Capstone Project Final Report - Genomics Project

Tomas Meade

Contents

1 Introduction

Gynecologic and breast cancers have a strong prevalence throughout the world and in the United States. In 2017, there were 350,000 cases of Pan-Gyn cancers in the United States and many more outside of the U.S. Specifically, breast cancer alone has the highest rate of diagnosis of any other cancer in the world according the World Health Organisation. Breast cancer accounted for 12% of new cancer cases in the world in 2021 and resulted in 650,000 deaths in 2020. However, survival rates have shown improvement in the past few decades. The American Cancer Society reported in 2017 that over the last 25 years, the number of deaths has dropped by 40%.

This is a promising improvement that can be partially accredited to clinical advancements like targeted therapies. These targeted therapies focus on the molecular make up of tumors unlike chemotherapy which targets the cancer cells. In order to continue to develop and improve targeted therapies, recent research has attempted to identify molecular features and therapeutic targets based on genomic data.

The aim of this project is to replicate some of the recent molecular research on breast cancer, specifically parts of the research done in the paper titled "A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers" by Berger et al. [1]. We will attempt to find shared and unique molecular features of breast cancer tumors as well as identify significant clinical characteristics and their relation to the genetic make up of patients. Of particular interest to the project is genomic features of patients that relate to the treatment of breast cancer like the HER2 gene which is known to have a strong relationship to breast cancer based on previous research. To examine this relationship, we account for general HER2 mutation and whether the gene exhibited amplification or deletion in each patient. The reason HER2 is relevant is because it produces proteins which regulate the growth of breast cells. Having amplified HER2 means that a patient is HER2 positive and that breast cells grow too rapidly. This amplification can result in breast cancer cells spreading faster, resulting in more severe cases. Therefore, it is a important factor in understanding and treating breast cancer. Other important characteristics we examine are estrogen and progesterone receptor positive breast cancers. Cancer cells with these receptors need estrogen and progesterone to grow, which makes them important factors since treatments can target these receptors to inhibit tumor growth.

In total, the project will extract clinical and genomic data of breast cancer patients, efficiently group the patients into clinically relevant subtypes, identify potential treatments and provide insight into survival outcomes.

Clinical Data	N=1098
Sex	
Male	12 (1.1%)
Female	1085 (98.9%)
Vital Status	
Alive	945 (86.1%)
Dead	152 (13.9%)
AJCC Stage	
Stage 1	183 (16.8%)
Stage 2	621 (57.2%)
Stage 3	249 (22.9%)
Stage 4	20 (1.9%)
Stage X	13 (1.2%)

Table 1: Summary Statistics of the TCGA-BRCA dataset.

2 Data Description

The data source for our project is the Genomic Data Commons or GDC. The GDC is part of the National Cancer Institute (NCI) and is a cancer research database with a specific focus on genomic data. The data sets are generated by NCI designated cancer centers around the U.S like university medical centers and hospitals. It includes data sets from two of the largest programs devoted to examining genomic data related to cancer research and treatment. The two programs being The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Therapies (TARGET). The mission of the GDC is to provide data to assist with genomic cancer research projects.

For our project we choose to narrow down the focus and concentrate on the TCGA and specifically the TCGA-BRCA project which contains data on patients with breast cancer. The TCGA-BRCA project data is organized by case or patient with breast cancer. There are a total of 1,098 cases in the TCGA-BRCA project. Of the 1098 cases, 1,085 are females and 12 are males, 945 are alive and 152 are dead, and there are 21,058 different genes present (see Table 1). Of special interest to us is the HER2 gene (see Figure 1). In the data set there are 233 cases with HER2. As for the general structure of the data, each observation is a case and the features are different types of genomic data, clinic data and bio specimen data which contains information about the actual physical sample that was taken from the patient. The other features contain information about mutations and genes present in the sample and also images of the samples. To conduct our analysis we will utilize only the clinical data and parts of the genomic data. Since there are many different types of genomic data, we decided to focus on a few key types that were identified as significant by the paper mentioned earlier. For the first part of our analysis we use copy number variation (CNV) data. CNV data can be converted into focal scores which are a centralized way of determining if variation in a genome resulted in amplification or deletion of a gene, or if no significant variation occurred.

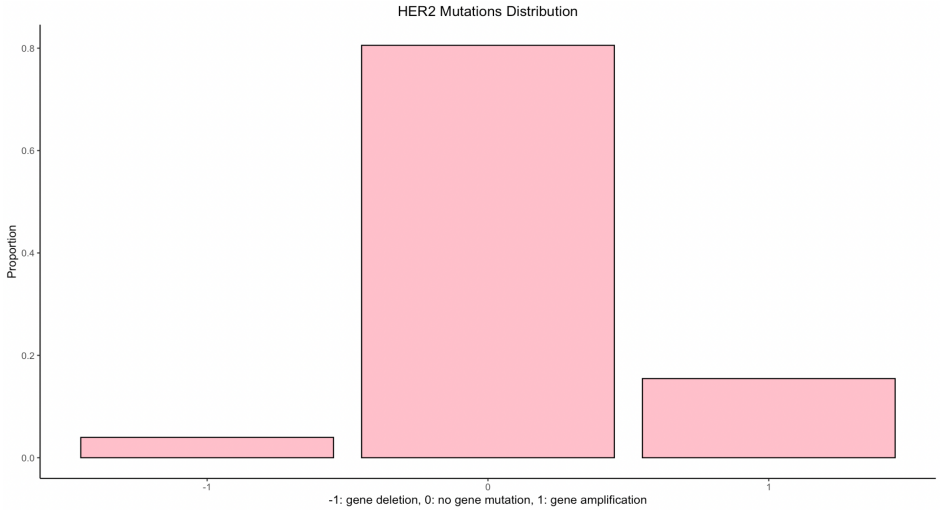


Figure 1: Proportion of HER2 amplifications, deletions and none significant mutation amoung the patients.

For the second part of our analysis we use other features of genomic data that were identified as significant in previous literature. The key variables are protein expression of ER and PR, ERBB2 amplification status (HER2 is also known as ERBB2), hyper mutator phenotype, CNV load, immune score and mutation status of PTEN, TP53, H-RAS/K-RAS/N-RAS, ERBB2, PIK3CA, and POLE. The mutation status of these genes is shown in more detail in Figure 2. We then look at similarities and differences between cases based on these genomic features and then connect them to clinic data to examine how they relate to things like treatment and survival. To conduct the survival analysis we examined the built-in time-based bias since patients can be in different stages of cancer when they are first diagnosed. The plots below show how the survival rates differ by AJCC pathological stage which is a system to describe the amount and spread of cancer in a patient's body. In stage IV, patients with HER2 amplification and deletion have survival curves that drop much faster than those who have no HER2 mutation (see Figure 3).

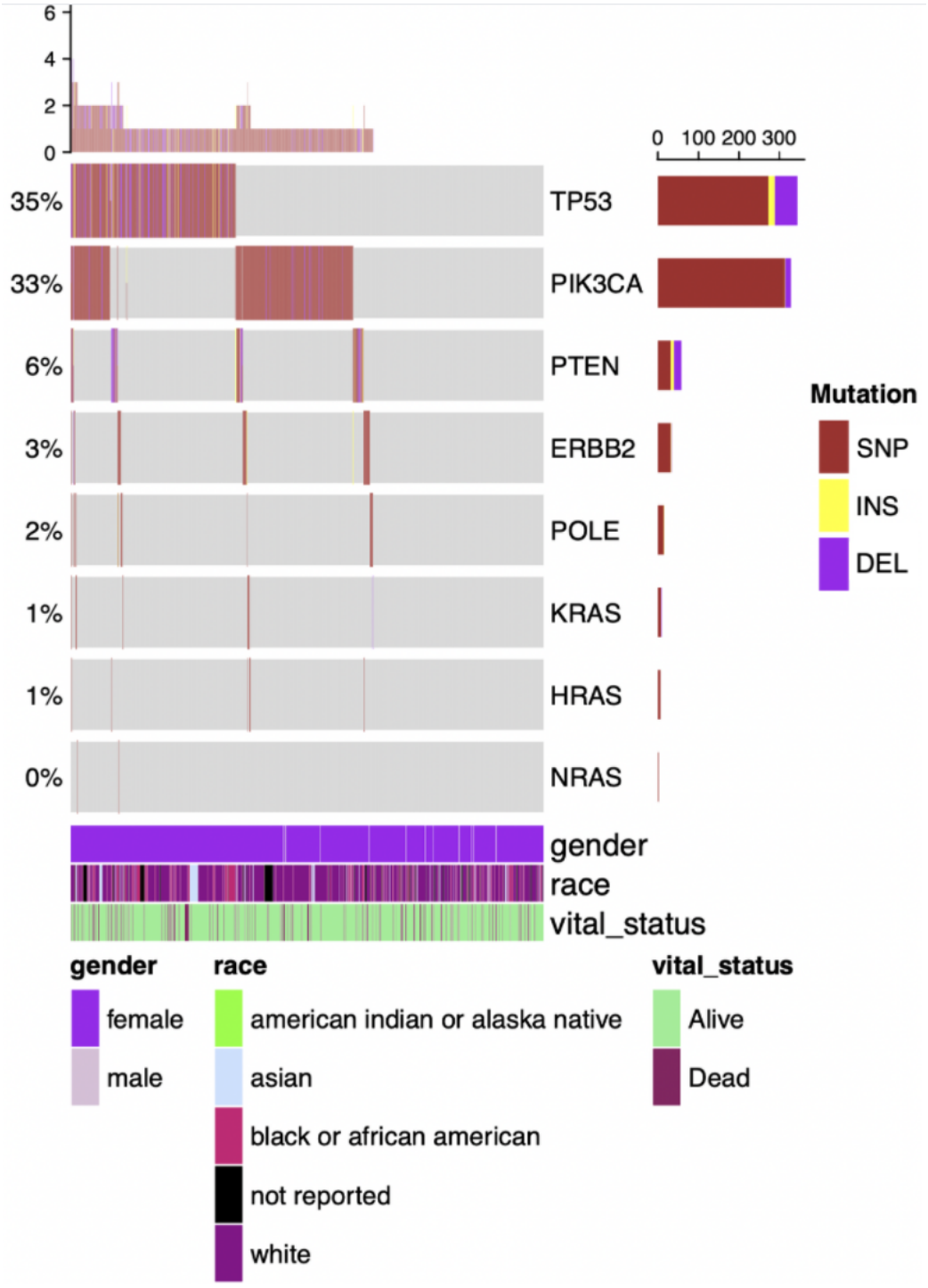


Figure 2: Diagram showing the types of mutations for the key genes present in the dataset. Each of the thin colored lines represent a patient and the color specifies the type of mutation that occured for the genes listed on the right. The gender, race and vital status for each of the patients is also listed. The percentages are the proportion of times the gene exhibited a mutation out of the total samples the gene was present in.

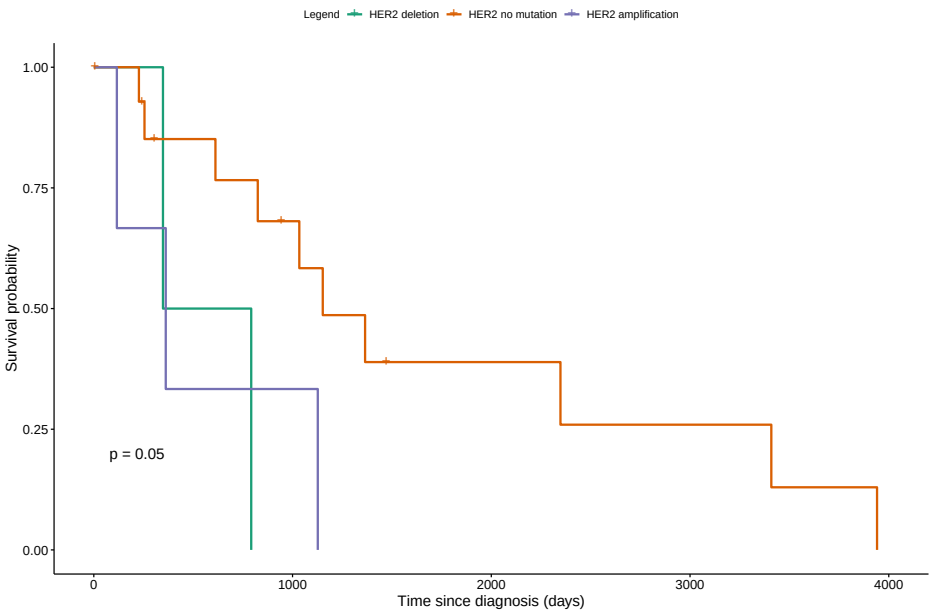


Figure 3: Survival Plot for patients in the AJCC stage IV split by HER2 mutation.

In order to access the data, we queried the GDC platform from an R environment. The GDC processes data using bioinformatics pipelines which then can be accessed by researchers. One way to access this is using the Bioconductor software package which is based primarily in R, but can be used with other software as well. In order to access the TCGA data we used an R package called TCGAbiolinks which utilizes the Bioconductor package and allows you to write queries directly in R to to extract TCGA project specific data. For the first part of our analysis, we pulled the clinical and CNV focal score data from the TCGA-BRCA project into R using TCGAbiolinks.

3 Methods

The first step in our analysis involved fitting a preliminary model that grouped patients using only one genomic feature and clinic data. More discussion of the preliminary model and the methods used can be found in Appendix A.

We then moved to our main clustering model, which focused only on the genomic features of patients in the TCGA-BRCA dataset. The features used were the 12 genomic features stated in the Data Description section. Before fitting our model, we processed the data in the same manner as in [1] by converting all features to binary variables. This entailed representing HER2 amplification as 1 if HER2 was amplified in the patient and 0 otherwise. We also converted gene mutations of PTEN, TP53, HRAAS/KRAS/NRAS, PIK3CA, ERBB2

and POLE into 1 if mutation occurred in a patient and 0 otherwise. For ER and PR expression we simply converted positive to 1 and negative to 0. We changed the remaining variables to binary by representing high amounts as 1 and low amounts as 0 with a threshold set at the median. So if the patient had higher than the median amounts of CNV (CNV load), mutation (hyper-mutation), or immune cell infiltration (immune score), we input the feature as a 1 for the patient and 0 otherwise.

To analyze shared and unique molecular features of breast cancer tumors, we used unsupervised learning to group similar patients together based on these 12 genomic features. Specifically, we used agglomerative hierarchical clustering to group the patients into similar clusters. To preform the clustering, we again followed the same method described in [1], which was hierarchical clustering with Ward linkage and 1-Pearson's correlation coefficient as the distance metric. More details on this can be found in Appendix B. We choose an optimal number of clusters based on the number of clusters found in [1].

After completing the clustering with all 12 features, we then fit a decision tree using only 4 features to attempt to group the patients into those same established 5 clusters. The reasoning behind this is because grouping the patients using all 12 features is valuable in a clinical setting, but also unrealistic since testing for this many genomic features in expensive and time consuming. Plus, a significant amount of these tests are not commonly available. We used a train-test split of 80-20% to implement the decision tree and test its accuracy.

We then performed survival analysis by cluster to examine differences in clinical outcomes. First, we generated Kaplan Meier plots that showed the survival curves of the 5 clusters and conducted a hypothesis test to determine if survival rates differed significantly between clusters. To account for the stage at diagnosis bias, we separated the plots by early and late stage. Last, we fit a Cox proportional hazards regression of survival on tumor stage and cluster to confirm our results.

Figure 4 below shows an overall flowchart of the steps we took to complete the project and the different methods we used for our analysis.

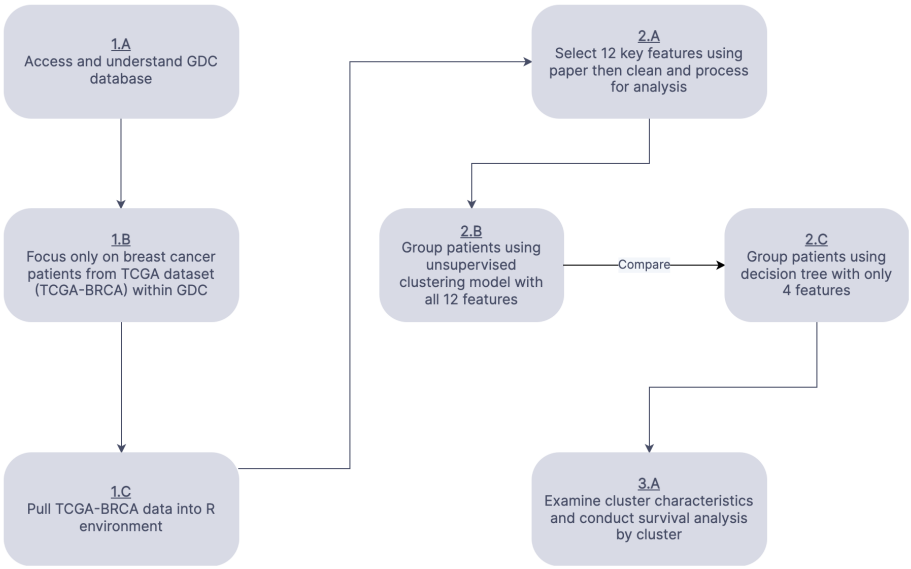


Figure 4: Flowchart overview of the Project and Methods used.

4 Results

Our final cluster model resulted in the dendrogram shown in Figure 5. We cut the dendrogram at 5 clusters, resulting in 5 unique groupings of patients based on their genomic features relating to breast cancer. Table 2 shows the makeup of the clusters and the weights of some of the genomic features that characterize them. We can see that cluster 1 has a high proportion of patients with HER2 amplification while clusters 2, 3 and 5 all have a high proportion of patients who are ER positive.

clus	pr	er	her2_amp	cnv_load	immune_score	hyper_mutator
1	0.50	0.62	0.98	0.79	0.59	0.59
2	0.81	0.97	0.01	0.86	0.46	0.60
3	0.85	0.94	0.00	0.01	0.82	0.26
4	0.02	0.16	0.02	0.72	0.61	0.67
5	1.00	1.00	0.00	0.09	0.03	0.00

Table 2: Genomic characteristics of clusters.

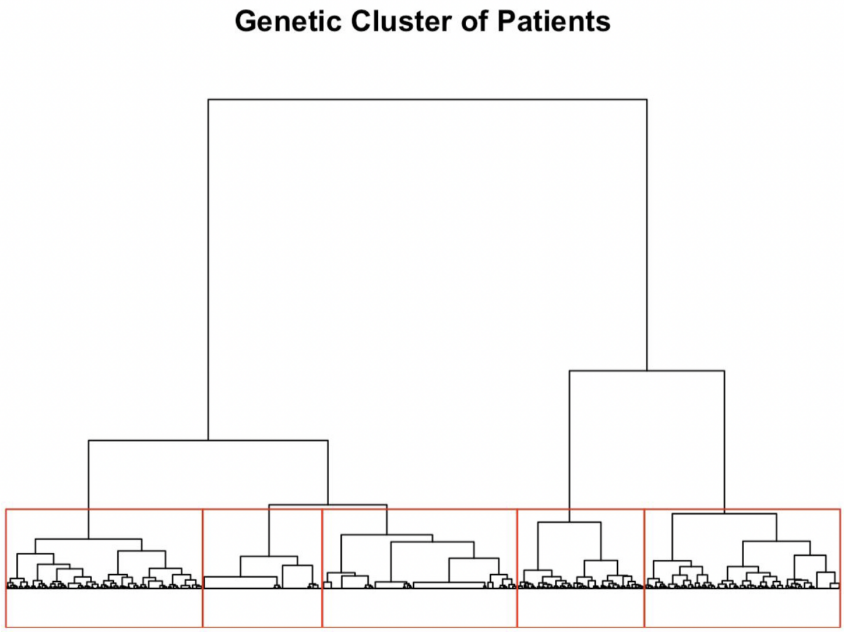


Figure 5: Final model dendrogram cut at 5 clusters.

We then trained a decision tree to classify the patients into the 5 clusters. The decision tree using 4 features achieved an accuracy of 86% when it came to placing the patients into the correct clusters. It is shown below in Figure 6. The decision tree provides a natural way to highlight feature importance by examining the order of the splits. The first split is the most important and our decision tree uses HER2 amplification as the first split, verifying the strongly established connection between the HER2 gene and breast cancer. The decision tree also provides an intuitive structure for testing in a clinical setting where the order of tests would be done in alignment with the order of the tree splits. This would be beneficial because it could potentially prevent excess or unneeded testing. For example, if a patient tested positive for HER2, they would immediately be placed in cluster 1 and would not necessarily require the other tests. Our decision tree is similar to the decision tree created in [1] with the major difference being the initial split on HER2 amplification. This is due to the fact that the study done in [1] incorporates many types of cancer, while our project focuses only

on breast cancer. To view the decision tree made in [1] see Appendix C.

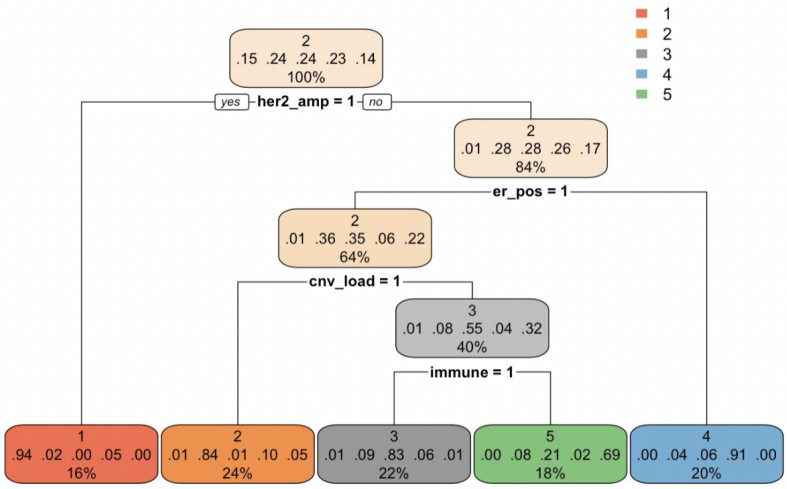
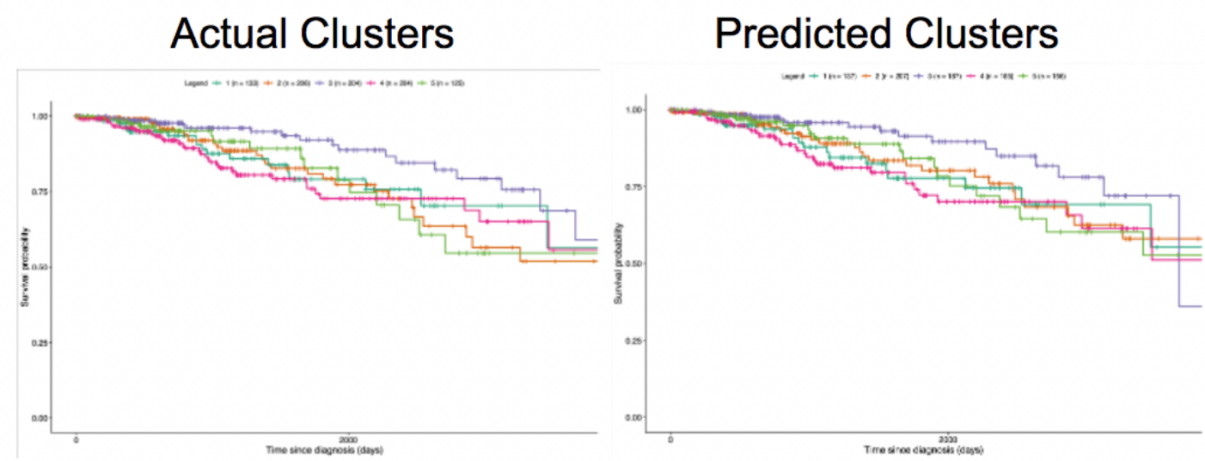


Figure 6: Decision tree using 4 features to classify patients into 5 established clusters.

To conclude our analysis we examined clinical outcomes and how they differed by cluster. We first plotted overall survival curves for the different clusters found through unsupervised learning. We also plotted survival curves for the predicted clusters found using the decision tree for comparison. These two plots are shown below in Figure 7 and demonstrate that the predicted clusters have similar survival curves to the actual clusters. The null hypothesis that there is no difference in survival curves of clusters was rejected at the 0.1 level using a log-rank test.



	5-year	10-year
C1	25.00%	10.00%
C2	34.48%	3.45%
C3	55.00%	15.00%
C4	12.12%	3.03%
C5	41.18%	5.88%

	5-year	10-year
C1	23.81%	9.52%
C2	31.03%	3.45%
C3	52.63%	15.79%
C4	16.13%	3.23%
C5	42.11%	5.26%

Figure 7: Overall survival curves of patients by cluster using actual clusters and clusters predicted by the decision tree.

To account for the AJCC stage basis, we plotted survival curves for the different clusters separated by early and late stage (see Figures 8 and 9). These two survival curves that account for stage exhibit even greater separation between clusters, highlighting further the clinical relevance of these patient groups.

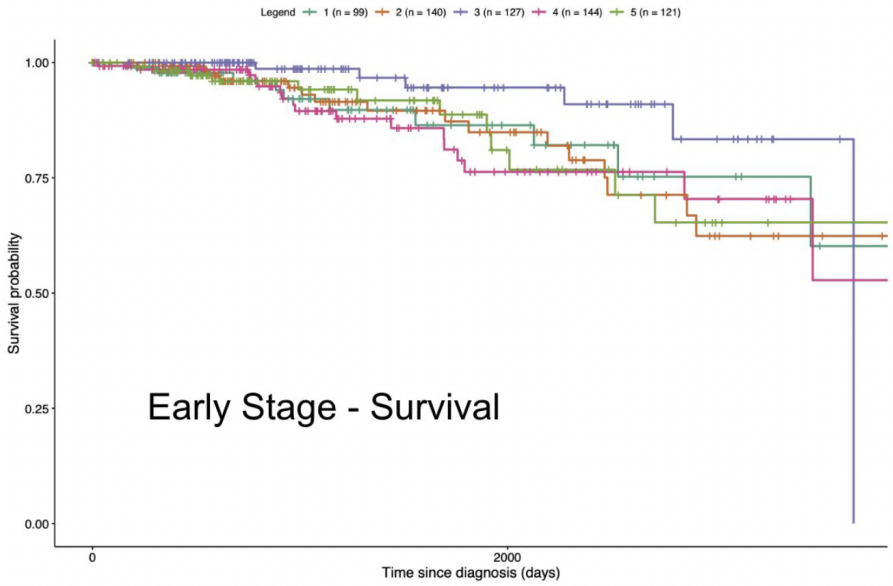


Figure 8: Survival curves of early stage patients by cluster.

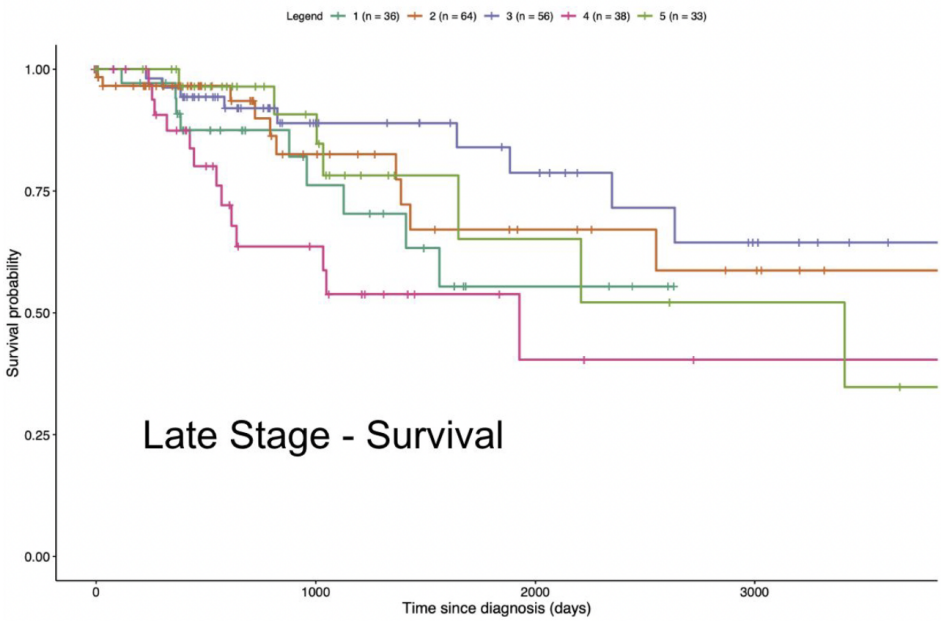


Figure 9: Survival curves of late stage patients by cluster.

We also fit a Cox proportional hazard regression model of survival on AJCC stage and cluster and compared it to a model fit on only AJCC stage. The model fit with the addition of clusters improved the AIC, R-squared and Concordance when compared to the base model with stage alone (see Appendix D). The Martingale residuals of the model using cluster and stage were also closer to a zero mean.

These results show the clinical relevance of our analysis in terms of survival outcomes and also treatment. For example, if we pick a late stage breast cancer patient who tests negative for HER2, positive for ER, negative for high CNV load and positive for high immune score this would place them in cluster 3 based on our decision tree. Since the patient is HER2 negative and has low CNV load these do not provide direct use for treatment, but allow us to focus on the two other features. High immune score is a positive since this means the patient has a higher number of immune cells that can identify and attack the cancer cells. This leaves ER as a potential clear focus for treatment of this patient. As we already established ER positive breast cancers have receptors that use the hormone estrogen to grow. A type of treatment to combat this is called Hormone therapy and a specific example of this therapy is a drug called Tamoxifen which can be used to treat women or men diagnosed with advanced-stage, hormone receptor-positive breast cancer. It works by preventing estrogen's ability to stimulate the growth of breast cancer cells. So for the example patient we know their breast cancer uses estrogen to grow and spread and we could use a hormone therapy to prevent it. From the plot of late stage survival curves shown in Figure 13, we can see that this patient would have the highest expected survival rate since cluster 3 is significantly higher than the other clusters. In support of this,

a study in the national library of medicine published in 2017 states that therapies such as tamoxifen have revolutionized the treatment of breast cancer, resulting in significant decreases in cancer-related mortality [2].

5 Conclusion

In this project we were able to replicate some of the findings in the paper "A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers" and group breast cancer patients into clinically relevant subtypes. These subtypes provide insight into potential treatments and expected survival outcomes. In the future, a larger sample size of breast cancer patients would be beneficial since our sample size was limited. There is also a crucial need for biological research into these subtypes to understand there significance and biological implications. The features we choose for our analysis were all based on previous literature stating a known association with cancer so there is also a need for more research into potential new features that relate to breast cancer and improving clinical outcomes. Finally, there is also a challenge to implementing this research in a clinical setting since genomic cancer research is still somewhat new and evolving so tests and treatments are not widely available. In our case, the findings of our project seem to be important and align with other literature, but there still exist clear challenges to implementation like testing for all the features that are needed to use our decision tree. However, targeted therapies applied using information from quantitative genomic research have demonstrated great promise in reducing mortality and will hopefully continue to be improved upon in the future.

6 Appendix A

To prepare the data for the preliminary model, all the variables were converted to factors (categorical) or integers/numeric (continuous). To preform the clustering on both the clinical and CNV data, pairwise dissimilarities between observations, in this case patients, are calculated. This is because the clinical data contained some categorical features so we had to use a slightly different method for the distance function. These dissimilarities or distances are computed using the 'daisy()'. function in R. Since we have both categorical and continuous variables, we use a generalization of Gower's formula which is the Gower's metric in hclust(). This means that the variables will be standardized and then the distance between observations is the sum of all the distances for each variable.

Gower's method starts with standardization which is done by dividing each individual value by the column's range after subtracting the minimum value. This means that each variable has a range of between 0 and 1. Then the dissimilarity is calculated as the weighted mean of the contributions of each variable or

$$d_ij = d(i,j) = sum(k = 1 : p; w_k \delta(ij;k) d(ij,k))/sum(k = 1 : p; w_k \delta(ij;k)).$$

The equation above shows that d_ij is a weighted mean of $d(ij,k)$ with weights w_k and $\delta(ij;k)$, w_k are the weights for k, $delta(ij;k)$ is 0 or 1, and $d(ij,k)$ is the distance between $x[i,k]$ and $x[j,k]$. The weight $delta(ij;k)$ is zero when $x[,k]$ is missing in one of the rows or when the variable is asymmetric binary and both values are zero, otherwise it is always one. The contribution of binary variables given by $d(ij,k)$ is zero if both values are equal and one otherwise. The contribution to the distance for the other variables is the absolute difference of both values, divided by the range of that variable.

After the distances are calculated then the clustering is performed using the distances between observations. For bottom up or agglomertive clustering, each patient starts out as its own cluster and similar patients are grouped together. The 'hclust()' function preforms the clustering in R. The algorithm joins the two closest clusters and then repeats until there is only one cluster containing all the patients. The cluster grouping is done using the complete linkage method calculates the distance between two clusters by the largest distance between a point in the first cluster and a point in the other cluster. For each step of the algorithm the distances between clusters are computed using the Lance-Williams dissimilarity update formula. Given two clusters, K and L the dissimilarity between them and another cluster M is given by

$$D(K \cup L, M) = \alpha_1 * D(A,C) + \alpha_2 * D(B,C) + \beta * D(A,B) + \phi * |D(A,C) - D(B,C)|$$

where the four coefficients $(\alpha_1, \alpha_2, \beta, \phi) are (.5, .5, 0, +.5)$.

7 Appendix B

To preform the clustering on the binary genomic data, we used 1-Pearson's correlation coefficient to quantify the distances between observations which is given by

$$d_{i,j} = 1 - r_{i,j}$$

where

$$r_{i,j} = \frac{Cov(i,j)}{\sigma_i \sigma_j}.$$

The dist() function is used to compute the distances in R.

We then used these distances to perform the hierarchical clustering using Ward Linkage. Ward linkage works by specifying the distance between clusters as the increase in the sum of squares if they were to be merged. The sum of squares of a cluster X is given by,

$$SS(X) = \sum_{i=1}^{n_X} |x_i - n_X^{-1} \sum_{j=1}^{n_X} x_j|^2.$$

The algorithm works by starting with each observation as its own cluster then clusters are merged together at each iteration by minimizing the sum of squares increase when new clusters are formed. The sum of squares is calculated using the specified distance metric, which in our case is 1-Pearson's correlation coefficient. The 'hclust()' function with Ward linkage specified preforms the clustering in R.

After the clustering is complete, a dendrogram can be constructed using the groups at each iteration. Then the number of clusters can be decided by visualizing the dendrogram and choosing the number where there is a large separation between steps of the dendrogram. We split our dendrogram into 5 clusters using 'cutree()' which cuts the dendogram and labels each point into its respective cluster. We chose 5 clusters because that is what the paper we are replicated choose as the optimal number.

8 Appendix C

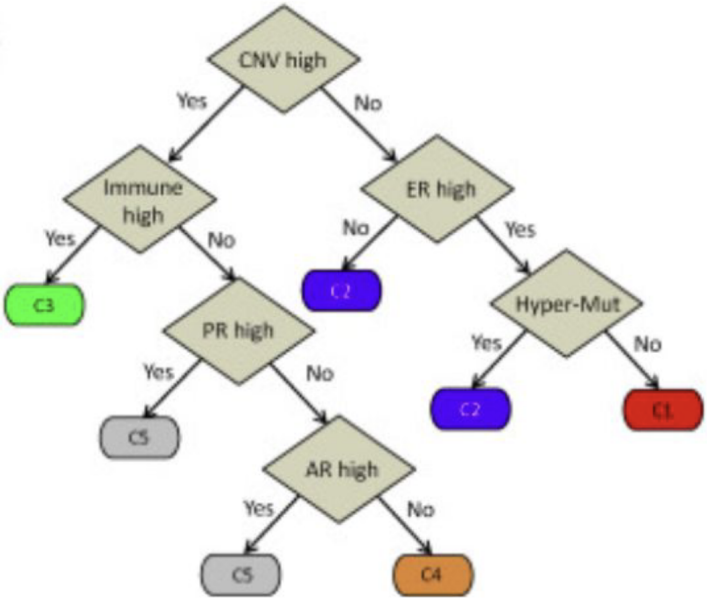


Figure 10: Decision tree constructed by Berger et al. [1].

9 Appendix D

```
Call:
coxph(formula = Surv(days_to_death, vital_status) ~ ajcc_pathologic_stage,
    data = cox_data)

  n= 114, number of events= 114
   (745 observations deleted due to missingness)

                                coef exp(coef) se(coef)     z Pr(>|z|)
ajcc_pathologic_stageStage II  0.02903   1.02946  0.31333 0.093   0.9262
ajcc_pathologic_stageStage III 0.84046   2.31744  0.33291 2.525   0.0116 *
ajcc_pathologic_stageStage IV  0.43523   1.54532  0.39486 1.102   0.2704
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                               exp(coef) exp(-coef) lower .95 upper .95
ajcc_pathologic_stageStage II      1.029     0.9714    0.5571     1.902
ajcc_pathologic_stageStage III     2.317     0.4315    1.2068     4.450
ajcc_pathologic_stageStage IV      1.545     0.6471    0.7127     3.351

Concordance= 0.599  (se = 0.026 )
Likelihood ratio test= 13.28  on 3 df,   p=0.004
Wald test            = 14.04  on 3 df,   p=0.003
Score (logrank) test = 14.7  on 3 df,   p=0.002

[1] "AIC: 851.145538880524"
```

Figure 11: Summary R output for Cox proportional hazard regression of survival on AJCC stage.

```
Call:
coxph(formula = Surv(days_to_death, vital_status) ~ clus + ajcc_pathologic_stage,
    data = cox_data)

  n= 114, number of events= 114
   (745 observations deleted due to missingness)

                               coef exp(coef) se(coef)     z Pr(>|z|)
clus                        0.09487   1.09952  0.07545 1.257  0.20858
ajcc_pathologic_stageStage II  0.09978   1.10493  0.31895 0.313  0.75439
ajcc_pathologic_stageStage III 0.89466   2.44651  0.33660 2.658  0.00786 **
ajcc_pathologic_stageStage IV  0.52024   1.68243  0.40183 1.295  0.19543
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                            exp(coef) exp(-coef) lower .95 upper .95
clus                            1.100     0.9095    0.9484     1.275
ajcc_pathologic_stageStage II      1.105     0.9050    0.5913     2.065
ajcc_pathologic_stageStage III     2.447     0.4087    1.2648     4.732
ajcc_pathologic_stageStage IV      1.682     0.5944    0.7654     3.698

Concordance= 0.616  (se = 0.03 )
Likelihood ratio test= 14.87  on 4 df,   p=0.005
Wald test            = 15.61  on 4 df,   p=0.004
Score (logrank) test = 16.26  on 4 df,   p=0.003

[1] "AIC: 851.562822270238"
```

Figure 12: Summary R output for Cox proportional hazard regression of survival on AJCC stage and cluster.

10 References

[1] Berger et al. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers, 2018

[2] Tremont et al. Endocrine Therapy for Early Breast Cancer: Updated Review, 2017