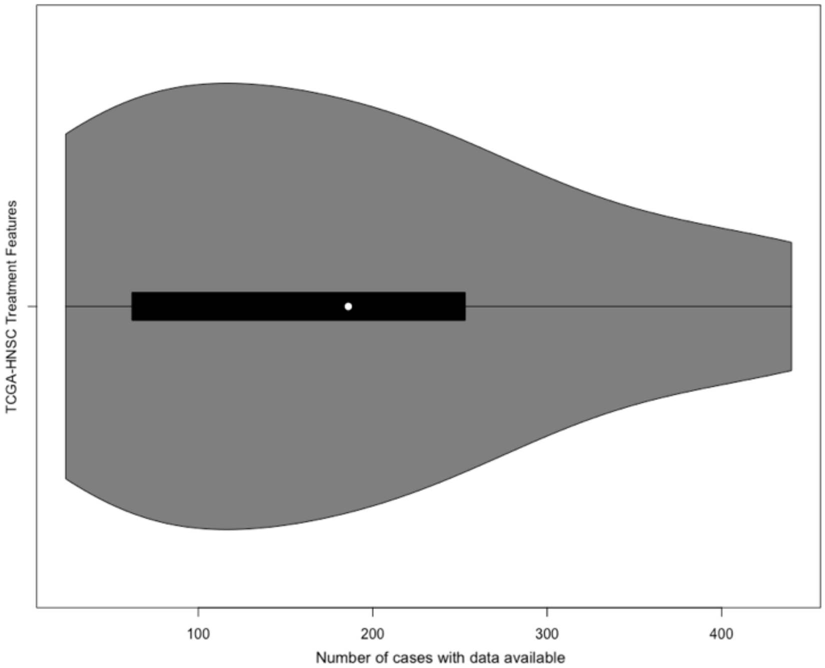

Figure 1: Violin Plot Describing Sparsity of Clinical Treatment Features

The distribution of the number of patient cases with available data for fifteen descriptors of treatment regimen.

In addition to the problem of missing data, several fields were populated inconsistently, with responses varying both due to human error (e.g. leading zeros in numeric fields) and varying convention (e.g. “External” vs. “EXTERNAL BEAM”). Such complications required extensive preprocessing and an expert system built using domain-specific knowledge to determine whether each patient had received a specific type of therapy. Even after preprocessing and condensing of treatment fields, issues of missing data persisted. Whether a patient had received radiotherapy and/or chemotherapy was unclear for 47% and 27% of cases, respectively. One possible technique for handling such problems is to exclude cases or variables with missing data, as was done previously with this dataset (Mroz, Patel and Rocco 2018). Due to the relevance of these features to our decision support goals, as well as the limited number of cases from which to draw, we attempt to maximize utilization of the available data by imputing missing values.

Molecular datatypes are often extremely high-dimensional. Feature selection and dimensionality reduction techniques are necessary steps when utilizing such data to best employ available computational resources. There are several strategies for selection and dimensionality reduction, including feature filtering, feature transformations, and wrapper methods such as sequential selection (Saeys, Inza and Larrañaga 2007). In this work, feature filtering and an unsupervised sparse PCA feature space transformation of 20,531 solid-tumor RNA expression variables were employed and evaluated in the context of TCGA-HNSC.

Broad Firehose Data

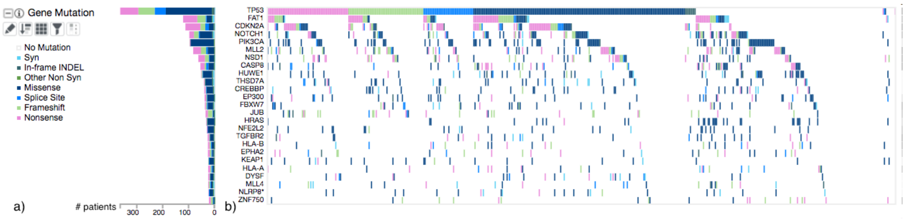

Figure 2: MutSig2CV Data Visualization

Portion (a) contains the legend and a histogram, the former listing the different types of mutations identified and the latter displaying the number of patients found with each type of mutation in each gene. Genes are vertically sorted by mutation incidence. Portion (b) displays the MutSig output itself, with each column representing a single patient in the TCGA-HNSC dataset, sorted based on TP53 mutation, followed by FAT1 mutation, then CDKN2A mutation, etc.

The Broad Institute's Genomic Data Analysis Center (GDAC) has created a large analysis pipeline, Broad GDAC Firehose, to systematically analyze data from TCGA. The results of these analysis runs are published as online interactive figures through "Firebrowse" (https://firebrowse.org/). One tool in this pipeline, MutSig2CV (or MutSig), analyzes annotated exomic tumor-normal variant data (Broad Institute TCGA Genome Data Analysis Center 2016).

For each patient in the dataset, MutSig2CV identifies genes that are significantly mutated above an expected baseline and reports the most deleterious disruptions to those genes (see Figure 2).

CHAPTER 3: APPROACH

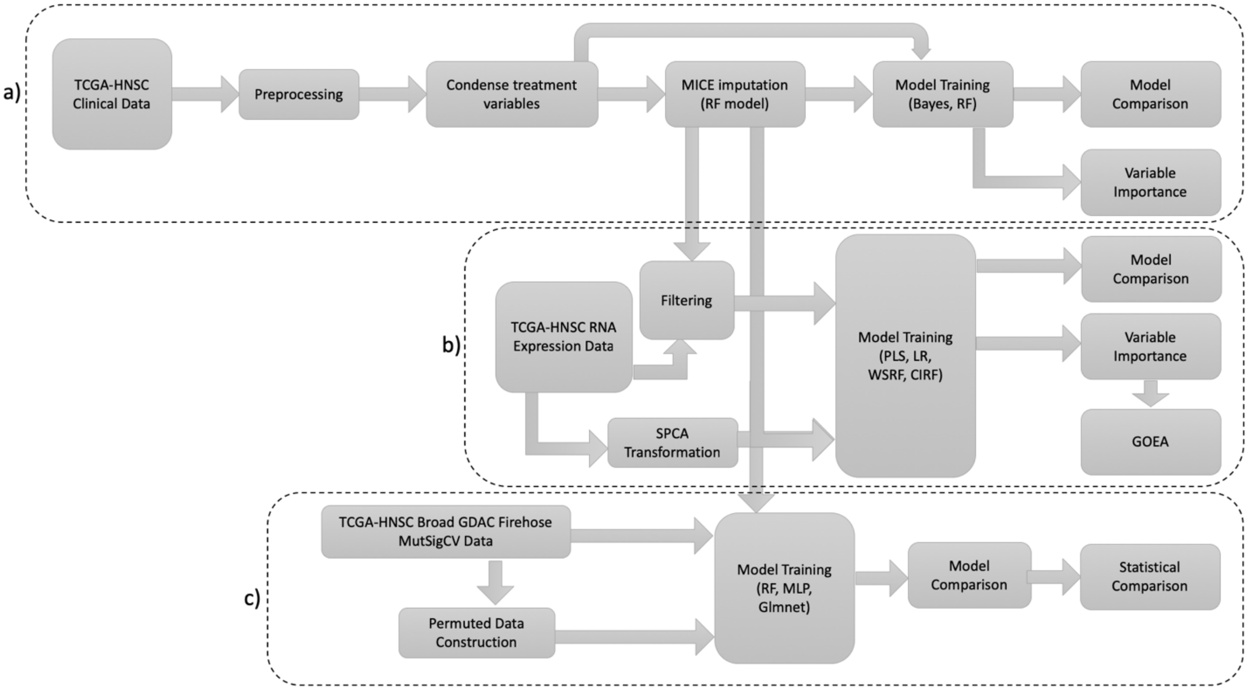
a)
TCGA-HNSC Clinical Data
Preprocessing
Condense treatment variables
MICE imputation (RF model)
Model Training (Bayes, RF)
Model Comparison
Variable Importance
b)
TCGA-HNSC RNA Expression Data
Filtering
SPCA Transformation
Model Training (PLS, LR, WSRF, CIRF)
Model Comparison
Variable Importance
GOEA
c)
TCGA-HNSC Broad GDAC Firehose MutSigCV Data
Permuted Data Construction
Model Training (RF, MLP, Glmnet)
Model Comparison
Statistical Comparison

Figure 3: Flow Diagram Outlining the Approach of this Work

Figure 3a) Clinical data preprocessing, imputation, and classification experiments. b) RNA expression preprocessing, classification experiments, and subsequent analyses. c) Evaluation of mutation significance features as a supplement to clinical prognostic models. Arrows between section a) and the other two sections indicate that clinical data (including imputed variables) are utilized in the models of sections b) and c).

Figure 3 illustrates a high-level view of the methods employed in this thesis. Section a) describes the processing and imputation of clinical data, as well as evaluation of the imputation method. Section b) details the filtering and transformation of RNA expression data followed by the model-building process and subsequent analyses. In section c), the investigation of MutSig2CV data for potential clinical biomarkers is depicted.

Preprocessing, Condensing, and Missing Data Imputation (Figure 3a)

Performing imputation on large datasets requires development of an expert model. Careful examination and correction of inconsistent and missing values was performed in collaboration with expert oncologist Dr. John Buatti (Chair of Radiation Oncology, University of Iowa). A rubric for consistent preprocessing and condensing of the fifteen relevant treatment fields resulted in a much more concise and usable dataset. However, a significant fraction of the TCGA-HNSC patients still had uncertain status in their treatment regimens. To address this, Multivariate Imputation by Chained Equations (MICE) (Groothuis-Oudshoorn 2011) was utilized, and the resulting changes to classifier performance were measured.

MICE builds predictive models for each missing variable to realistically impute entries based on the remaining predictors. For each missing entry, five intermediate imputations were performed using random forest models. Imputation models were trained using the clinical characteristics listed in Table 1, excluding the Surgery, Chemotherapy, and Radiation Therapy features. Patient outcomes were also excluded from the imputation to prevent information leakage. As the imputed treatment variables are binary, a majority vote was conducted of the five imputations to yield a final imputation. Using the imputed variables, the clinical characteristics utilized in imputation, and the outcome of two-year recurrence-free survival, two types of model were trained on the pre-imputation and post-imputation datasets. Missing values in the pre-imputation set were assigned a third category, “Unavailable”.

Naïve Bayes (NB) and Random Forest (RF) classifiers were selected for this evaluation, trained on the 24 clinical features listed in Table 1. These two were chosen to cover two major types of classifier: a pure conditional-statistical effort to predict survival (the Bayesian model), and a recursive partitioning-focused model (the Random Forest). NB does not consider interaction effects, whereas the RF model extensively leverages them. Earlier work suggests RF models are effective for this classification problem (Rendleman 2017). The pre- and post-imputation models were compared with respect to both predictive performance and variable importance.

Table 1: List of Clinical Data Features

Age	Alcohol consumption per day	Organ of Tumor Origin
Ethnicity	Lymphovascular Invasion	Margin Status
Tobacco Smoking History	Tobacco Pack-Years Smoked	Perineural Invasion
Inferred HPV Status	Extracapsular Spread	Smokeless Tobacco Average per day
Gender (sex)	Race	Tumor Grade
AJCC Pathologic Nodes (PN)	AJCC Clinical Tumor (CT)	AJCC Clinical Nodes (CN)
AJCC Clinical Metastasis (CM)	AJCC Pathologic Metastasis (PM)	AJCC Pathologic Tumor (PT)
Radiation Therapy (imputed or non-imputed)	Chemotherapy (imputed or non-imputed)	Surgery

RNA Expression Experiments (Figure 3b)

In the TCGA-HNSC dataset, solid-tumor expression was available for 520 of the 528 patients. With a feature set of 20,531 solid-tumor RNA expression variables, seven tumor grading variables, and the random forest-imputed treatment variables, several RF classifiers were trained to predict two-year recurrence-free survival. The classifiers varied in feature sampling and tree construction procedures: a standard RF, a weighted subspace RF (WSRF) (Zhao, Williams and Huang 2017), and a conditional inference random forest (CIRF) (Strobl et al. 2007). The WSRF weights randomly sampled variables based on their correlation with the output procedure, increasing the probability that a given tree will sample variables with high univariate correlation to patient survival. The CIRF utilizes a conditional inference procedure for tree construction that aims to eliminate bias in recursive partitioning and reduce computation time with stopping criteria.

With the full set of RNA expression data, the feature set was first refined through two filters: a univariate, near-zero variance filter to remove uninformative features and a multivariate correlation filter to remove features with correlation greater than 0.9. These filters reduced the feature set from 20,531 gene expression variables to approximately 18,000.

In addition, a dimensionality reduction was performed via Sparse Principal Component Analysis (SPCA) (Zou et al. 2006) which has the potential to improve interpretability of the model and reduce training time. Interpretability is improved because each sparse principal component (SPC) has only a handful of genes that contribute to it, allowing connections to be drawn between individual SPCs and the biological processes related to their constituent genes. One significant problem of PCA-based data reduction is choosing the number of components. If too many components are retained, this transformation may be amplifying noise. If too few are included, valuable predictive information may be excluded. To estimate information inclusion, percent explained variance is examined in Figure 4. Here, we chose the number of principal components to be ten, as this number of components yielded the best classifier performance over the three RF classifiers while explained approximately 90% of the variance. The resulting ten SPCs (below labeled X1-X10) were constructed and the feature set supplemented with the same grading and treatment features as used with the full set of RNA variables. The same set of RF classifiers was trained on this data to predict two-year recurrence-free survival.

After training, variable importance for the 10-component SPCA feature set was evaluated for each of the classifiers. The genes that comprise the most and least important variables were examined with a gene ontology enrichment analysis (GOEA) (The Gene Ontology Consortium et al. 2000, The Gene Ontology Consortium 2017). Analysis was conducted with the biological process annotation dataset from the PANTHER Classification System (ontology database released 2019-02-02), using their enrichment analysis tool (Mi et al. 2017). Enriched terms for high- and low-importance SPCs are identified and compared, and processes associated with high-importance SPC gene sets but not associated with low-importance SPC gene sets are described.

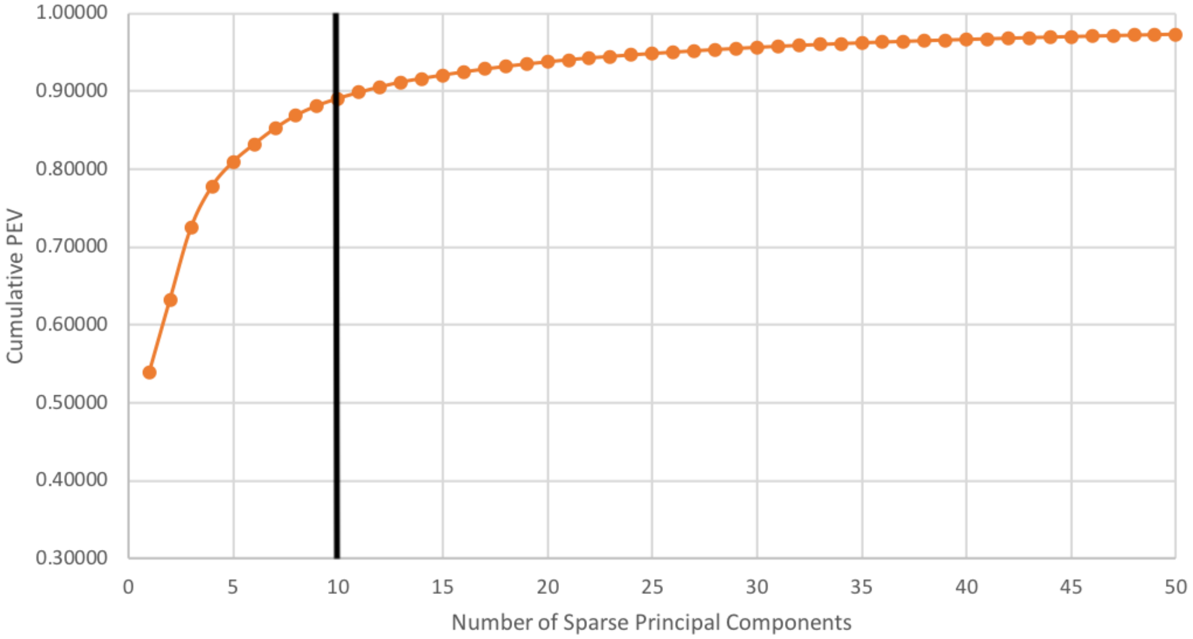

Figure 4: SPCA Percent Explained Variance

Cumulative percent explained variance is reported as it relates to the number of sparse principal components. The black vertical line indicates the value used for transforming RNA expression into the SPCA feature set for these experiments.

Mutation Significance Experiments (Figure 3c)

Comparison of Classifiers Trained on Clinical and MutSig Data

Classifiers were trained on datasets consisting of the ten gene-based MutSig features and also a combination of the imputed clinical data and MutSig features. The full list of clinical and mutation significance features used in this thesis can be found in Tables 1 and 2, respectively. The legend for Table 2 also describes some feature selection and preprocessing of the mutation significance features. The classifier performance metric for these each of the classifiers is reported.

Table 2. List of MutSig2CV Genes

The ten most-populated genes were chosen from the MutSig2CV data. All gene features other than TP53 were considered too sparsely-populated for informative categorical distinction (77%-93% with “No Mutation”), and as such were collapsed into binary variables (“Mutation” or “No Mutation”) to prevent overfitting. The TP53 mutation significance feature itself exhibits a distribution within which individual categories are relatively well-populated, with only 32% of examples having “No Mutation”.

TP53	FAT1
CDKN2A	NOTCH1
PIK3CA	MLL2
NSD1	CASP8
HUWE1	THSD7A

Comparison of Classifiers Trained on MutSig Data and Shuffled MutSig Data

To examine the mutation significance data more directly, the mutation significance features were permuted by shuffling values randomly within each feature. These shuffled mutation significance features were then used both alone and with the actual clinical features to train classifiers. A total of ten shuffled mutation significance datasets were created and used in this manner. The mean and standard deviation of the classifier metrics for these classifiers were calculated, and one-sided t-tests were performed with the null hypothesis that the classifiers trained with the actual mutation significance data were not more predictive than the classifiers trained with randomly-shuffled mutation significance data. This experimental design results in two comparisons of the MutSig data with the shuffled data, one including the clinical features and one considering only the MutSig features.

T-values were calculated with $t = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$, where $\bar{x}$ is the classifier performance for the classifiers trained with actual mutation significance data, μ_0 is the mean of the metrics calculated for the classifiers trained with shuffled mutation significance data, σ is the standard deviation of the shuffled metrics, and n is the number of samples (in this case 10). P-values were obtained for the one-sided significance tests with the Microsoft Excel function TDIST.

Classifier Training and Hyperparameter Tuning

A nested cross validation (CV) is used to tune hyperparameters and estimate out-of-sample performance, a measurement of how well a classifier would generalize if it were to be trained on the entire dataset. Figure 5 further describes nested CV. Classifier performance is measured using AUC.

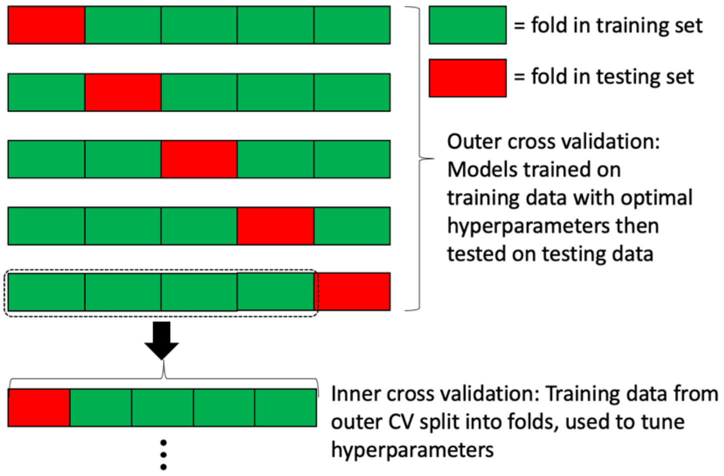
🟩 = fold in training set
🟥 = fold in testing set
Outer cross validation: Models trained on training data with optimal hyperparameters then tested on testing data
Inner cross validation: Training data from outer CV split into folds, used to tune hyperparameters

Figure 5: Diagram of Nested Cross Validation

A five-fold nested CV is depicted. Before training a model on the training data in each iteration of the outer CV, an inner CV is performed on the training set with a repeated grid-search to tune the optimal hyperparameters. These hyperparameters are then used to train the classification model in the outer CV and generalization performance is estimated using the testing set.

For the missing data imputation, models were trained and tested in Weka 3.9.1 (Frank, Hall and Witten 2016) with ten-times repeated ten-fold cross validation as the internal CV procedure using the “CVParameterSelection” wrapper method. Repeated CV reduced bias due to random partitioning (Borra and Di Ciaccio 2010). One hyperparameter was tuned for the Random Forest, the number of randomly chosen predictors to be considered for each split.

In the RNA expression experiments, three different RF classification procedures are considered. Classifier training and evaluation was handled using the R package caret (Kuhn 2015). Classifiers trained on the full-RNA data were evaluated with the internal CV procedure as ten-fold CV, and those trained with the dimensionality-reduced data were evaluated (within each fold) using five-times repeated ten-fold CV. Preprocessing was handled within each CV iteration with the R recipes package (Kuhn and Wickham 2018). For these RF models, the number of randomly-sampled predictors for each tree was varied over a span of values appropriate for each feature set.

In the MutSig analyses, Naïve Bayes and Random Forest classifiers were trained on the datasets using Weka 3.9.1 and a ten-times repeated ten-fold CV internal CV procedure, performing nested CV with the “CVParameterSelection” wrapper method. As in the imputation experiment, only the one hyperparameter for the Random Forest model was tuned.

Variable/Feature Importance

Feature importance is a measurement of how perturbations to variables affect classifier performance. A conditional variable importance procedure has been applied in this work. Conditional importance involves not only univariate perturbations, but conditional perturbation of variables and the variables with which they correlate (Strobl et al. 2008). For the imputation experiments, the correlation threshold was set at 0.2 for computational viability. In analysis of SPCA variables, this threshold was set to 0.05 as the feature space is smaller. Importance of categorical variables can also be biased in this scenario (depending on the number of categories), so a conditional inference random forest model is used to reduce this bias (Strobl et al. 2007). Reported importance values are relative to the most important variable in each case and were averaged over 50 runs to ensure stability.

CHAPTER 4: RESULTS

Imputation Evaluation

As shown in Table 3, imputation of treatment fields using MICE yielded no significant change in AUC. Changes in relative importance values can be seen in Table 4.

Table 3. Effect of Imputation on Classifier Performance Using the Imputed and Non-imputed Datasets

Classifier	Dataset	AUC
Naïve Bayes	Pre-imputation	0.633 ± 0.077
	Post-imputation	0.675 ± 0.063
Random Forest	Pre-imputation	0.668 ± 0.062
	Post-imputation	0.658 ± 0.075

Considering Figure 6, the relative importance of treatment features doubled as a result of imputation. Interestingly, changes were observed in non-imputed features as well, with some features (HPV status, margin status) becoming more important and others (Pathologic Tumor status grade, tumor grade, gender, ethnicity, alcohol consumption) dropping in importance.

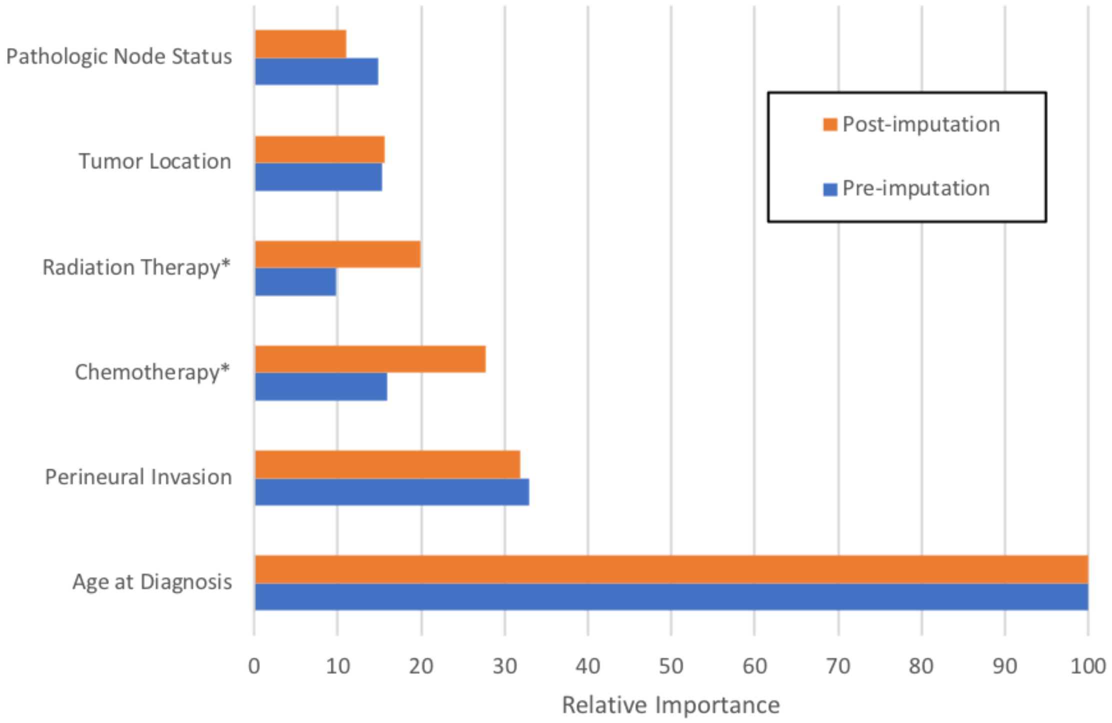

Figure 6: Effect of MICE Imputation on CIRF Conditional Variable Importance When Predicting Two-Year Recurrence-Free Survival

Importance values are relative to the most important variable. Imputed treatment features are denoted with *, and several common prognostic clinical variables are shown for comparison.

RNA Expression Experiments

Table 4 shows that classifier performance was slightly higher (though not significantly so) with the dimensionality-reduced dataset. The best-performing classifier overall appears to be the CIRF, which was middling in runtime. A drastic difference in evaluation runtime is observed (as expected) between the Full RNA feature set (20,541 predictors) and the SPCA feature set (20 predictors). With both feature sets, the non-standard RF variants required more compute time and computational resources than the standard RF classification procedure.

Table 4: Classifier Performance with Full RNA Expression Data and SPCA Features

AUC and approximate runtime values for the RNA expression feature sets. The best value for each row is bolded. Here, runtimes are evaluation times for a given classifier on a given feature set via ten-fold nested cross validation with the internal cross validation procedures as described in Chapter 3. Computations performed on the University of Iowa's Argon High-Performance Computing Cluster.

Datasets Classifiers:	RF	WSRF	CIRF
AUCs			
Full RNA	**0.632 ± 0.106**	0.596 ± 0.038	0.629 ± 0.105
SPCA	0.640 ± 0.128	0.626 ± 0.114	**0.658 ± 0.044**
Nested CV Runtimes	---		---
Full RNA	**52 hr**	185 hr	85 hr
SPCA	**12 min**	1.9 hr	30 min

Table 5: Percent Explained Variances for the Sparse Principal Components

The 10 SPCs account for 89.05% of the original data's variance. * denotes SPCs chosen for further analysis based on variable importance (see Figure 7).

SPC	Percent Explained Variance
X1*	53.84%
X2*	9.43%
X3*	9.19%
X4	5.31%
X5	3.14%
X6*	2.27%
X7*	2.04%
X8	1.67%
X9*	1.24%
X10	0.93%

Considering Figure 7, SPC X6 is favored most by the conditional inference importance metric. SPCs X9 and X2 are the next-highest ranked. X7, X1, and X3 were the least important variables to the CIRF classifier, indicating they had little-to-no effect on classification performance. The genes composing these six SPC features were selected for further examination via GOEA (see Table 6). It is worth noting that within the gene sets constituting the SPCs, many repeats of genes and gene families are present. This is an artifact of gene family co-expression, and the tendency of SPCA to focus on genes with high variance.

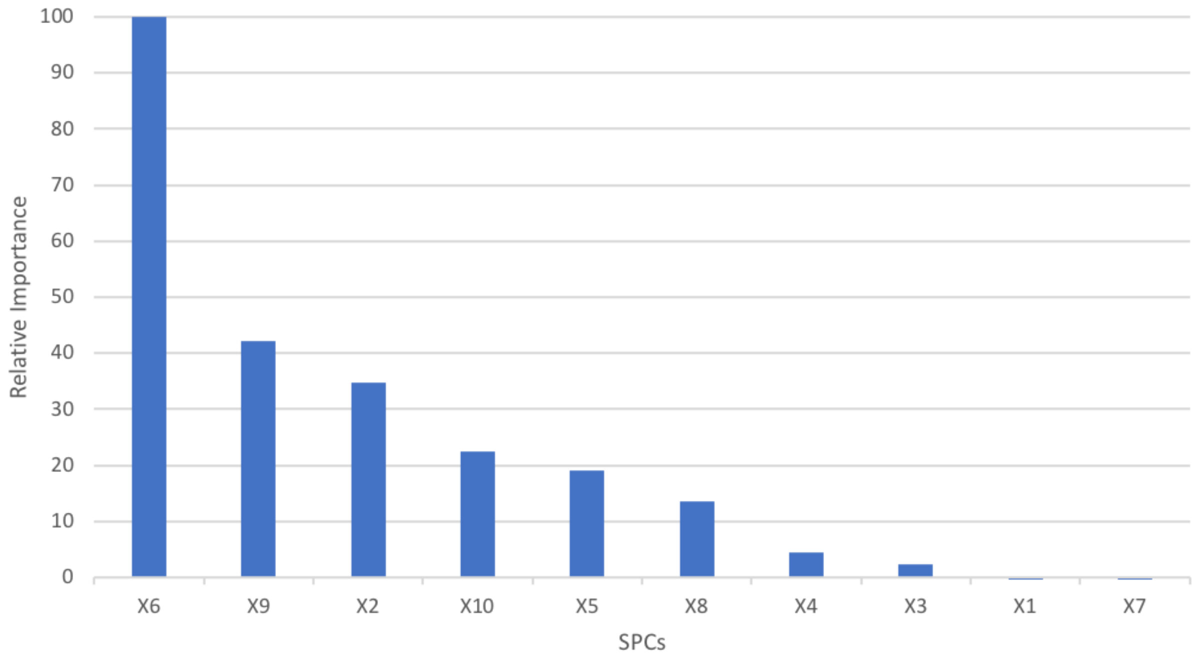

Figure 7: SPCA Relative CIRF Conditional Variable Importance

Importance values for the 10 SPCs, labeled X1-10, are reported. In cases where no importance is reported for an SPC, its effect on classifier performance is negligible.

Table 6: SPCA Gene Set Enriched Gene Ontology Terms

For each SPC gene set, gene names and Entrez gene IDs are listed alongside the PANTHER annotation terms found to be enriched in the corresponding gene set are reported. Genes are listed in order of increasing order of p-value, with all p < 0.001. * indicates that some lower level hierarchical GO terms were omitted for brevity.

SPC	Contributing Genes (Gene Name|GeneID)	Enriched GO Biological Processes
X6	ADAM6|8755, FBP4|2167, FN1|2335, GAPDH|2597, KRT13|3860, KRT16|3868, KRT17|3872, LOC96610|96610	Cornification
X2	COL1A1|1277, COL1A2|1278, COL3A1|1281, FN1|2335, KRT13|3860, KRT14|3861, KRT16|3868, KRT17|3872, KRT5|3852, KRT6A|3853, SPARC|6678	Cornification*, keratinocyte differentiation, wound healing, cell-substrate junction assembly*, collagen fibril organization*
X9	ACTB|60, ADAM6|8755, COL1A1|1277, COL1A2|1278, FN1|2335, LAMC2|3918, TGFBI|7045	Skin morphogenesis, protein heterotrimerization, platelet activation*, cell junction assembly, cell junction organization, extracellular matrix organization, extracellular structure organization, blood vessel development*, cell adhesion
X7	ADAM6|8755, FABP4|2167, KRT16|3868, KRT17|3872, KRT5|3852, KRT6B|3854, LOC96610|96610, PI3|5266	Cornification*, programmed cell death, cell death, keratinization, skin development
X1	KRT14|3861, KRT16|3868, KRT17|3872, KRT5|3852, KRT6A|3853, KRT6B|3854, KRT6C|286887, S100A9|6280	Cornification*, intermediate filament cytoskeleton organization*, cell death, hair cycle
X3	COL1A1|1277, COL1A2|1278, COL3A1|1281, KRT13|3860, KRT14|3861, KRT16|3868, KRT17|3872, KRT5|3852, KRT6A|3853, KRT6B|3854, KRT6C|286887, SFN|2810	Cornification*, multicellular organism development, intermediate filament cytoskeleton organization*, collagen fibril organization

Mutation Significance Experiments

Comparison of Classifiers Trained on Clinical and MutSig Data

From Table 7, the classifiers trained with the clinical data outperform those trained with only the MutSig data. Additionally, there is negligible change in classifier performance when the clinical features are supplemented with the MutSig features.

Table 7: MutSig2CV Experiment 1: Classifier performance (AUC) with the Clinical, MutSig, and Clinical + MutSig feature sets.

Dataset	Naïve Bayes	Random Forest
Clinical Only	0.675 ± 0.063	0.658 ± 0.072
MutSig2CV Only	0.573 ± 0.087	0.609 ± 0.084
Clinical + MutSig2CV	0.679 ± 0.069	0.660 ± 0.063

Comparison of Classifiers Trained on MutSig Data and Shuffled MutSig Data

In the second experiment, the comparison of classifiers trained with original MutSig data and the shuffled MutSig data in the presence of clinical data shows inconclusive results. The Bayesian statistical model showed a significant difference, but the Random Forest model results suggest that RF classifiers trained with the real data performed no better than those trained with the manufactured data. When this comparison was performed with the MutSig and shuffled data in isolation, the Naïve Bayes classifier showed around the same significance and the RF models showed that the real MutSig data significantly outperformed the shuffled data.

Table 8: MutSig2CV Statistical Test Results

Results of t-tests comparing performance of classifiers trained on original datasets to those trained with permuted mutation significance data. P-values where the null hypothesis was rejected are denoted with *.

Dataset	Naïve Bayes	Random Forest
Clinical + MutSig2CV	$*p = 4.84 * 10^{-5}$	$p = 0.306$
MutSig2CV Only	$*p = 4.63 * 10^{-5}$	$*p = 5.72 * 10^{-8}$

CHAPTER 5: DISCUSSION

Imputation Evaluation

With imputation, classifier performance is not negatively affected, which is expected based on other studies using the MICE imputation technique (Ambler et al. 2007, Deng et al.

2016). Increases in variable importance after imputation indicate that the treatment variables more effectively predict patient outcomes after application of MICE. Because importance is calculated with a random forest, the importance changes in non-imputed variables might indicate that MICE imputation of the treatment variables modifies the landscape of variable interactions to a high enough degree that feature selection within trees is affected.

RNA Expression Experiments

For this prediction problem, the dimensionality-reduced features (SPCs) allow comparable classifier performance while drastically reducing runtime and necessary computation. Though not quantified here, memory requirements were also much lower for the dimensionality-reduced data. Additionally, this reduction allowed us to identify gene set candidates for GOEA. In both important and not-important SPCs, the GO term “cornification” (a form of cell death in squamous epithelial cells) is found, indicating that this biological process is related to high-variance genes in this dataset. Terms found only in the high-importance SPC gene sets are related to cell motility (cell adhesion, extracellular interactions), immune response, cell growth, and blood vessel development. Activity of genes involved in these processes could be indicative of a cancer’s ability to survive, grow, and metastasize, suggesting that these SPCA transformed RNA data contain useful information about underlying relationships between solid-tumor expression and two-year recurrence-free survival.

Mutation Significance Experiments

From Table 8, it is clear that the MutSig2CV data has some prognostic value. The results in Table 7, however, suggest that the predictive power in these features is significantly less than that found in the clinical data. Considering that the Naïve Bayes classifier found statistically significant differences though the RF classifier showed no difference between the models compared in the presence of clinical data, it appears that the shuffled data is more of a detriment to the Naïve Bayes model than in the RF model, which directly considers interaction effects. Therefore, it is likely that the information provided by the MutSig2CV features is also provided by the clinical data, as the model considering variable interactions was not significantly more predictive with real data than shuffled data.

CHAPTER 6: CONCLUSIONS

In modern oncological research, TCGA datasets present significant large data analysis challenges, from clinical parameter sparsity to high dimensionality. Facing these problems requires significant preprocessing and machine learning modeling to uncover new knowledge. A multivariate imputation method (MICE), SPCA dimensionality reduction, and an SPC-focused GOEA are presented in the context of TCGA-HNSC clinical and RNA expression variables to improve usability of data for future HNSCC decision support.

As others (Ambler et al. 2007, Deng et al. 2016) have found, MICE is an effective method for imputing data while introducing minimal bias. In this case, it improved the variable importance of imputed features while altering the importance of other variables through interaction effects. Most importantly, the imputation provided a complete set of treatment variables to incorporate into our models, furthering our ability to evaluate the effectiveness of potential biomarkers in later analyses.

Unsupervised transformation of RNA expression data via SPCA was extremely useful in improving interpretability of survival models and biomarker identification by limiting the number of genes contributing to each principal component and allowing for a nuanced examination of the underlying biological processes. The biological processes found to be associated with high-importance SPCs may be useful in future feature selection for biomarker discovery. Additionally, the SPCA functioned well as a dimensionality reduction technique, as the dimensionality reduced features allowed for significantly lower computation time without significantly affecting classifier performance. From the literature and these analyses, unsupervised transformations of RNA expression data seem a viable option for future integration of molecular data into HNSCC clinical predictive models.

The output of MutSig2CV from the Broad GDAC Firehose pipeline was examined as a potential feature set for improving predictive power of prognostic models. It appears that while the data has value for predicting two-year recurrence-free survival, this value is either already present in the supplied clinical data or simply dwarfed by the predictive value in the clinical features. Therefore, the mutation significance features, in their current state with these methods, cannot be utilized to improve clinical prognosis estimation.

Future work will consider the effect of clinical imputation on models also utilizing molecular data, both with SPCA transformations and other unsupervised feature transformations methods such as denoising autoencoders. Additionally, biomarker evaluation will be expanded to directly consider right-censored survival.

REFERENCES

Ambler, G., R. Z. Omar & P. Royston (2007) A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical Methods in Medical Research,* 16, 277-298.

Borra, S. & A. Di Ciaccio (2010) Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Computational Statistics & Data Analysis,* 54, 2976-2989.

Bratman, S. V., J. P. Bruce, B. O'Sullivan & et al. (2016) Human papillomavirus genotype association with survival in head and neck squamous cell carcinoma. *JAMA Oncology,* 2, 823-826.

Broad Institute TCGA Genome Data Analysis Center. 2016. Mutation Analysis (MutSig 2CV v3.1). Broad Institute of MIT and Harvard.

Chakravarthy, A., S. Henderson, S. M. Thirdborough, C. H. Ottensmeier, X. Su, M. Lechner, A. Feber, G. J. Thomas & T. R. Fenton (2016) Human Papillomavirus Drives Tumor Development Throughout the Head and Neck: Improved Prognosis Is Associated With an Immune Response Largely Restricted to the Oropharynx. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 34, 4132-4141.

Chaudhary, K., O. B. Poirion, L. Lu & L. X. Garmire (2018) Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clinical Cancer Research,* 24, 1248.

Deng, Y., C. Chang, M. S. Ido & Q. Long (2016) Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data. *Scientific Reports,* 6, 21689.

Frank, E., M. A. Hall & I. H. Witten. 2016. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques". Morgan Kaufmann.

Groothuis-Oudshoorn, K. (2011) mice: Multivariate Imputation by Chained Equations in R. *2011,* 45, 67.

Grossman, R. L., A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe & L. M. Staudt (2016) Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine,* 375, 1109-1112.

Hu, S., H. Yuan, Z. Li, J. Zhang, J. Wu, Y. Chen, Q. Shi, W. Ren, N. Shao & X. Ying (2017) Transcriptional response profiles of paired tumor-normal samples offer novel perspectives in pan-cancer analysis. *Oncotarget,* 8, 41334-41347.

Huang, H., C. Lin, C. Yang, C. Ho, Y. Chang & J. Chang. 2016. An Integrative Analysis for Cancer Studies. In *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, 387-389.

Kim, S. Y., T. R. Kim, H.-H. Jeong & K.-A. Sohn (2018) Integrative pathway-based survival prediction utilizing the interaction between gene expression and DNA methylation in breast cancer. *BMC medical genomics,* 11, 68-68.

Krempel, R., P. Kulkarni, A. Yim, U. Lang, B. Habermann & P. Frommolt (2018) Integrative analysis and machine learning on cancer genomics data using the Cancer Systems Biology Database (CancerSysDB). *BMC Bioinformatics,* 19, 156.

Krstajic, D., L. J. Buturovic, D. E. Leahy & S. Thomas (2014) Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics,* 6, 10-10.

Kuhn, M. (2015) Caret: classification and regression training. *Astrophysics Source Code Library*.

Kuhn, M. & H. Wickham. 2018. recipes: Preprocessing Tools to Create Design Matrices.

Liu, J., Y. Wu, Q. Wang, X. Liu, X. Liao & J. Pan (2018) Bioinformatic analysis of PFN2 dysregulation and its prognostic value in head and neck squamous carcinoma. *Future Oncology*.

Mi, H., X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang & P. D. Thomas (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*.

Mroz, E. A., K. Patel & J. W. Rocco (2018) TCGA Data on Head and Neck Squamous Cell Carcinoma Suggest Therapy-Specific Implications of Intratumor Heterogeneity. *International Journal of Radiation Oncology*Biology*Physics,* 100, 1309.

Rendleman, M. 2017. Technical Report 2017-01: Clinical and molecular feature evaluation with TCGA-HNSC. Center for Bioinformatics and Computational Biology, University of Iowa.

Saeys, Y., I. Inza & P. Larrañaga (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics,* 23, 2507-2517.

Strobl, C., A.-L. Boulesteix, A. Zeileis & T. Hothorn (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics,* 8, 25.

Strobl, C., A. L. Boulesteix, T. Kneib, T. Augustin & A. Zeileis (2008) Conditional variable importance for random forests. *Bmc Bioinformatics,* 9.

The Gene Ontology Consortium, M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin & G. Sherlock (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics,* 25.

The Gene Ontology Consortium (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*.

Wang, L., Y. Jia, Z. Jiang, W. Gao & B. Wang (2017) FSCN1 is upregulated by SNAI2 and promotes epithelial to mesenchymal transition in head and neck squamous cell carcinoma. *Cell Biology International,* 41, 833-841.

Zhang, Z. (2016) Multiple imputation with multivariate imputation by chained equation (MICE) package. *Annals of translational medicine,* 4, 30-30.

Zhao, H., G. J. Williams & J. Z. Huang (2017) wsrf: An R Package for Classification with Scalable Weighted Subspace Random Forests. *Journal of Statistical Software,* 77.

Zou, H., T. Hastie & R. Tibshirani (2006) Sparse principal component analysis. *Journal of computational and graphical statistics,* 15, 265-286.