

Predicting Breast Cancer Survival Rates and Cancer Type

Arthur Chen
Beth Choi
Christine Chun
Danrui Wang

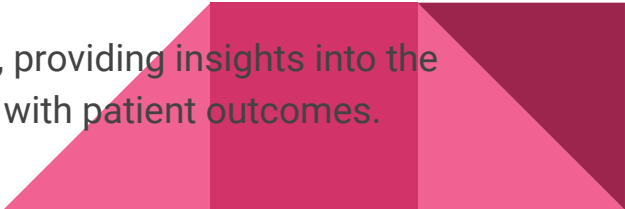
Introduction

- Objective: discover ways to predict survival rates by analyzing and finding patterns within existing clinical and genomic data
- Approach:
 - Preprocessing and cleaning of data
 - Understanding attributes and terminology
 - Training, validating and evaluating model performance
- Other goals considered:
 - predicting response to treatments
- Key stakeholder/target audience:
 - Patients
 - Clinicians
 - Researchers
 - Insurance Providers
 - Pharmaceutical Companies
 - Patient Advocacy Groups

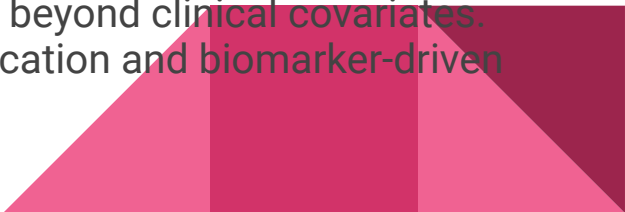


Literature Review

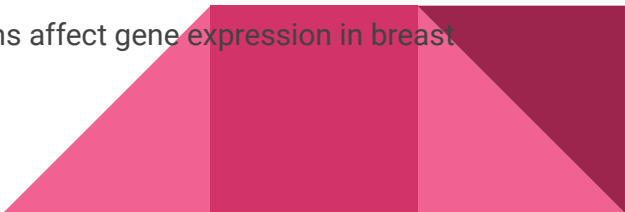
The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes

- Breast cancer is a heterogeneous disease with varied outcomes and responses to treatment, even in patients with similar clinical characteristics.
 - An integrated analysis of somatic copy number aberrations (CNAs) and gene expression profiles in 2,000 primary tumors was performed to better understand the genomic drivers behind this heterogeneity.
 - The analysis resulted in the identification of 10 Integrative Clusters (IntClusters) with distinctive copy number profiles and clinical courses.
 - A copy number-based classification has been validated in 7,500 tumors.
 - Somatic SNVs and indels in driver genes also contribute to tumor biology in addition to CNAs.
 - 173 of the most frequently mutated breast cancer genes were sequenced in 2,433 primary tumors to investigate the clinical significance of mutations in these genes.
 - Associations between genomic and clinical features were identified, providing insights into the mutation profiles of key breast cancer genes and their associations with patient outcomes.
- 

Dynamics of breast-cancer relapse reveal late-recurring ER-positive genetic subgroups

- Breast cancer rates and routes of lethal systemic spread are poorly understood due to the lack of molecularly characterized cohorts with long-term follow-up.
 - A statistical framework was presented to model distinct disease stages and individual risk of recurrence predictions.
 - The model was applied to 3240 breast cancer patients, including 1980 with molecular data, across various subtypes (IHC, PAM50, IntClust).
 - Four late-recurring integrative subtypes of ER+, HER2- tumors were identified, each with characteristic genomic copy number driver alterations and high risk of recurrence up to 20 years post-diagnosis.
 - A subgroup of triple-negative breast cancers rarely recur after 5 years, while a separate subgroup remains at risk.
 - Integrative subtypes improve prediction of late distant relapse beyond clinical covariates.
 - These findings offer opportunities for improved patient stratification and biomarker-driven clinical trials.
- 

The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups

- A population-based molecular subgrouping of breast cancer based on multiple genomic views was generated
 - CNAs and SNPs influence expression variation, with CNAs dominating the landscape in cis and trans
 - Joint clustering of CNAs and gene expression profiles resolves heterogeneity of expression-only subgroups and highlights high-risk 11q13/14 cis-acting subgroup and several other strong cis-acting clusters and a genomically quiescent group
 - Reproducibility of subgroups with molecular and clinical features in a validation cohort of 995 tumors suggests that integrating multiple genomic features can derive more robust patient classifiers
 - Subtype-specific trans-acting aberrations modulate concerted transcriptional changes, such as TCR deletion-mediated adaptive immune response characterizing the CNA-devoid subgroup and chromosome 5 deletion-associated cell cycle program in the basal cancers
 - Integrated CNA-expression landscape highlights limited number of genomic regions probably containing driver genes, including ZNF703, somatic deletion events affecting key subunits of the PP2A holoenzyme complex and MTAP, IGF1R, KRAS and EGFR amplifications, and CDKN2B, BRCA2, RB1, ATM, SMAD4, NCOR1, and UTX homozygous deletions
 - Focusing sequencing efforts on representative numbers from these groups will help establish comprehensive breast cancer somatic landscape at sequence-level resolution
 - A significant number of breast cancers are devoid of somatic CNAs and are ripe for mutational profiling (approximately 17% in the discovery cohort)
 - Provides a definitive framework for understanding how gene copy number aberrations affect gene expression in breast cancer and reveals novel subgroups for future investigation
- 

Methodology

Data Cleaning

- Step 1: Understanding the data
 - First by identifying all the attributes and deciphering any medical jargon, we created a glossary of feature definitions
 - By understanding each attribute we can now utilize the attributes in our code in a way that makes sense for our analysis.
- Step 2: Identify any inconsistencies, errors, or missing values
 - Our dataset began with 2,509 patient samples, once missing data were identified they were then removed
 - We also identified a subtype that was generic and only labeled as “breast”. This vague inconsistency was also removed in the cleaning process
 - After the aforementioned inconsistencies, errors and missing values were removed the dataset was reduced to 1078 patient samples
- Step 3: Validate the data
 - In this step the data was reviewed thoroughly for additional errors or inconsistencies
- Step 4: Verify the data
 - By using coding tools we ensure our data is usable and workable.

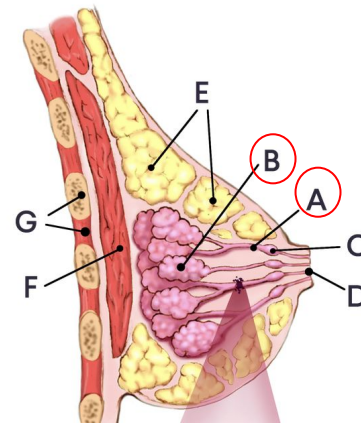




Clinical Survival Analysis Results

Definitions

Breast Cancer types	Number of Patients	Percentage
Breast Invasive Ductal Carcinoma	852	79%
Breast Mixed Ductal and Lobular Carcinoma	137	12.7%
Breast Invasive Lobular Carcinoma	76	7.1%
Breast Invasive Mixed Mucinous Carcinoma	13	1.2%



- Lobular: cancer started in the milk-producing glands
- Ductal: cancer started in the milk ducts
- Invasive: cancer has spread into surrounding breast tissues
- Mucinous: cancer cells are surrounded by mucin

Definitions

- **Tumor Stages**

- **1:** tumor size is 2 cm (3/4 of an inch) or less across
- **2:** tumor size is more than 2 cm but not more than 5 cm (2 inches) across
- **3:** tumor size is more than 5 cm across
- **4:** tumor of any size growing into the chest wall or skin (includes inflammatory breast cancer)

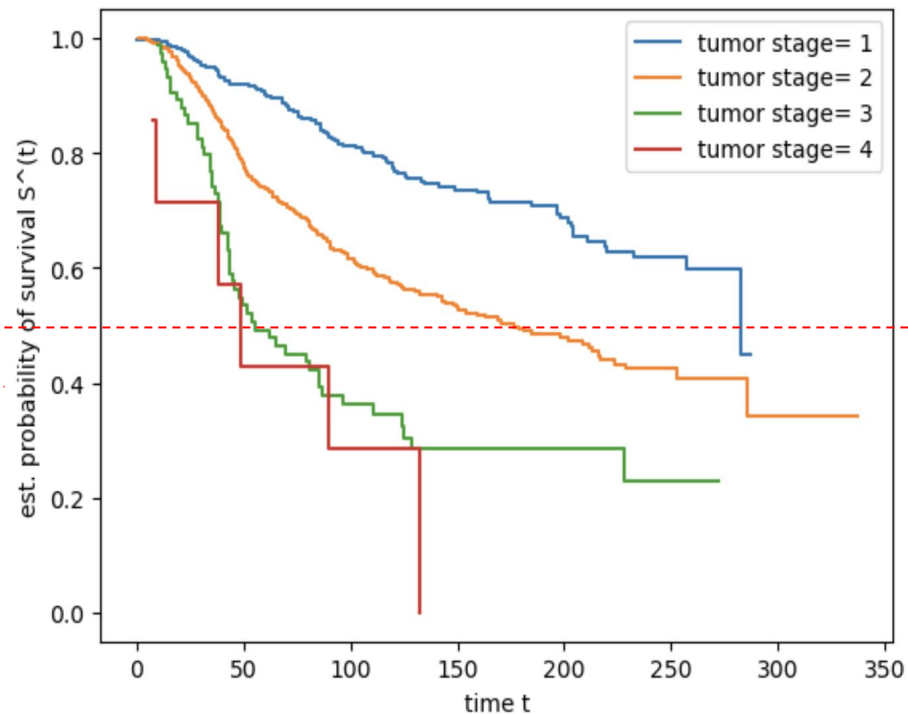
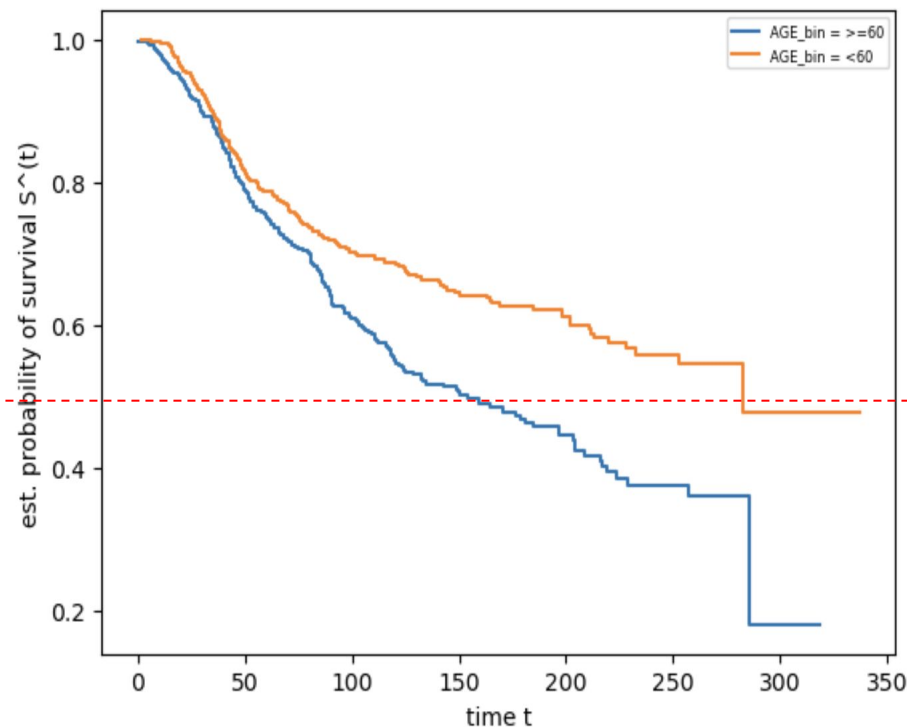
- **ER Status (Positive vs. Negative)**

- Presence or absence of estrogen receptors in a tumor. Tumors that are estrogen receptor positive are more sensitive to certain treatments, such as hormone therapy, and are associated with a better prognosis than tumors that are estrogen receptor negative.

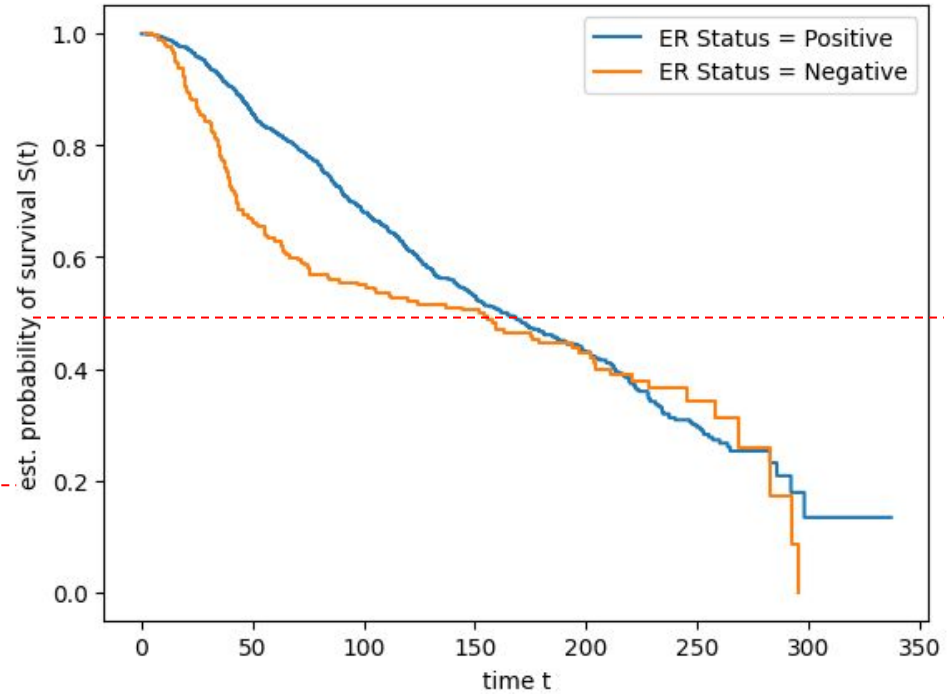
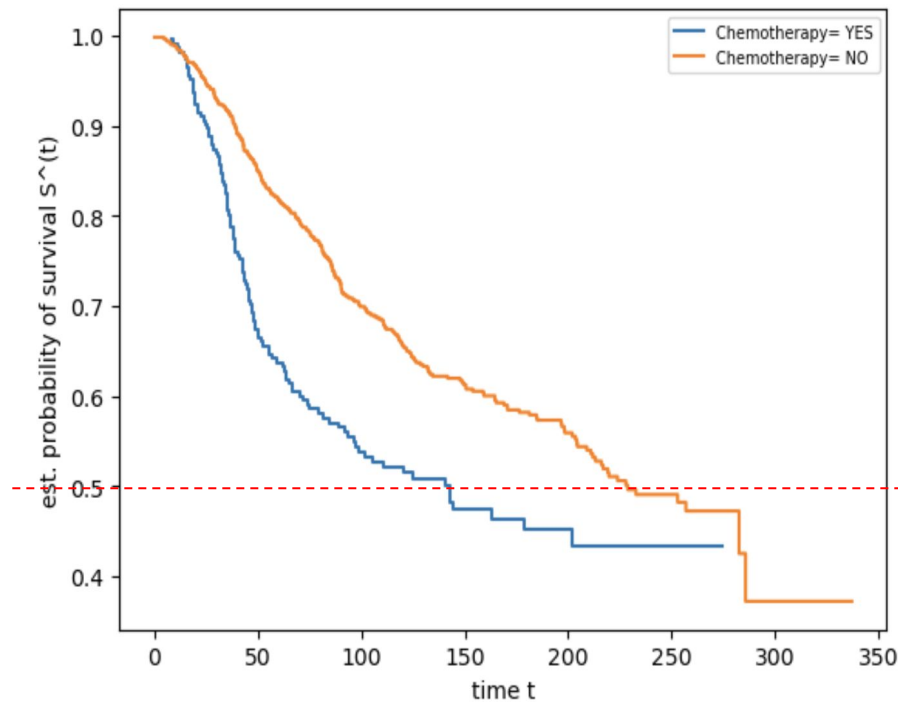
- **Relapse Free Status (Recurred vs. Not Recurred)**

- Recurred, if cancer is found after treatment and after a period of time when the cancer couldn't be detected

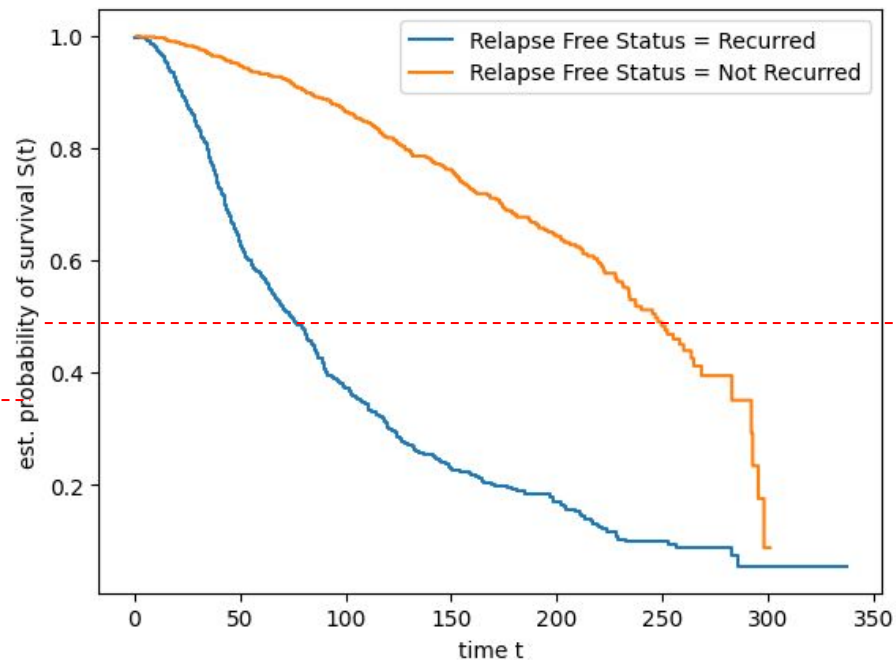
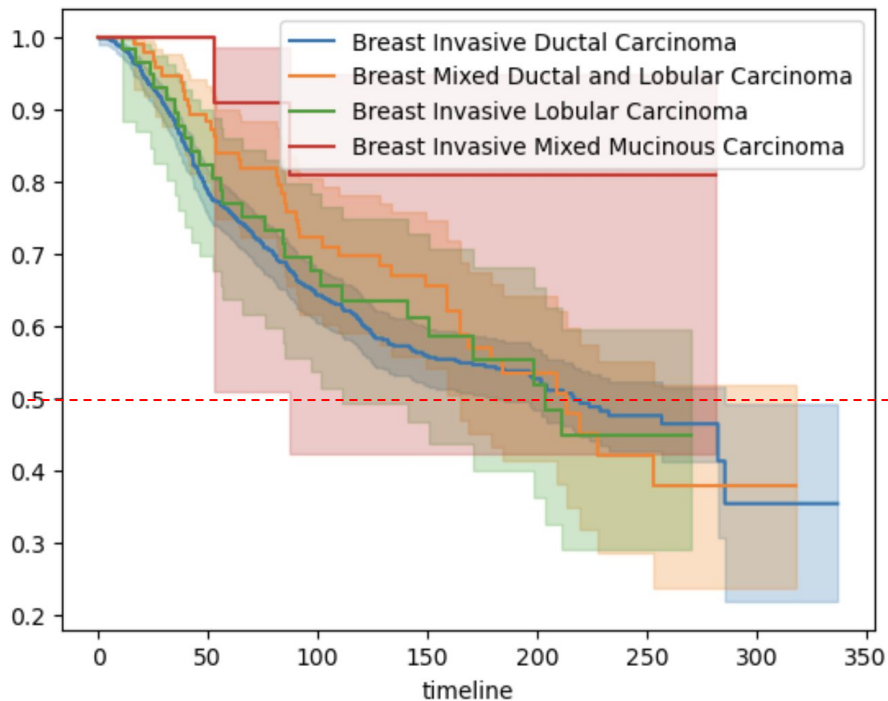
Preliminary Results - Age, Tumor Stage



Preliminary Results - Chemotherapy, ER Status



Preliminary Results - Subtypes of Breast Cancer, Relapse Free Status



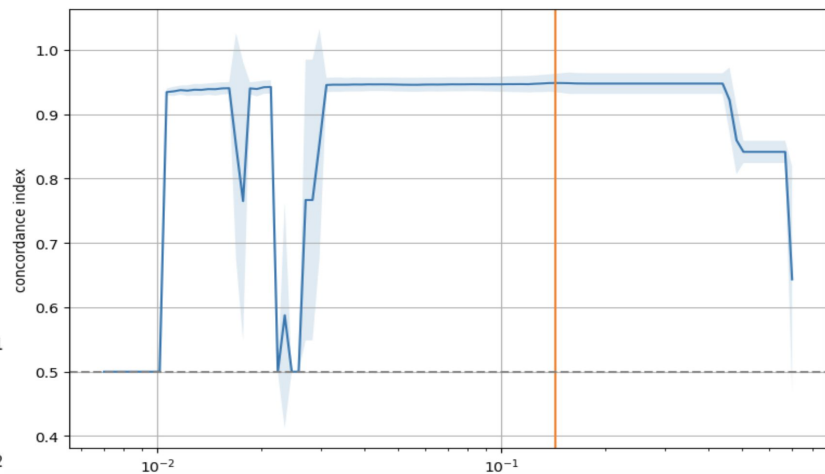
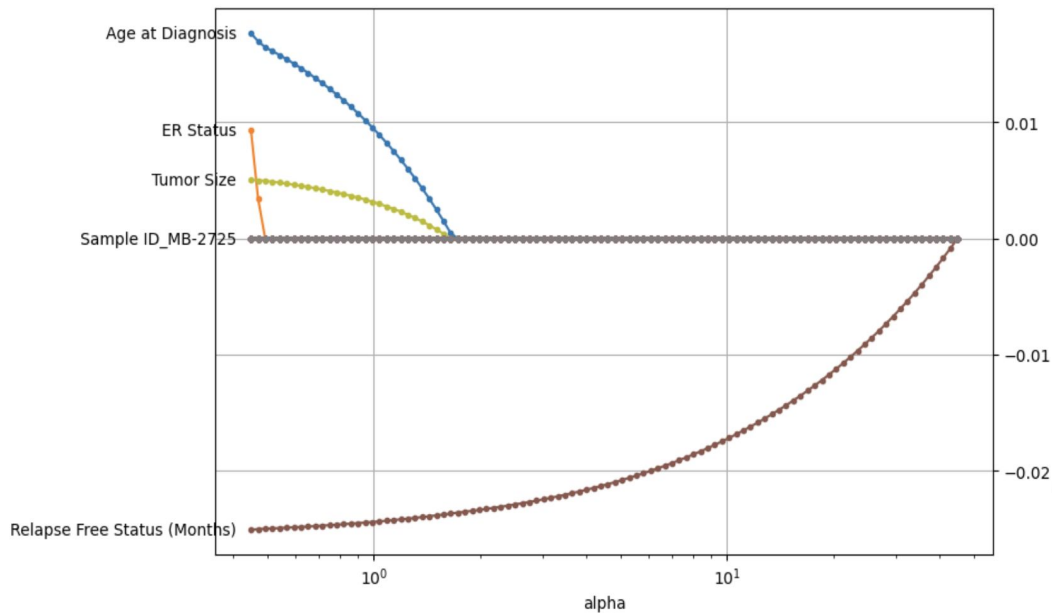
Clinical Machine Learning Model

Methodology

- **Models/Approaches**
 - Cox Proportional-Hazards Model (Elastic Net Penalized)
- **QA/QC Work**
 - Focusing on death by disease only
 - Only chose variables with p-values lower than 0.05
 - Age at Diagnosis
 - ER Status: 0 - positive, 1 - negative
 - Tumor Size
 - Relapse Free Status (Months)



Penalized Cox Model & Best Alpha



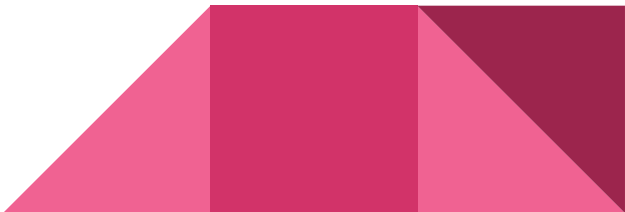
Genomic Cancer Type Prediction

Methodology

- Models/Approaches
 - mRNA: Log-transformed mRNA z-scores standardize gene expression values for comparison across samples. This results in a normalized value with mean 0 and standard deviation 1, enabling comparisons across genes and samples.
 - DNA: Copy-number alterations from DNA copy. Calls made after normal contamination correction and CNV removal using thresholds
- QA/QC work
 - mRNA: Our dataset began with 1980 patient samples, and 20603 types of genes, after the missing values were removed the dataset was reduced to 1964 patient and 20603 types of genes
 - DNA: Our dataset began with 2173 patient samples, and 22544 types of genes, after the missing values and type 6th & 7th cancer were removed the dataset was reduced to 2091 patient and 22544 types of genes

DNA data explain –

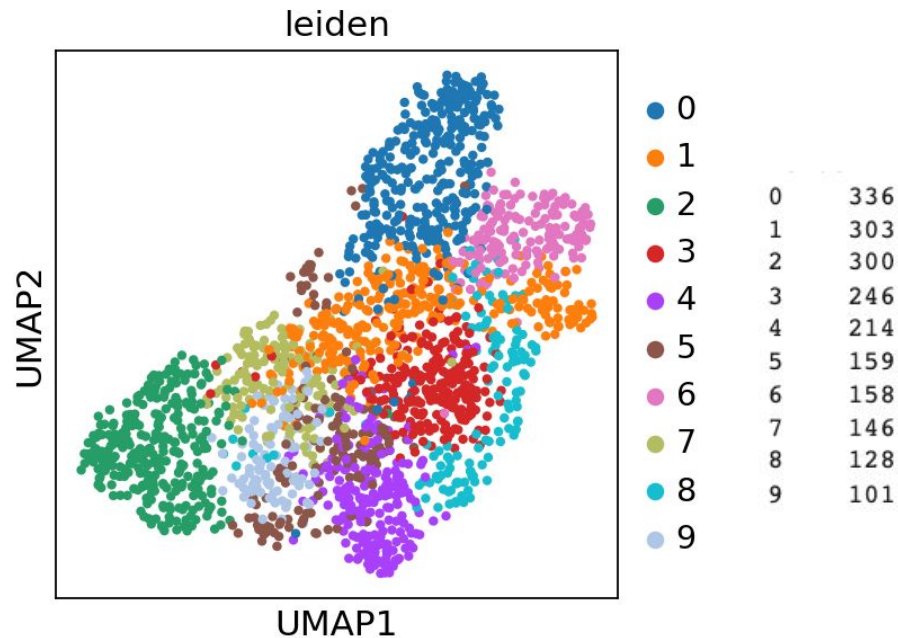
- 2 : Deep Deletion, indicates a deep loss, possibly a homozygous deletion
- 1 : Shallow Deletion, indicates a shallow loss, possibly a heterozygous deletion
- 0 : diploid
- 1 : Gain indicates a low-level gain (a few additional copies, often broad)
- 2 : Amplification indicate a high-level amplification (more copies, often focal)



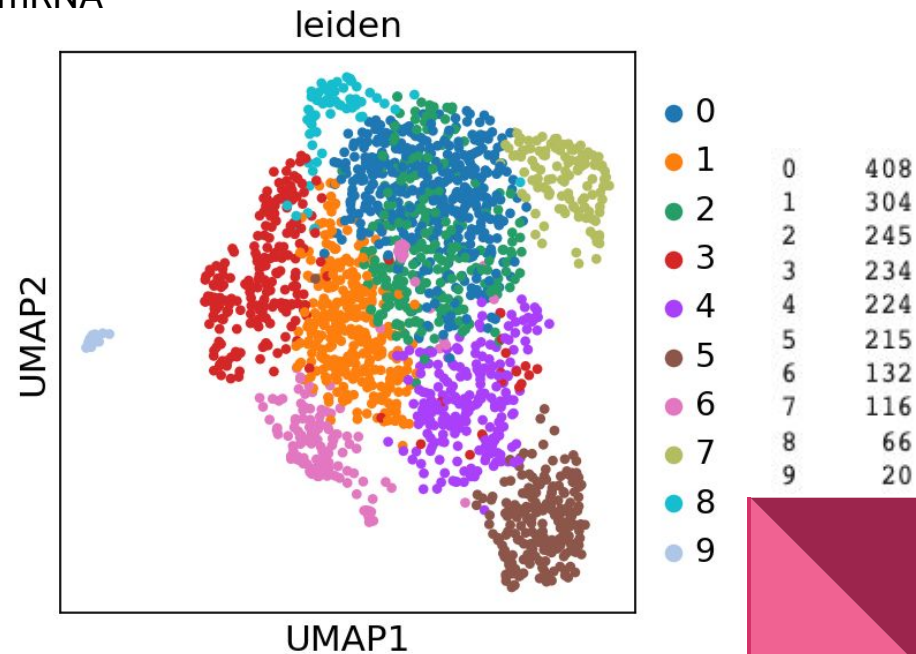
Preliminary Analysis - Gene Clustering

Leiden + UMAP

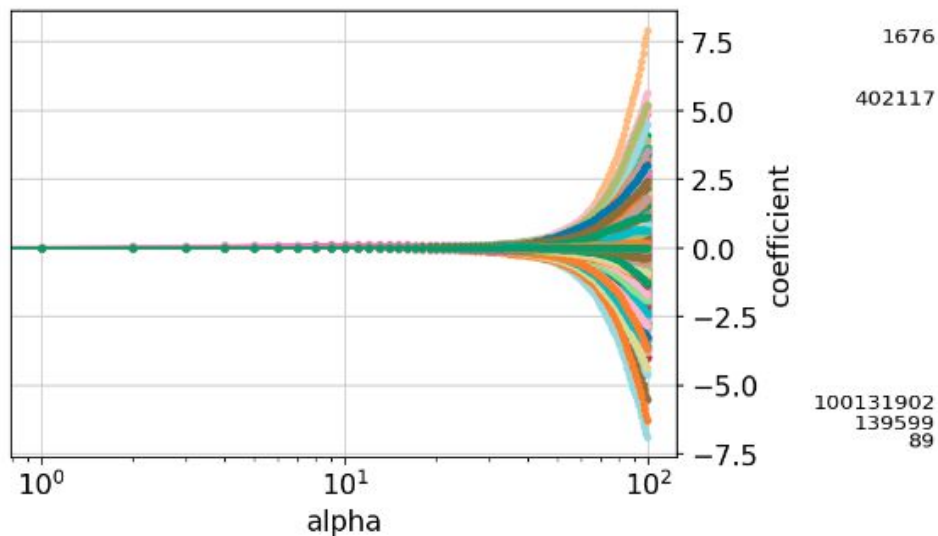
DNA



mRNA



Preliminary Results - mRNA Survival Analysis



Top 5 genes that have highest coefficient :

- DFFA(1676)
- VWC2L (402117)
- KRTAP 25-1 (100131902)
- MAGEE2 (139599)
- ACTN3 (89)



Genomic Data Models

Methodology (Models/Approaches)

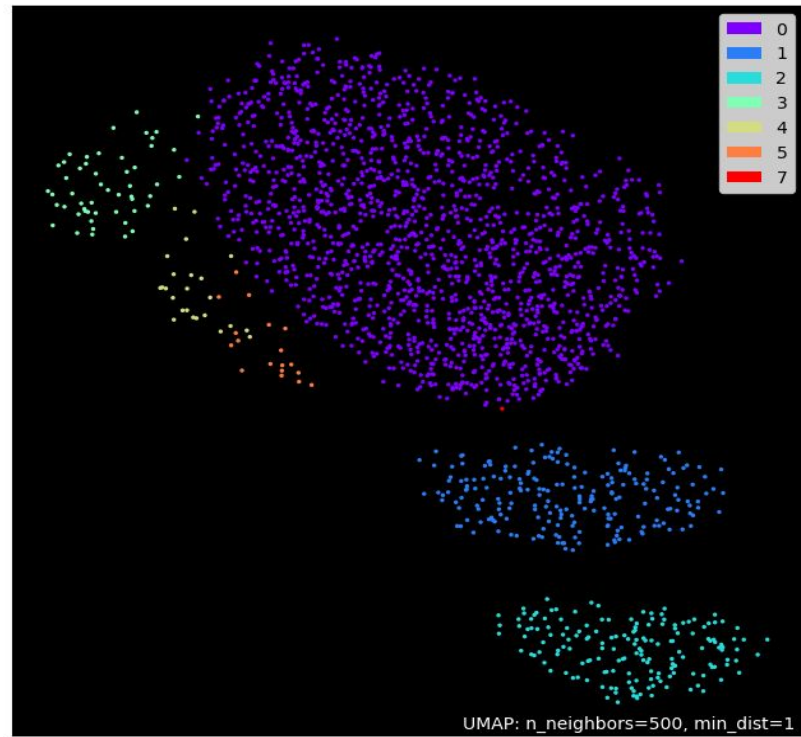
- Models/approaches:
 - Random Forest
 - Logistic Regression
 - Logistic Regression + PCA
 - SVM with UMAP
 - Knn with UMAP
- QA/QC work
 - Dropped 6th & 7th breast cancer types (as dataset was skewed)



Genomic Data (Process)

Reasons of why we dropped
6th & 7th breast cancer types

Breast Invasive Ductal Carcinoma	1865
Breast Mixed Ductal and Lobular Carcinoma	269
Breast Invasive Lobular Carcinoma	192
Invasive Breast Carcinoma	133
Breast Invasive Mixed Mucinous Carcinoma	25
Breast	21
Breast Angiosarcoma	2
Metaplastic Breast Cancer	2



Cancer Type Prediction - Random Forest

```
forest = RandomForestClassifier(n_estimators=1000, class_weight='balanced').fit(x_train, y_train)
```

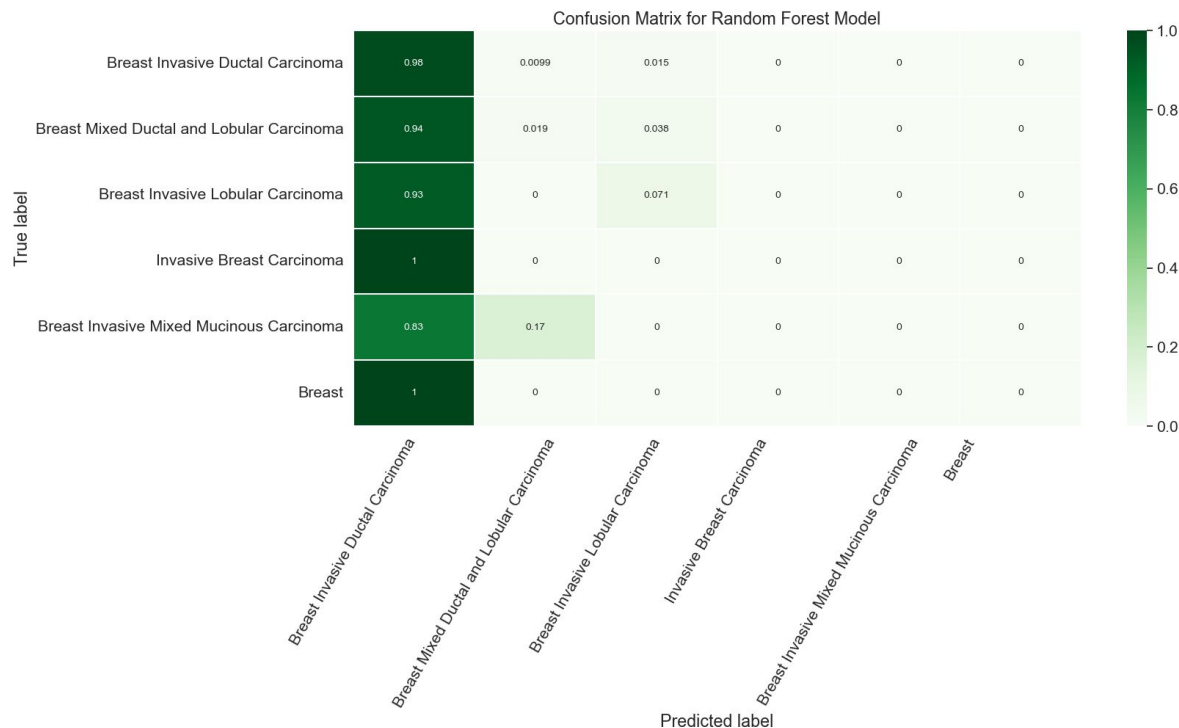
Performance

Accuracy : 0.77

Roc_auc_score : 0.94

Confusion Matrix

	precision	recall	f1-score	support
0	0.78	0.98	0.87	404
1	0.20	0.02	0.03	53
2	0.22	0.05	0.08	42
3	0.00	0.00	0.00	13
4	0.00	0.00	0.00	6
5	0.00	0.00	0.00	5
accuracy			0.76	523
macro avg	0.20	0.17	0.16	523
weighted avg	0.64	0.76	0.68	523



Cancer Type Prediction - Logistic Regression

```
lr = LogisticRegression(solver="lbfgs", class_weight='balanced', max_iter = 2500).fit(x_train, y_train)
```

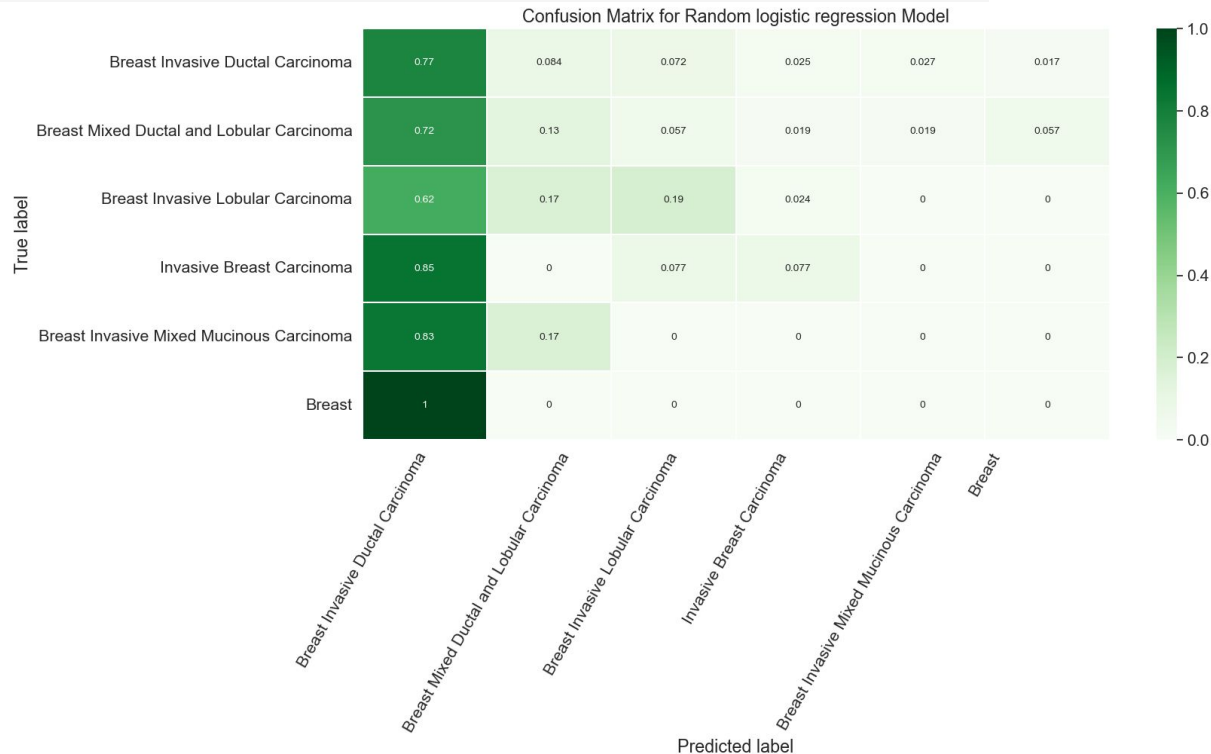
Performance

Accuracy : 0.63

Roc_auc_score : 0.89

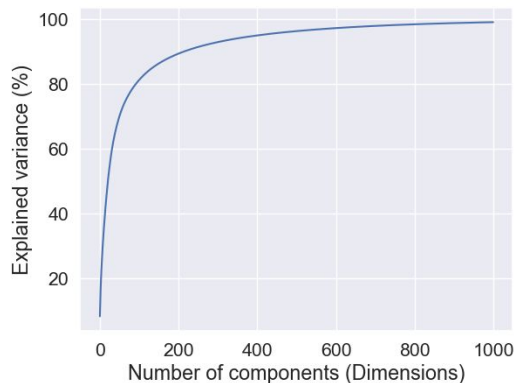
Confusion Matrix

	precision	recall	f1-score	support
0	0.79	0.77	0.78	404
1	0.14	0.13	0.14	53
2	0.20	0.19	0.19	42
3	0.08	0.08	0.08	13
4	0.00	0.00	0.00	6
5	0.00	0.00	0.00	5
accuracy			0.63	523
macro avg	0.20	0.20	0.20	523
weighted avg	0.64	0.63	0.63	523



Cancer Type Prediction - Logistic Regression + PCA

250 components explain nearly **90%**
of the variance in the data



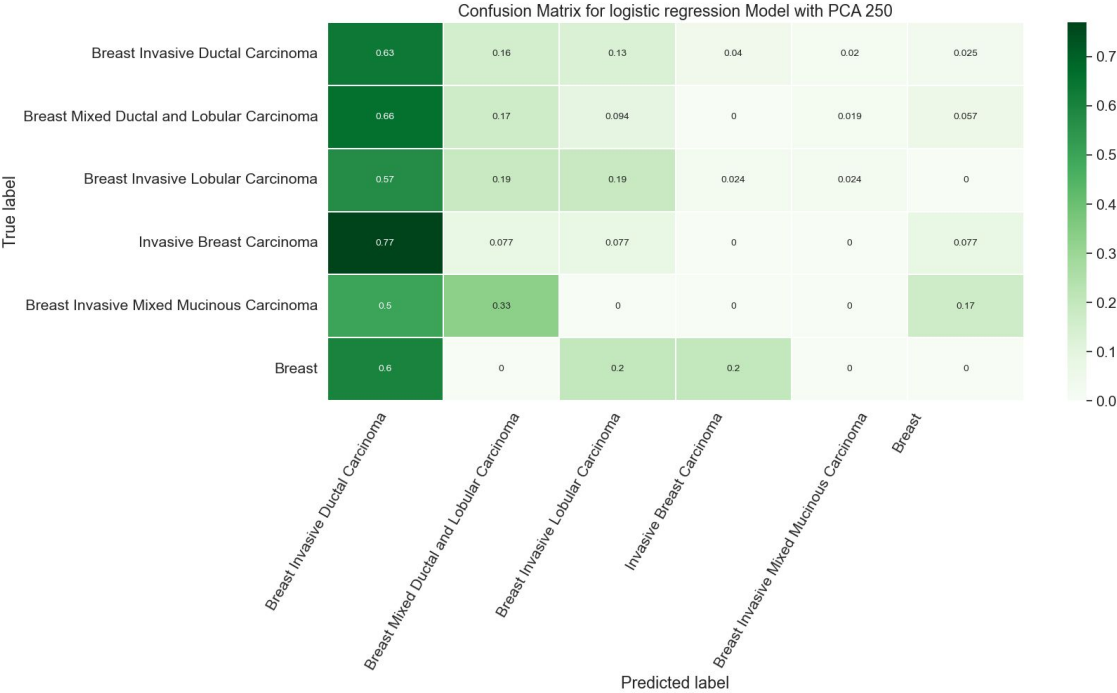
```
pca = PCA(n_components=250)
pca.fit(X_scaled)
```

```
X_pca = pca.fit_transform(X_scaled)
X_pca = pd.DataFrame(X_pca)
X_train_pca, X_test_pca, y_train, y_test = train_test_split(X_pca, y, test_size=0.25,
                                                             shuffle=True, random_state=2, stratify=y)
lr = LogisticRegression(solver="lbfgs", class_weight='balanced', max_iter = 5000)
```

```
[8.21199067e+00 4.76682367e+00 4.16871214e+00 3.07839456e+00
2.67028646e+00 2.38990215e+00 2.30381343e+00 2.05469544e+00
1.94401882e+00 1.83436344e+00 1.75118577e+00 1.68465203e+00
1.63801384e+00 1.53450433e+00 1.50355317e+00 1.47980112e+00
1.38550191e+00 1.35334060e+00 1.32990753e+00 1.31114232e+00
1.23404610e+00 1.15880977e+00 1.10490843e+00 1.03363342e+00
1.01035074e+00 9.80002507e-01 9.51456304e-01 9.46204073e-01
8.80922095e-01 8.53797698e-01 7.82142663e-01 7.56186643e-01
7.38520983e-01 6.95372214e-01 6.71926480e-01 6.37045150e-01
6.21065421e-01 5.92751365e-01 5.54816263e-01 5.36490958e-01
5.27369197e-01 5.17697177e-01 4.98226352e-01 4.77713918e-01
4.59842513e-01 4.57801106e-01 4.41073953e-01 4.26847174e-01
4.16561979e-01 4.07100463e-01 3.89591914e-01 3.74583567e-01
3.71823934e-01 3.51809947e-01 3.40771337e-01 3.33004506e-01
3.30003244e-01 3.13450790e-01 3.07968657e-01 3.02548331e-01
2.94325050e-01 2.85409103e-01 2.70339659e-01 2.64900433e-01
2.56025746e-01 2.55914790e-01 2.44247718e-01 2.41349520e-01
```

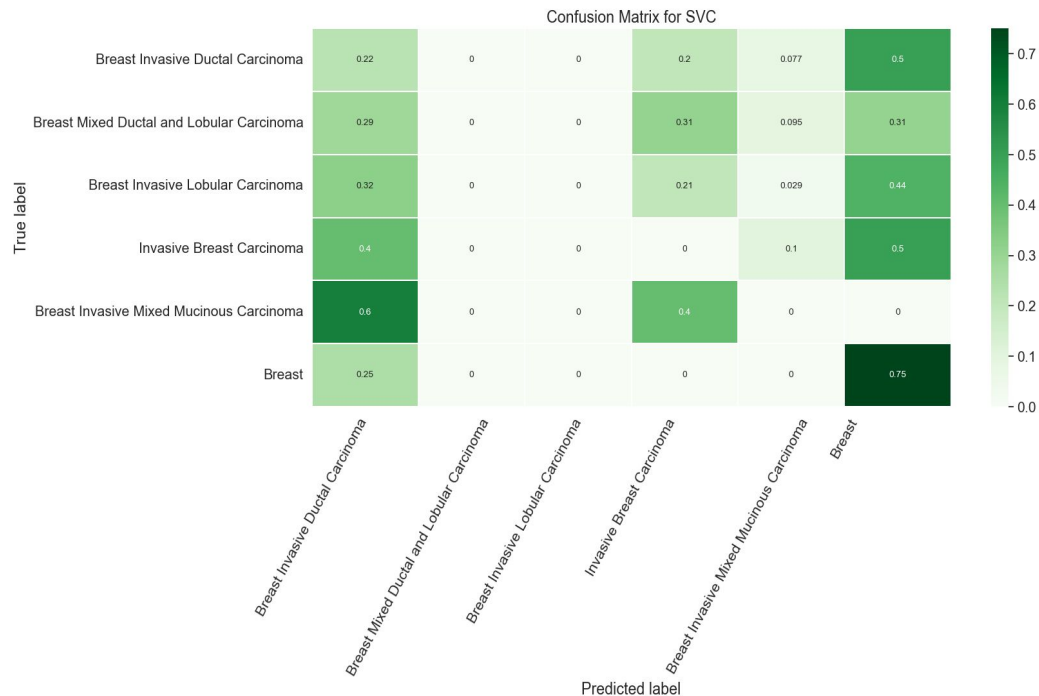
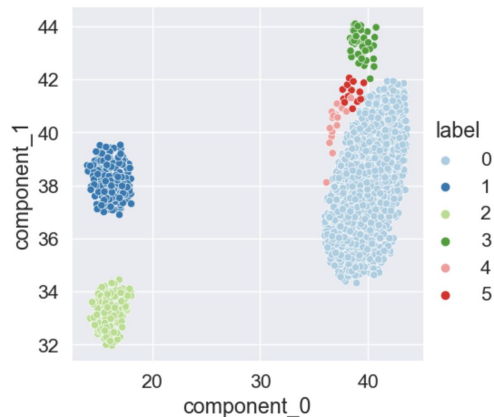
Cancer Type Prediction - Logistic Regression + PCA

	precision	recall	f1-score	support
0	0.81	0.54	0.65	404
1	0.15	0.28	0.20	53
2	0.12	0.21	0.15	42
3	0.03	0.08	0.04	13
4	0.05	0.17	0.08	6
5	0.00	0.00	0.00	5
accuracy			0.47	523
macro avg	0.19	0.21	0.19	523
weighted avg	0.65	0.47	0.54	523



Cancer Type Prediction - SVM with UMAP

```
mapper = umap.UMAP(n_neighbors = 1000, min_dist = 1).fit(x_train, y_train)
```



	precision	recall	f1-score	support
0	0.70	0.22	0.33	323
1	0.00	0.00	0.00	42
2	0.00	0.00	0.00	34
3	0.00	0.00	0.00	10
4	0.00	0.00	0.00	5
5	0.02	0.75	0.03	4
accuracy			0.18	418
macro avg	0.12	0.16	0.06	418
weighted avg	0.54	0.18	0.26	418

Cancer Type Prediction - KNN with UMAP

```
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=6)
knn.fit(x_train_umap, y_train_umap)
y_pred_knn=knn.predict(X_test_umap)
accuracy_score(y_pred_knn,y_test_umap)
```

	precision	recall	f1-score	support
0	0.77	0.84	0.80	323
1	0.00	0.00	0.00	42
2	0.00	0.00	0.00	34
3	0.00	0.00	0.00	10
4	0.00	0.00	0.00	5
5	0.00	0.00	0.00	4
accuracy			0.65	418
macro avg	0.13	0.14	0.13	418
weighted avg	0.59	0.65	0.62	418



Conclusions

Takeaways/Potential Improvement

- Key Takeaways:

- Clinical Model


- Older patients have higher risk of death than younger patients.
 - Patients with negative ER status have higher risk of death than patients with positive ER status.
 - Patients with bigger tumor size have higher risk of death than patients with smaller tumor size.
 - Patients with longer duration of relapse free status have less risk of death than patients with shorter duration of relapse free status.

- Genomic Model

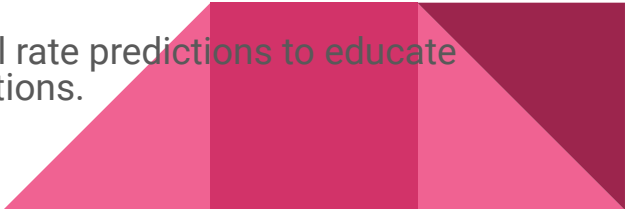
- When doing research in the future, many genes can possibly be omitted because 250 genes explain nearly 90% of the variance in the data.
 - Simply using DNA information to predict cancer type may not be sufficient because there are many breast cancers that are non-hereditary, and may require the combination of mRNA information or other personal-related information.
 - Random Forest algorithm has been observed to outperform other algorithms in dealing with imbalanced datasets. It can adaptively adjust to the distribution of the dataset, making the classifier more accurate in identifying minority categories when handling imbalanced data. Therefore, Random Forest can be a suitable choice for such scenarios.

- Potential Improvement:

- Genomic Model

- The data we were working with was imbalanced, lowering the accuracy
 - The data after lowering the dimension still somehow inseparable
 - The sample size of the dataset is too small.
- 

How would the stakeholder/target audience use the takeaways?

- **Patients:** Accurate survival rate predictions can help them make informed decisions about their treatment options and post-treatment care.
 - **Clinicians:** They use survival rate predictions to develop personalized treatment plans and monitor patient progress.
 - **Researchers:** Predictive models based on clinical and genomic data can help them identify new treatment targets and develop more effective therapies.
 - **Insurance providers:** Insurance providers may use survival rate predictions to make coverage decisions and determine the cost of care.
 - **Pharmaceutical companies:** Pharmaceutical companies may use survival rate predictions to develop and test new drugs and therapies.
 - **Patient advocacy groups:** Patient advocacy groups may use survival rate predictions to educate patients and their families about breast cancer and its treatment options.
- 

Learnings from the project

- What we would do differently:
 - Explore other datasets on breast cancer, available on cBioPortal
- Unexpected challenges:
 - Difficult medical terminologies
 - Dataset was very skewed, in terms of breast cancer types
- New skills that were developed during the project:
 - Code collaboration
 - Exploring new machine learning models
 - Working with medical datasets
- Things we wish we got to do but didn't:
 - Using MSK's DeepLIIF project for translating and segmenting tissue images with immunohistochemical (IHC) staining



References

- Center for Molecular Oncology. "Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016)." CBioPortal for Cancer Genomics, Memorial Sloan Kettering Cancer Center, http://www.cbioportal.org/study/clinicalData?id=brca_metabric.
- Pereira B, Chin SF, Rueda OM, Vollan HK, Provenzano E, Bardwell HA, Pugh M, Jones L, Russell R, Sammut SJ, Tsui DW, Liu B, Dawson SJ, Abraham J, Northen H, Peden JF, Mukherjee A, Turashvili G, Green AR, McKinney S, Oloumi A, Shah S, Rosenfeld N, Murphy L, Bentley DR, Ellis IO, Purushotham A, Pinder SE, Børresen-Dale AL, Earl HM, Pharoah PD, Ross MT, Aparicio S, Caldas C. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat Commun. 2016 May 10;7:11479. doi: 10.1038/ncomms11479. PMID: 27161491; PMCID: PMC4866047.
- Rueda OM, Sammut SJ, Seoane JA, Chin SF, Caswell-Jin JL, Callari M, Batra R, Pereira B, Bruna A, Ali HR, Provenzano E, Liu B, Parisien M, Gillett C, McKinney S, Green AR, Murphy L, Purushotham A, Ellis IO, Pharoah PD, Rueda C, Aparicio S, Caldas C, Curtis C. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. Nature. 2019 Mar;567(7748):399-404. doi: 10.1038/s41586-019-1007-8. Epub 2019 Mar 13. PMID: 30867590; PMCID: PMC6647838.
- Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S; METABRIC Group; Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowitz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale AL, Brenton JD, Tavaré S, Caldas C, Aparicio S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012 Apr 18;486(7403):346-52. doi: 10.1038/nature10983. PMID: 22522925; PMCID: PMC3440846.