– Independent Work Report Spring 2023 –

Investigating the Effects of Racial/Ancestral Bias in The Cancer Genome Atlas Breast Cancer Dataset on Classification and Survival Prediction

Andres Blanco Bonilla
Adviser: Olga Troyanskaya

May 1st, 2023

Abstract

It is imperative that we understand the effects of White/European biased data on machine learning (ML) models. Using the The Cancer Genome Atlas (TCGA) breast cancer dataset, I create a variety of machine learning tumor classifiers and survival predictors in R, trained and tested on the genomic data of different subsets of the population based on patients' genetic ancestry, and evaluate the predictive power of the models on a number of performance metrics. Specifically, I try to answer whether ancestry-specific models outperform others. Genetic ancestry appears to have little to no impact on classifier performance, and ancestry-specificity seems to lowers survival prediction performance. More research is needed to assess the feasibility of "-omics"-based, ancestry-specific ML models.

1. Introduction

1.1. Motivation and Goal

While the innovation of machine-learning (ML) shows great promise for use in the medical field, unfortunately, the technology may be limited by biased datasets that overwhelmingly feature data from White European individuals, such as The Cancer Genome Atlas (TCGA) dataset [1]. I aim to investigate more precisely what the effects are of this racially skewed data on machine learning classification accuracy. Specifically, the goal of my project is to calculate and discuss the changes in different prediction metrics when tumor classifiers, and ML survival predictors, are trained and

tested on RNA-Seq and CNV datasets with different racial/ancestral compositions. In particular, I study the TCGA breast cancer dataset, and wish to understand the performance of specific classifiers, trained and tested only on samples with certain genetic ancestry. It is important to analyze how unbalanced datasets can affect machine learning models before they become a widespread and common tool of medicine, and how we may be able to improve our models. Biased classifiers run the risk of further discrimination against vulnerable minorities, leading to worsening healthcare for groups who already have worse health outcomes. A survival classifier may achieve high accuracy when tested on the majority White/European population, but may yield incorrect predictions of survival time when used on a Black patient, and this is quite troublesome, because early risk assessment is so important for cancer treatment. People of color who are high-risk cancer patients may be erroneously misclassified and thus not receive the additional treatment they desperately need, which would be devastating to their health [2][3]. We must understand what kinds of tools are borne out of biased datasets, and learn how to best create machine learning tools that can serve not just White/European patients, but any patient in need.

1.2. Background and Related Work

One of the most infamous examples of algorithmic racial bias is the failure of classifiers to predict skin cancer on dark skin when trained on images of mostly White skin lesions. Researchers have found that prediction accuracy is halved when a classifier like this is used on darker skin [2]. This problem, though, goes beyond classification errors due to visually different skin tones, and unfortunately even a classifier based on "-omics" data, if "only trained with genetic data of white patients, it may fail to generalize to patients of other ethnicities," [2]. Genetic profiles and genomic aberrations of cancer have indeed been found to differ along racial lines (or possibly between groups of people with different genetic ancestry) [4][5]. In particular, a 2018 analysis of TCGA data, which also sought to further our understanding of racial disparities in cancer care, revealed that "breast, head and neck, and endometrial cancers of African Americans (AA) have higher levels of chromosomal instability than those of European Americans," [4]. The paper first uses a classifier

to sort TCGA patients into genetic ancestry groups, rather than group by patients' self-reported race, and their labelings are publically available at the TCGAA (The Cancer Genetic Ancestry Atlas) [9]. The authors then use various TCGA "-omics" data to examine the frequency of genetic alterations between African Americans and European Americans, noting that "these alterations in cancer genomes dominantly and intrinsically influence the transcriptional phenotypes of cancers" [4]. A natural extension of this work, it seems, is to investigate how the variations in the cancer transcriptome may impact class prediction that is based on these cancer transcriptomics of TCGA individuals (that is, the individuals' RNA-Seq data, which measures expression levels of their genes). The authors also find that find that African Americans have "significantly higher" [4] levels of somatic-copy number alterations/variations (CNA/CNV), and somatic copy number alterations of tumors and tissues are one of the many forms of omics data available through TCGA, motivating the use of ML models trained on this data, in addition to RNA-Seq models.

Another related paper, which I am drawing heavy inspiration from, notes the racial/ethic differences in breast cancer and experiments with training/testing survival machine learning on varying subsets of data, in order to "examine the feasibility of race/ethnicity- specific ML models that may outperform the general model trained with all races/ethnicity" [3]. The authors focus on White, Black, and Hispanic patients, and use US National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) data (with variables like total number of tumors, tumor grade, age at diagnosis, etc), to train classifiers to predict patients' survival time in months. They conclude that race-specific ML approaches are more accurate and conducive to better care for underrepresented patients [3]. This work is very similar, but I use TCGA molecular data, specifically RNA-Seq and CNV data, to make predictions, rather than predict based on clinical/statistical SEER data.

There is ongoing debate about whether race is socially constructed or whether it is biological, and whether it is a valid genetic marker that has use in the medical field. Self-reported race is often used as a proxy for genetic ancestry, which refers to the geographic human population group that someone biologically descends from, but they are not exactly the same, and often do not map on to

each other smoothly or accurately. There is evidence that socially constructed races do not exhibit genetic similarity, and that purported racial differences actually capture geographical differences along the lines of genetic ancestry [6][7]. To clarify, lack of genetic ancestral diversity in datasets is still a pressing issue, perhaps even more-so than racial diversity due to the biological significance of ancestry [6]. I will henceforth refer to genetic ancestry, or just ancestry, in this paper.

Researchers have explored the genomic differences of cancer between groups of patients, and suspect that this may be a pitfall for ML models, if those models are based on insufficiently diverse data. Other researchers have further explored these challenges using ML trained on racially biased data such as cancer images or statistics and clinical data. I plan to, empirically, show how biased "-omics" datasets affects the accuracy and performance of classification and survival prediction, and gauge the possible benefits of a ancestry-specific ML models for cancer analysis.

2. Approach

The main idea is to expand upon the research done by [3], with some modifications. My approach is to train ML models using genomic data from the TCGA dataset, rather than clinical data from the SEER dataset, and evaluate the changes in performance when these models are trained and tested on different groups of patients. As discussed previously, the genetic makeup of cancer patients' samples may differ along ancestral lines, so it is instructive and worthwhile to discern if the benefits purported by the authors of the aformentioned paper "Evaluation of race/ethnicity-specific survival machine learning models for Hispanic and Black patients with breast cancer" [3] exist on a genetic level as well.

In addition, I group patients by their genetic ancestry, rather than their race. Because many researchers believe that genetic ancestry more accurately captures humans' underlying biology than self-identified race might (though it is a heated topic)[6][7], I think research into variations between human population groups should reflect this. Grouping by ancestry also provides more data to train and test on, because many TGCA patients' races are not reported.[1]

Lastly, I train not just survival predictors but also tumor classifiers divided in varying ways by ancestry. Tumor classification is a simpler prediction than survival time, but it still has use and merit in the medical field, so I examine both, as it may be the case that ancestry-specificity helps resolve one problem but not the other.

3. Implementation

The majority of this work was done in R. R [8] is built for data analysis, and there are many existing packages meant for working specifically with TCGA data, so this was a natural choice. There are three main steps to the R workflow: data processing, classification, and survival prediction.

3.1. Data Processing

3.1.1. RNA-Seq and Survival Data TCGA genomic data needed to be downloaded and aggregated together with TCGAA ancestral data into a usable format before any analysis could proceed. Unfortunately, any attempts to download files from the TCGA portal, or the recommended GDC Data Transfer Tool, and extract genomic and clinical data caused my RStudio to run out of memory and subsequently crash. Instead, the RTCGA [10] package in R was used to obtain preprocessed TCGA RNA-Seq and survival data. This includes the gene expression levels of every recorded gene for each sample, each patient's recorded survival time in days, and each patient's vital status. Importantly, the RTCGA package associates each TCGA patient with their patient/sample barcode, as opposed to the TCGA portal itself which marks downloaded files with a 36-character-long UUID. The TCGA UUID is especially inconvenient when integrating with others' TCGA research that identifies patients by TCGA barcode, such as the TCGAA.

3.1.2. Genetic Ancestry Data Unfortunately, the TCGAA data portal does not have any native option to download patients' ancestral data, and manually recording the ancestries of over 1000 individuals is not practical. To overcome this, I wrote a webscraping program using the Beautiful-

Soup [11] package in Python to identify each patient's TCGA barcode and genetic ancestry, and match them together in a csv file. The ancestries were read into R and merged by patient barcode with the survival and RNA-Seq data provided by the RTCGA package.

3.1.3. Classification Data I could not find a clear source that indicated if a given sample originated from a tumor or normal tissue, but this information is encoded in a sample's TCGA barcode [12]. With indexing starting at 1, the 14th and 15th characters of the barcode correspond to the sample type. Primarily of interest is that "01" encodes "Primary Solid Tumor", and "11" encodes "Solid Tissue Normal". "06" technically encodes a Metastatic tumor sample, but this is still a tumor, so a "06" sample in this paper is counted as Solid Tissue Normal for classification purposes. I matched the encoding substring of each sample's barcode to the appropriate type and added the type to the complete BRCA RNA-Seq dataframe. Unfortunately, only a fraction of tumor samples have RNA-Seq data available for corresponding normal tissue samples, which introduces a greater risk of overfitting the classifiers. In particular, out of 1121 total EA and AA samples, there are only 6 normal tissue samples from AA patients.

3.1.4. CNV Data While RNA-Seq data for normal tissue may be limited, the vast majority of patients in the TCGA dataset have CNV data available for both their tumor and normal tissue samples, including patients with African ancestry. Training another set of classifiers and survival predictors afforded additional protection against overfitting, and also provides more models with results to evaluate, compare, and discuss, improving the robustness of this study.

Processing TCGA CNV data was a much more lengthy process than expected, though, because the CNV data returned from the RTCGA package was structured completely unlike the RNA-Seq data, meaning I would either have to engage in heavy data processing to make the dataset compatible with the RNA-Seq analysis pipeline I had already written, or rewrite my existing code. The TCGA2STAT[13] package in R provides preprocessing of the CNV dataset into a clean gene by patient matrix, similar to the RNA-Seq data from RTCGA, but this package was removed from

the CRAN repository and had not been updated since 2016. After downloading the tar file for the package and building it locally, the package would not work, and returned poorly documented errors when trying to use it. The output of the package if it did work, though, was exactly the data I needed in the way I needed it. So, I decided to fix it myself, iteratively manually editing the R source script and rebuilding the tar file until no more errors were returned.

The main issue was that the code relied on webscraping the data portal using outdated XML tools, which I was able to resolve by updating all the XML usage to the appropriate xml2 methods. Once successful, this CNV data was merged with the ancestral, classification, and survival data to form the complete CNV dataset.

3.1.5. Data Subsetting Methodology When feeding the data to the ML models, I subset the data in 5 different ways, with each subset having a distinct purpose that allows for detection of possible patterns in machine learning performance, and clearer evaluation of the whole of the study's results. Note that when training and testing data are split 80:20, the split conserves a relative balance in both sets between the factors being predicted. For classification, this means that both the training and testing set will have an amount of normal tissues that is as even as possible with the amount of tumor tissue samples in that set. For survival prediction, it likewise means that both sets attempt to have a balanced proportion of censored individuals to dead individuals within the set.

All* Data means that the ML model has been trained on a random subset of 80% of all the data being used, and predicts on the remaining 20% of the data kept for holdout testing. Genetic ancestry is not taken into account when splitting training and testing data, providing a baseline of sorts for comparison to other ML models. However, "all" actually refers to all the African or European samples, but not literally all the samples, which is why there is asterisk. This work aims to evaluate ancestry-specific models, but there is not enough data from patients of other ancestries to build these models without blatant overfitting, so samples of patients of Asian, Native American, or other ancestries, are ignored. They are excluded as training or testing data for other models for the sake of simplicity and consistency.

AA Only means that the ML model was trained on a random subset of 80% of samples from African-ancestry patients, and tested on the other fifth. This is an African-specific model that aims to demonstrate the performance gains, or lack thereof, from using ancestry-specific machine learning.

Trained EA/AA and tested AA means that the training set for this ML model was composed of the same training data as the AA-specific model, in addition to the entirety of the European samples. The model is tested on the same test dataset as the AA-specific model, but the difference is it was trained on additional EA data. In conjunction with the AA-specific model, this model allows for direct comparison of the performance changes that may result from ancestry specificity. This subset was added because I realized the AA model performing better than the All* Data model may not necessarily mean that excluding EA data makes for better AA predictions. For example, suppose the All* Data model predicts every AA sample correctly, but erroneously predicts every EA sample, causing a significant drop in performance metrics for the whole model. The AA model may score better, but the absence of EA samples did not actually make a difference toward that performance. So, this model helps elucidate what is occurring when EA samples are added or removed, and if there are in fact resulting changes in performance toward AA samples.

EA Only means that the training and testing sets for this model were subset 80:20 from only European samples. The "Evaluation of race/ethnicity-specific survival machine learning models" paper did not consider an EA-specific model [3], but even though disadvantaged groups are more in need of improvements in healthcare and are the priority, I believe we should still investigate if we can improve the health of anyone and everyone using specific machine learning. If Black-specific and Hispanic-specific models yield enhanced performance for those groups, enhanced performance from European-specific models is a natural hypothesis to make.

Trained EA/AA and tested EA means that the same training and test set as the EA model were used, with the addition of. This complements the EA model and makes clearer whether the addition of AA samples improve, worsen, or do not change the performance of EA predictions.

Trained EA and tested AA means that the entirety of the European samples were used as training data, and the entirety of AA samples were the testing data. This models a hypothetical worse case scenario with a dataset completely biased toward Europeans, and if there are benefits to an AA-specific model, this model is expected to see a drop in performance. Note that the reverse model was not trained or tested, because this is not reflective of real-world problems. Scientists and researchers are concerned with White/European bias in datasets[6], so evaluating a hypothetical dataset entirely biased towards Africans is not realistic and is likely not a useful scenario.

3.1.6. Classification For basic binary classification, the support vector machine (SVM) [14] and glmnet[15] machine learning models were chosen. SVM was selected due to my existing familiarity with the svm method of the caret package in R, and because SVMs can fit high dimensional data, which is necessary because this data contains 20,000+ gene variables for every sample. Though, this was rather slow in practice, taking approximately 40 minutes to fit the largest model. glmnet was selected because I encountered an existing tutorial for classifying tumors vs. non-tumors on TCGA data using glmnet, and as the authors explain, it is a happy middle ground between two other ML models, Lasso and Ridge [16]. Specifically, the cv.glmnet model automatically executes 20-fold cross-validates the training data it receives, lessening the possibility for poor performance due to the model itself and not the data subset. Though the format of my data was different than the tutorial, the glmnet documentation was quite helpful.

3.1.7. Survival Prediction

blockForest and RandomForest For survival prediction, the blockForest[18] model was originally chosen, because blockForest was found to be the best performing survival ML model on the TCGA dataset in terms of ibrier score in a large benchmark study[17], though the authors suggest to take this claim lightly. This selection was later changed to the random forest and and Cox proportional hazards models, due to unexpected difficulties with using the blockForest package, discussed in greater detail in Results. Furthermore, though the authors (hestitantly) find blockForest to outperform other ML models [17], my analysis gains no benefit from the "blocks" of blockForest

because I do not train on multi-omics data at once, but instead built two separate models for RNA-Seq and CNV data. Because of this, a standard random forest is a better suited model for this work. Though I did not use the same random forest package, in terms of ibrier performance the random forest model was in the upper half of models tested in [17]. Additionally, building the blockForests was very slow, taking hours of computation time, so I did not really have the time to train more blockForest models on additional subsets of data anyway.

RandomForestSrc and CoxPH Following the difficulties with calculating an ibrier score with the blockForest package, I used the randomForestSRC[18] and the CoxPH[19] packages for survival prediction because these have built in methods for calculating the c-index and ibrier score of a model. An rfsrc prediction automatically returns the ibrier score, which the package refers to as "CRPS", and returns $1 - $ c-index, referred to as "Requested Performance Error". These two measures can also be computed fairly easily from a coxph model as well, using the complementary PEC (Performance Error Curve) package. PEC was also used to plot the selected predicted survival curves. It is important to note that the coxph method consistently crashed and threw an error when trying to predict survival using all 20000 gene variables, on both RNA-Seq and CNV data. This may have been a hardware failure, or the model just may not be equipped to handle such high-dimensional data. Instead, I trained and tested coxph using only the data from 6 specific genes, found to be especially important for determining survival in TCGA BRCA patients, using the 6-gene signature discovered by [20]. This came with additional caveats, however. One of the genes, MRGPRX1, had 0 levels for all RNA-Seq AA samples, so it was excluded for any RNA-Seq AA predicitons. Similarly, the most relevant gene for the signature, "CD24", was inexplicably not present at all in the CNV data, and all samples had 0 levels for PRRG1 in that dataset.

4. Results

4.1. Classification

Prediction performance for classification was extremely high all around, so it is difficult to compare the models to each other because they are all quite good. I will clarify though, that the AA-only and

Data Subset	ML Model	Accuracy	Precision	Recall	Specificity
All* Data	svm	0.991	0.995	0.995	0.947
AA only	svm	1.000	1.000	1.000	1.000
Trained EA/AA and tested AA	svm	1.000	1.000	1.000	1.000
EA only	svm	0.995	1.000	0.994	1.000
Trained EA/AA and tested EA	svm	0.9892	0.994	0.994	0.9545
Trained EA and tested AA	svm	0.984	1.000	0.984	1.00
All* Data	glmnet	0.989	0.988	1.00	0.842
AA only	glmnet	1.000	1.000	1.000	1.000
Trained EA/AA and tested AA	glmnet	1.000	1.000	1.000	1.000
EA only	glmnet	0.989	1.000	0.988	1.000
Trained EA/AA and tested EA	glmnet	0.989	1.000	0.988	1.000
Trained EA and tested AA	glmnet	0.984	0.995	0.833	0.989

Table 1: Performance of ML models for classification trained on TCGA BRCA RNA-Seq data.

Data Subset	ML Model	Accuracy	Precision	Recall	Specificity
All* Data	svm	0.975	0.995	0.955	0.995
AA only	svm	0.959	1.000	0.919	1.000
Trained EA/AA and tested AA	svm	0.973	1.000	0.946	1.000
EA only	svm	0.970	0.994	0.945	0.994
Trained EA/AA and tested EA	svm	0.976	0.994	0.957	0.994
Trained EA and tested AA	svm	0.956	0.994	0.918	0.995
All* Data	glmnet	0.964	0.995	0.931	0.996
AA only	glmnet	0.973	1.000	0.946	1.000
Trained EA/AA and tested AA	glmnet	0.973	1.000	0.946	1.000
EA only	glmnet	0.970	1.000	0.939	1.000
Trained EA/AA and tested EA	glmnet	0.976	1.000	0.951	1.000
Trained EA and tested AA	glmnet	0.962	1.000	0.924	1.000

Table 2: Performance of ML models for classification trained on TCGA BRCA CNV data.

EA/AA trained, AA-tested RNA-Seq models (both svm and glmnet) are very likely overfitted due to lack of data. There are only 189 AA samples, compared to 932 EA samples, and only 6 of the AA samples are normal tissue.

4.1.1. RNA-Seq

svm performance offers mild evidence to support the boons of ancestry-specific classifiers, but not enough to be conclusive. Classification performance on EA samples does drop when AA data is added, but the difference is minimal. Though likely overfitted, AA classification performance is unaffected by the prescence of additional EA samples, and performance metrics are still very high

even when the svm is EA trained/AA tested.

glmnet has the worst performance metric of any classification model in this paper, with the EA trained, AA tested RNA-Seq based glmnet model having a recall of 0.833, but this is still rather high. Futhermore, both AA and EA specific glmnet models perform the same in the absence of data from the other ancestry group, which suggests that there is just little benefit from using an ancestry-specific classifier for tumor prediction, if there is any benefit at all.

4.1.2. CNV

svm s based on RNA-Seq data weakly suggested that the ancestry-specific division of a training and test set could only provide a small improvement, if any. Here, with CNV data, the opposite is true. The ancestry-specific models are outperformed on all four metrics by the models testing on the same set with additional differing-ancestry samples. The AA-specific svm is also only very marginally better than the EA trained AA tested svm, again implying that there are little to no performance gains from using an ancestry-specific classifier.

glmnet scores perfect precision and specificity (no false positives) on every data subset except All* Data, and even that is still extremely high. Accuracy and recall are either unchanged or are strictly worsened by the use of an ancestry-specific model, albeit slightly.

4.2. Analysis of Incorrectly Classified Samples

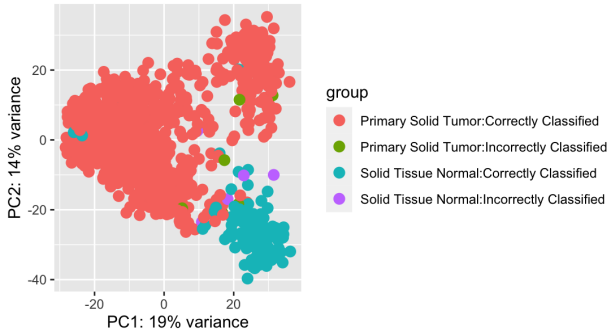

Figure 1: Visualization of primary component analysis of TCGA BRCA RNA-Seq data. Each point represents a sample. Points are colored by their true sample classification, and whether that classification was predicted accurately or not.

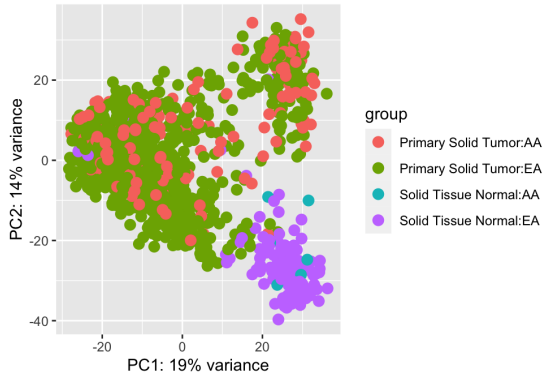

Figure 2: Visualization of primary component analysis of TCGA BRCA RNA-Seq data. Each point represents a sample. Points are colored by their true sample classification and their genetic ancestry.

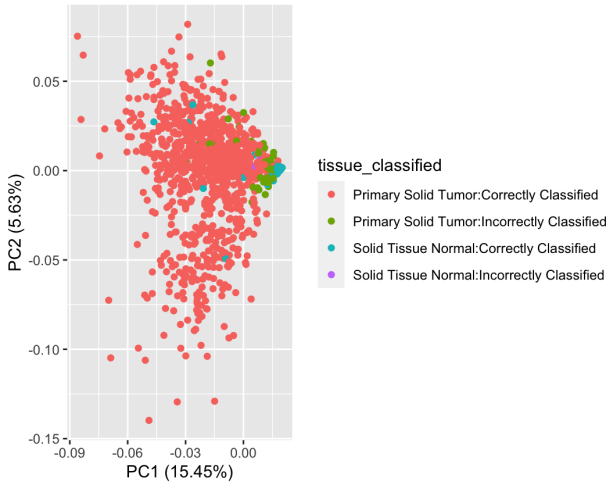

Figure 3: Visualization of primary component analysis of TCGA BRCA CNV data. Each point represents a sample. Points are colored by their true sample classification, and whether that classification was predicted accurately or not.

I recorded all of the samples that were misclassified by the RNA-Seq-based classifiers, and the CNV-based classifiers, labeled the samples according to whether their classification was correct or not, and performed principal component analysis (PCA) on the dataset using these labels. See Figure 1, Figure 3. This was done in order to determine if these samples had any characteristics in common that may be causing some, or all, of them to be misclassified for similar reasons. While not the most rigorous form of clustering, PCA allows for a clear and easily interpretable visualization of the data that can immediately show if the labeled samples may have characteristics in common [21]. I also created separate graphs that labeled the samples by sample type and genetic ancestry, to get a sense of how these genetic makeups might differ along ancestral lines, though obviously

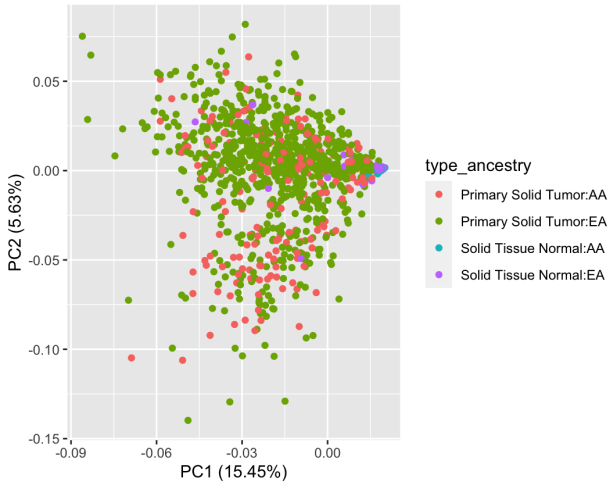

Figure 4: Visualization of primary component analysis of TCGA BRCA CNV data. Each point represents a sample. Points are colored by their true sample classification and their genetic ancestry.

PCA is not getting the whole picture, as evidenced by the low variance scores on the graphs. See Figure 2, Figure 3.

4.3. Survival Prediction

The survival data for this dataset is visualized by Figure 5. The integrated Brier (ibrier) score and concordance index (c-index) score are both widely used measures of survival prediction accuracy. A lower ibrier score indicates higher accuracy, with a perfect prediction having a score of 0 and a random prediction having a score of 0.25. A higher c-index indicates higher accuracy, with a perfect predicition scoring a c-index of 1 and a random prediction scoring a c-index of 0.5 [17]. Note that the ibrier score is considered a stronger measure of prediction accuracy (accuracy being used in the general sense here) than the c-index, which is not considered proper[17], so that column could be used as a weak tie-breaker of sorts when comparing two instances in which one model has a better ibrier score but worse c-index and the other has a worse ibrier score and better c-index.

4.3.1. blockForest The results of blockForest presented here should be taken with extreme caution, if not outright ignored. The calculated c-indexes of the blockForests support the hypothesis that an African-Specific model makes for better predictors of survival among people with primarily African ancestry, but the same is not necessarily true for a European-specific model, as the c-index drops

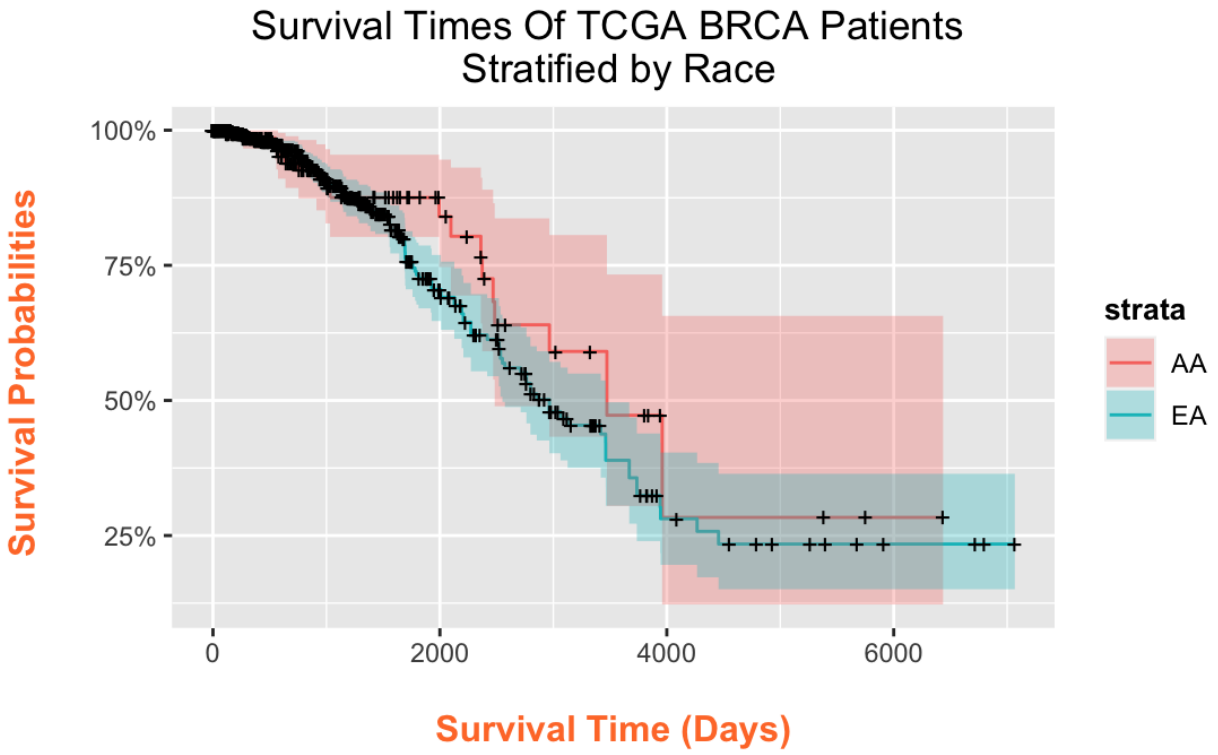

Figure 5: Kaplan-Meier curves for TCGA BRCA survival data, split by ancestry. The curve indicates the probability that a sample will survive up to a given amount of days. Crosses (+) on the curve represent censored data.

compared to training/testing on all data. A possible explanation is that the concordance of the all data model is brought up by AA predictions, but this was not tested. After computing c-indexes for these models, I computed ibrier scores as well, but found that all ibrier scores were greater than 0.25, so these scores signify that the blockForest models I trained have no predictive power for survival time, and are actually worse than guessing randomly. This may be due to the model itself, but the poor scores may also be due to my own user error when coding them (the blockForest package does not have much documentation, and it was not clear how to correctly compute the ibrier score). Either way, after discovering this, I did not continue further with any blockForest models.

4.3.2. RNA-Seq

rfsrc models trained on RNA-Seq have rather irregular performance. It is inconclusive whether an AA-specific rfsrc yields better or worse survival predictions for AA samples. The general All* Data model outscores the AA-specific model in terms of both c-index and ibrier score, and the c-index

Data Subset	ML Model	ibrier	c-index
All* Data	rfsrc	0.082	0.737
AA only	rfsrc	0.125	0.683
Trained EA/AA and tested AA	rfsrc	0.106	0.537
EA only	rfsrc	0.070	0.671
Trained EA/AA and tested EA	rfsrc	0.091	0.926
Trained EA and tested AA	rfsrc	0.100	0.588
All* Data	coxph	0.146	0.649
AA only	coxph	0.202	0.101
Trained EA/AA and tested AA	coxph	0.072	0.914
EA only	coxph	0.161	0.671
Trained EA/AA and tested EA	coxph	0.158	0.672
Trained EA and tested AA	coxph	0.154	0.560
All* Data	blockForest	0.359	0.790
AA only	blockForest	0.333	0.846
EA only	blockForest	0.336	0.764
Trained EA and tested AA	blockForest	0.285	0.733

Table 3: Performance of ML models for survival prediction trained on TCGA BRCA RNA-Seq data. ibrier and c-index scores are rounded to 3 decimal places for ease of reading and comparison. Note: lower ibrier scores and higher c-index scores indicate better performance.

drops, but the ibrier score improves when EA samples are added to training data, with even the extreme EA biased model outscoring by ibrier score. Interestingly, though the ibrier score worsens when AA samples are included for training for EA predictions, the c-index sees a large boost.

Data Subset	ML Model	ibrier	c-index
All* Data	rfsrc	0.096	0.554
AA only	rfsrc	0.153	0.574
Trained EA/AA and tested AA	rfsrc	0.134	0.557
EA only	rfsrc	0.120	0.606
Trained EA/AA and tested EA	rfsrc	0.115	0.972
Trained EA and tested AA	rfsrc	0.094	0.544
All* Data	coxph	0.175	0.576
AA only	coxph	0.113	0.733
Trained EA/AA and tested AA	coxph	0.101	0.848
EA only	coxph	0.192	0.553
Trained EA/AA and tested EA	coxph	0.185	0.559
Trained EA and tested AA	coxph	0.180	0.575

Table 4: Performance of ML models for survival prediction trained on TCGA BRCA CNV data. ibrier and c-index scores are rounded to 3 decimal places for ease of reading and comparison. Note: lower ibrier scores and higher c-index scores indicate better performance.

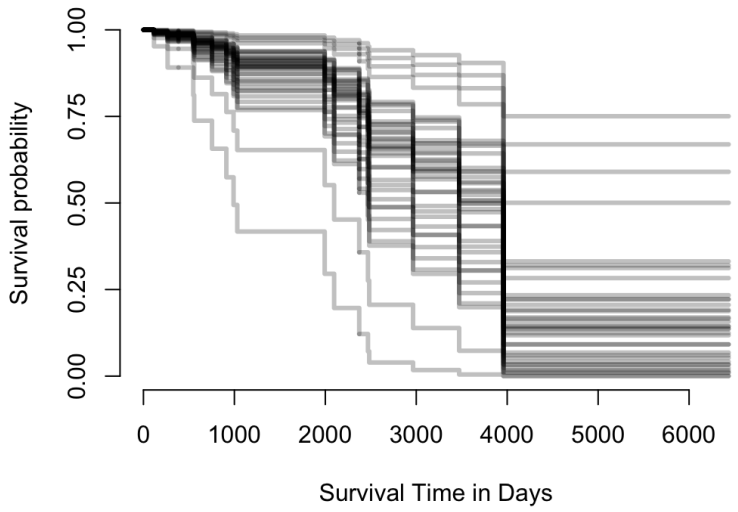

Figure 6: The RNA-based, AA only coxph model's predicted survival curve of the subset of AA samples.

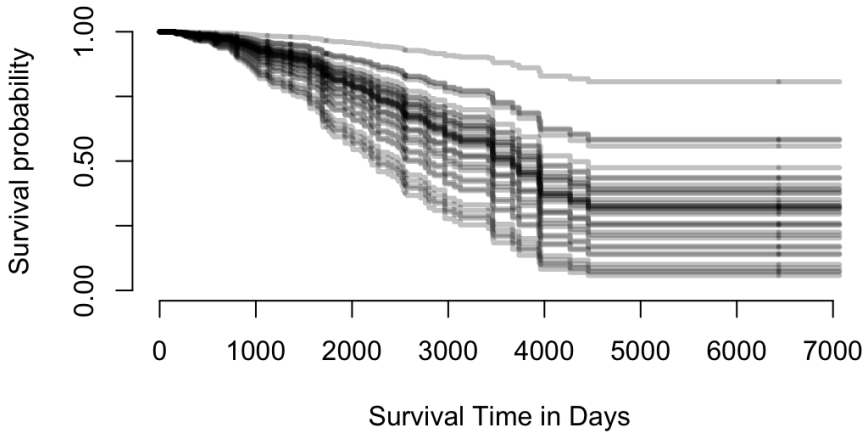

Figure 7: The RNA-based, EA/AA trained and AA tested, coxph model's predicted survival curve of the subset of AA samples. Note the significantly increased confidence of the prediction.

coxph models based on RNA-Seq data also see somewhat strange performances, however, there is a pattern emerging in these results. The AA-specific model scores a fairly bad ibrier score (close to 0.25), and an abysmal c-index, evidencing against its effectiveness in this setting. AA prediction performance gains a tremendous boost from the addition of EA data, though. This jump is visualized by the predicted survival curves of these two models: See Figure 6 and Figure 7 Both the ibrier and

c-index scores for AA and EA samples suggest that predictions are improved by the presences of other ancestries' samples in the training set, though the improvement for EA prediction is much slighter.

4.3.3. CNV

rfsrc scores poor c-indexes (close to .50) on most of the models when trained on CNV data. However, the notable exception is the EA/AA-trained EA-tested model, which achieves the highest c-index among all survival prediction models, and also improves on the ibrier score of the EA-specific model. AA prediction performance also sees a slight ibrier and c-index boost from the addition of EA data. Also notable are the fact that the lowest ibrier scores here come from the All* Data and EA-Trained AA-Tested models, all of which are more evidence against using ancestry specific models for survival prediction.

coxph models based on CNV data CNV-based coxph models should be viewed with extra caution because they were built only training and testing with 4 genes out of a 6-gene signature (out of over 20000 total genes), especially considering that one of the genes missing is the most important in the 6-gene survival signature. Still though, the ancestry specific models are outperformed when the other ancestry's data is used to bolster the training set.

5. Conclusion

5.1. Discussion

With classification, I believe that the conclusion to be drawn is that the genetic ancestry distribution of the training and test sets has a minimal impact on classifier performance either way. The RNA-Seq svm and glmnet models demonstrate very mild performance improvements from ancestry-specificity, while the CNV models show a slight performance drop from ancestry-specificity. The CNV results should arguably take priority over the RNA-Seq results when evaluating the models, because the CNV dataset has much more non-tumor normal samples available for training and testing. Regardless, though, every tumor classifier is still quite strong and the differences among

models are very small.

It appears a recurring pattern in the survival prediction models is that a larger dataset benefits prediction performance, even if that dataset is from a group of individuals with different genetic ancestry. No ancestry-specific ML model strictly outperforms their counterpart model that has additional data. The RNA-Seq rfsrc AA-specific model achieves a better c-index, but the ibrier score worsens, as does the CNV rfsrc AA-Specific mode. Also, the RNA-Seq rfsrc EA-specific model yields a better ibrier score (the lowest of all models), but worse c-index than its complement. Otherwise every other ancestry-specific model is strictly worse off for excluding the additional data. Additionally, it generally seems that RNA-Seq based survival prediction models perform better than CNV based survival prediction models, though this is expected for CNV coxph models due to having less data to work with.

In sum, I would conclude that, for the purposes of predicting whether a sample is a breast cancer tumor or not, genetic ancestry has a negligible, or almost negligible, impact on classifier performance, and ancestry-specific classifiers are of little consequence. For the purposes of survival prediction, however, it appears that the best performance results are attained by training on a mix of samples from different ancestries (including that of the test samples). Ancestry-specific classifiers, it seems, may actually be a detriment to survival prediction performance, and should be avoided unless further research conclusively shows otherwise.

5.2. Future Work

This research is partly limited by my computational power. For example, as mentioned previously, RStudio running on my MacBook can only generate a coxph model when using less than 10 genes as variables, otherwise, it fails. For future researchers, a more powerful computer may be able to create a more accurate coxph model, and thus better evaluate the benefits, or lack thereof, of ancestry-specific machine learning, especially in regards to the CNV-trained coxph models, which only used 4 genes as predictor variables.

It remains to be seen whether there are clear performance benefits to be gained from ancestry-

specific ML models that are trained on multi-omics data. Further investigation is needed into these cases, such as, for example, evaluating a blockForest model trained on both RNA-Seq and CNV data. Other ML models should be tested as well, such as CoxBoost and ipflasso. Additionally, further research should look into whether using patients' clinical data in addition to genomic data yields different results. On that note, it is also to be determined whether ancestry-specific ML models based on clinical TGCA data outperform general ones. It may be the case that the performance gains found by "Evaluation of race/ethnicity-specific survival machine learning models" [3] are only applicable in the SEER* dataset, but not the TCGA dataset.

In particular, this research looked only at the TCGA BRCA dataset, but it would be worth exploring the datasets of other types of cancer and determining the merits of ancestry-specific machine learning. Similarly, though TCGA is a large and easily accessible dataset, this analysis should also be conducted on different datasets and the findings compared, in order to discern if there is a trend in conclusions across these datasets, or if findings are limited and not generalizable to other data.

Note that the size and diversity of the TCGA BRCA dataset are limiting factors to this work, that could be resolved with a larger dataset. The dataset does not have enough samples to support a Hispanic-specific classifier as done by [3], and there are not enough patients with East Asian, Native American, or Pacific Islander ancestry to sufficiently analyze those groups either. Considering the conclusion that more data from any patient may possibly strengthen a ML model, an extension of this paper may look at how adding back in those patients with ancestries other than EA or AA into training data affects subsequent predictive power. For instance, how may Trained All and tested AA ML models compare to those that were trained EA/AA and tested AA? because this would be a rather contrived scenario that is not reflective of the real world.

In general, ML models for cancer that are specific to other factors, such as gender or age, may also be worth investigation and consideration. This concept extends beyond just cancer classification and survival prediction. For example, women's risk of developing adverse reactions from drugs are twice as high as men's risk, so there may be benefits to male and female specific ML models that

predict the risk of adverse effects when taking a drug. [22]

Lastly, though, given the ongoing debates about the use of race in medicine, even if future research concludes a clear and significance performance gain from ancestry-specificity in classification and survival prediction of cancer, there may be ethical and moral issues that come with these ML models that must be addressed. There are concerns that race-based medicine can perpetuate racist ideas and racial essentialism [6], and it is not clear whether substituting in "genetic ancestry" as essentially a euphemism for race would solve these problems. Moreover, due to the troublesome history of medicine and the exploitation of Black people by those in the field[23], there may be additional difficulties arising from hesitance and mistrust toward ancestry-specific ML models. Communication and transparency are of utmost importance if ancestry-specific machine learning begins to see wider use.

References

[1] National Cancer Institute, "The Cancer Genome Atlas Program," *National Cancer Institute*, 2019. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
[2] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, "Addressing bias in big data and AI for health care: A call for open science," *Patterns*, vol. 2, no. 10, p. 100347, Oct. 2021, doi: https://doi.org/10.1016/j.patter.2021.100347.
[3] Park JI, Bozkurt S, Park JW, et al Evaluation of race/ethnicity-specific survival machine learning models for Hispanic and Black patients with breast cancer BMJ Health & Care Informatics 2023;30:e100666. doi: 10.1136/bmjhci-2022-100666.
[4] J. Yuan *et al.*, "Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers," *Cancer cell*, vol. 34, no. 4, pp. 549-560.e9, Oct. 2018, doi: https://doi.org/10.1016/j.ccell.2018.08.019.
[5] N. Goel, D. Y. Kim, J. A. Guo, D. Zhao, B. A. Mahal, and M. Alshalalfa, "Racial Differences in Genomic Profiles of Breast Cancer," *JAMA Network Open*, vol. 5, no. 3, pp. e220573–e220573, Mar. 2022, doi: https://doi.org/10.1001/jamanetworkopen.2022.0573.
[6]T. Krainc and A. Fuentes, "Genetic ancestry in precision medicine is reshaping the race debate," *Proceedings of the National Academy of Sciences*, vol. 119, no. 12, 2022.
[7] Q. N. Spencer, "A racial classification for Medical Genetics," *Philosophical Studies*, vol. 175, no. 5, pp. 1013–1037, 2018.
[8] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[9] "TCGAA (the cancer genetic ancestry Atlas) - 52.25.87.215," *The Cancer Genomic Ancestry Atlas*. [Online]. Available: http://52.25.87.215/TCGAA/
[10] Kosinski M (2023). *RTCGA: The Cancer Genome Atlas Data Integration*. R package version 1.30.0, https://rtcga.github.io/RTCGA..
[11] L. Richardson, "Beautiful Soup documentation," *Beautiful Soup Documentation - Beautiful Soup 4.4.0 documentation*. [Online]. Available: https://beautiful-soup-4.readthedocs.io/en/latest/. [Accessed: 03-May-2023].
[12] "Sample type codes," *Sample Type Codes | NCI Genomic Data Commons*. [Online]. Available: https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes. [Accessed: 03-May-2023].
[13] Ying-Wooi Wan, Genevera I. Allen, Zhandong Liu, TCGA2STAT: simple TCGA data access for integrated statistical analysis in R, *Bioinformatics*, Volume 32, Issue 6, March 2016, Pages 952–954, https://doi.org/10.1093/bioinformatics/btv677
[14] "SVM: Support Vector Machines," *RDocumentation*. [Online]. Available: https://www.rdocumentation.org/packages/e1071/versions/1.7-13/topics/svm.
[15] "glmnet: fit a GLM with lasso or elasticnet regularization," *RDocumentation*. [Online]. Available: https://www.rdocumentation.org/packages/glmnet/versions/4.1-7/topics/glmnet
[16] T. Maie and M. Manolov, "Analysis of Cancer Genome Atlas in R - costalab.org," *Costa Lab*. [Online]. Available: https://www.costalab.org/wp-content/uploads/2020/11/R_class_D3.html.
[17] M. Herrmann, P. Probst, R. Hornung, V. Jurinovic, and A.-L. Boulesteix, "Large-scale benchmark study of survival prediction methods using multi-omics data," *Briefings in Bioinformatics*, vol. 22, no. 3, Aug. 2020, doi: https://doi.org/10.1093/bib/bbaa167.

[18] Ishwaran H. and Kogalur U.B. (2023). Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC), R package version 3.2.0.
[19] “Coxph: Fit proportional hazards regression model,” *RDocumentation*. [Online]. Available: https://www.rdocumentation.org/packages/survival/versions/3.5-5/topics/coxph.
[20] Mo W, Ding Y, Zhao S, Zou D, Ding X. Identification of a 6-gene signature for the survival prediction of breast cancer patients based on integrated multi-omics data analysis. PLoS One. 2020 Nov 10;15(11):e0241924. doi: 10.1371/journal.pone.0241924.
[21] “PCA Visualization - RNA-seq,” *PCA visualization - RNA-seq*. [Online]. Available: https://alexslemonade.github.io/refinebio-examples/03-rnaseq/dimension-reduction_rnaseq_01_pca.html
[22] Chandak P, Tatonetti NP. Using Machine Learning to Identify Adverse Drug Effects Posing Increased Risk to Women. Patterns (N Y). 2020 Oct 9;1(7):100108. doi: 10.1016/j.patter.2020.100108. Epub 2020 Sep 22. PMID: 33179017; PMCID: PMC7654817.
[23] Scharff DP, Mathews KJ, Jackson P, Hoffsuemmer J, Martin E, Edwards D. More than Tuskegee: understanding mistrust about research participation. J Health Care Poor Underserved. 2010 Aug;21(3):879-97. doi: 10.1353/hpu.0.0323. PMID: 20693733; PMCID: PMC4354806.

This paper represents my own work in accordance with University regulations. /s/ Andres Blanco Bonilla