

Breast Cancer Classification

ML 7641: Group 3 Final Project

By: Celine Al-Noubani, FNU Naga Nishkala, Abhishek Vijeev, Luis Andres Casavilca,
Nikhil Sundaram

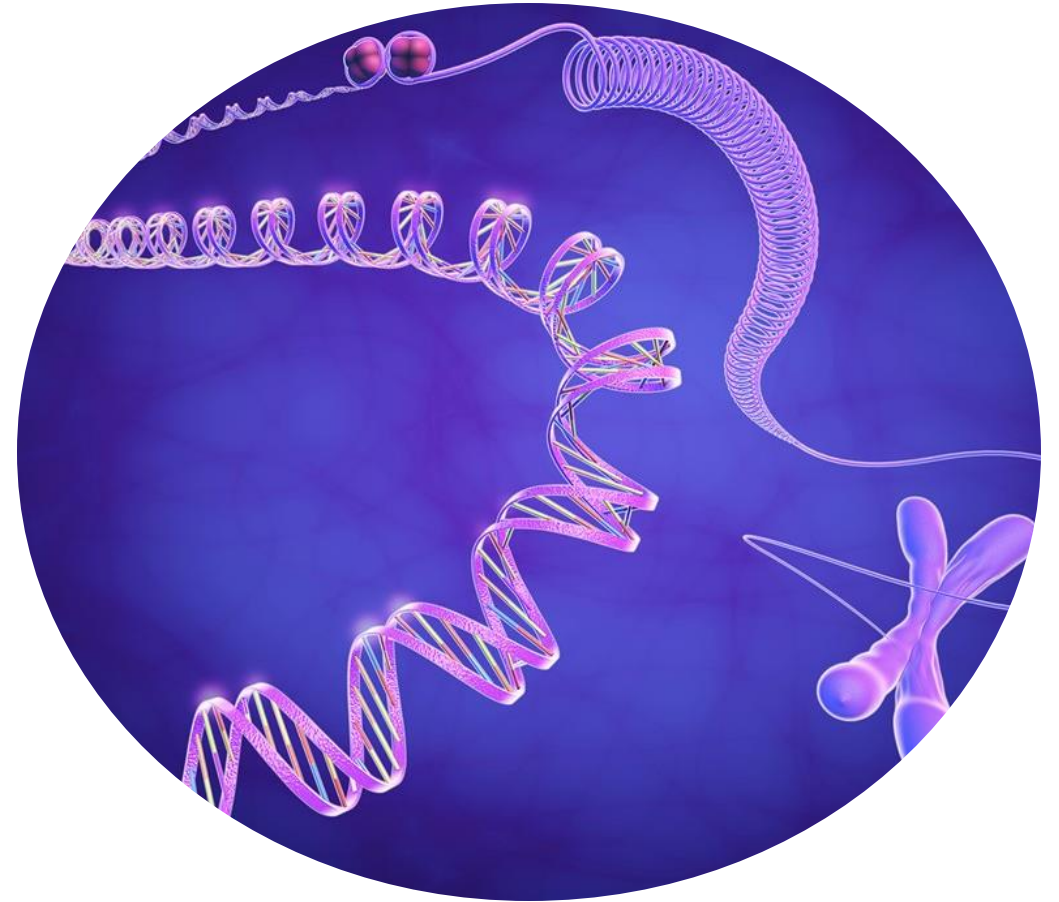
Breast Cancer Motivation

- One of the most prevalent and life-threatening forms of cancer among women worldwide
- 1 in 20 women worldwide will be diagnosed with breast cancer in their lifetime [1].
- High degree of heterogeneity @ molecular level [2].



METABRIC Dataset [3]

- Breast cancer gene expression profiles for 1,904 breast cancer patients
 - Includes 31 clinical attributes
 - m-RNA levels z-score for 331 genes
 - mutation data for 175 genes

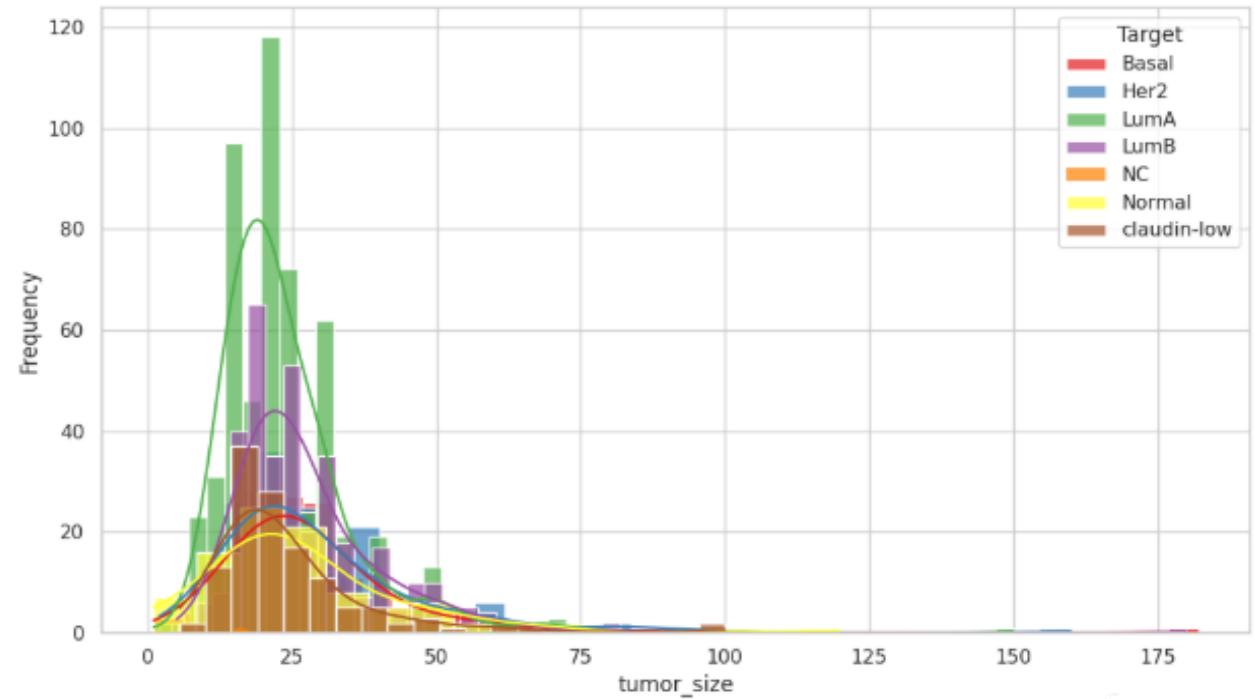
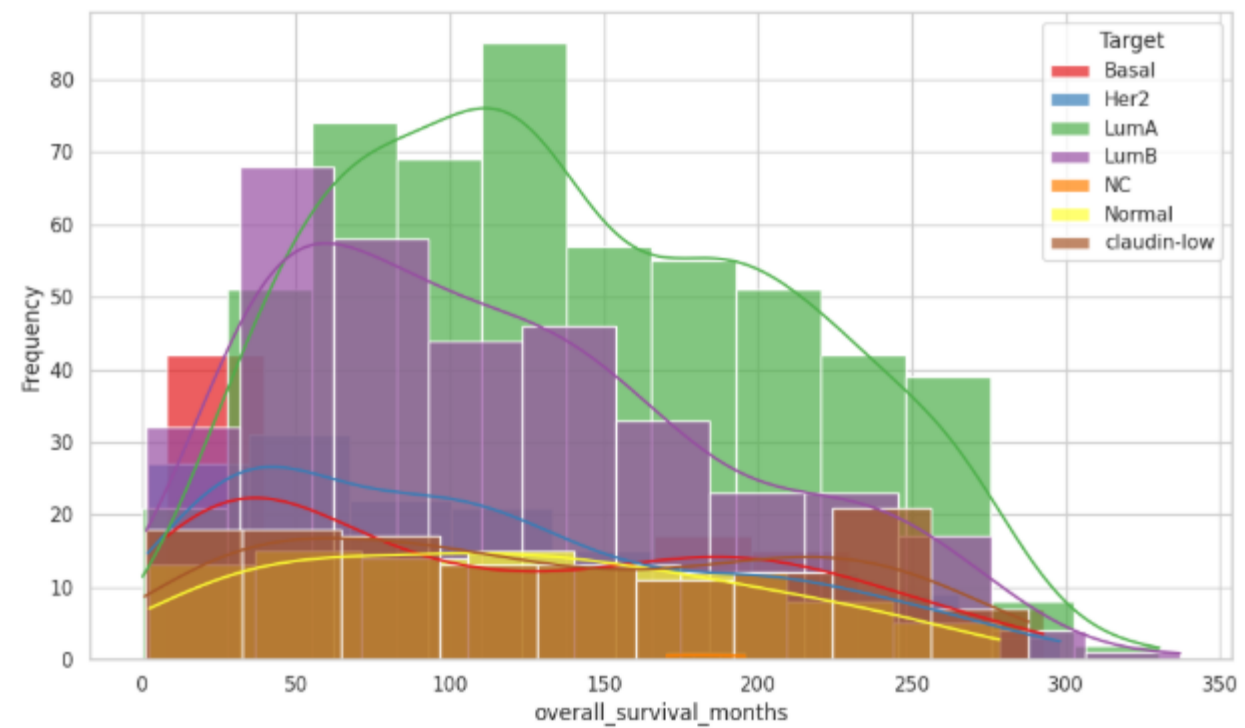
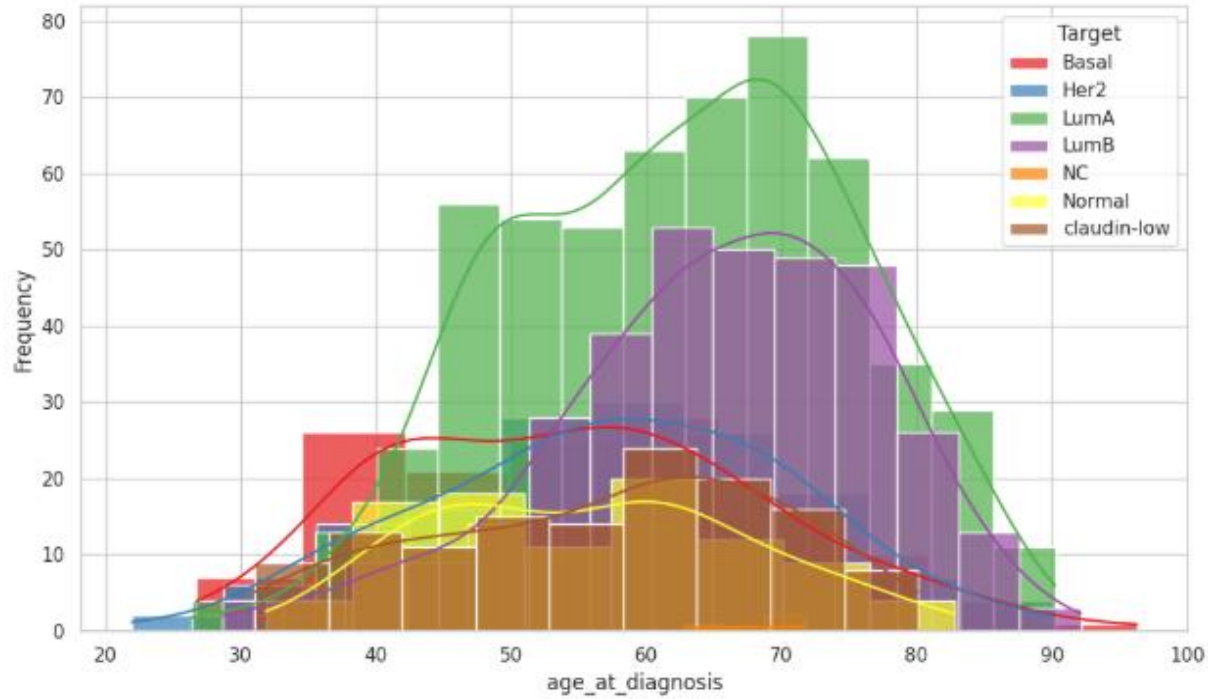


EDA 🔍

Exploratory Data Analysis

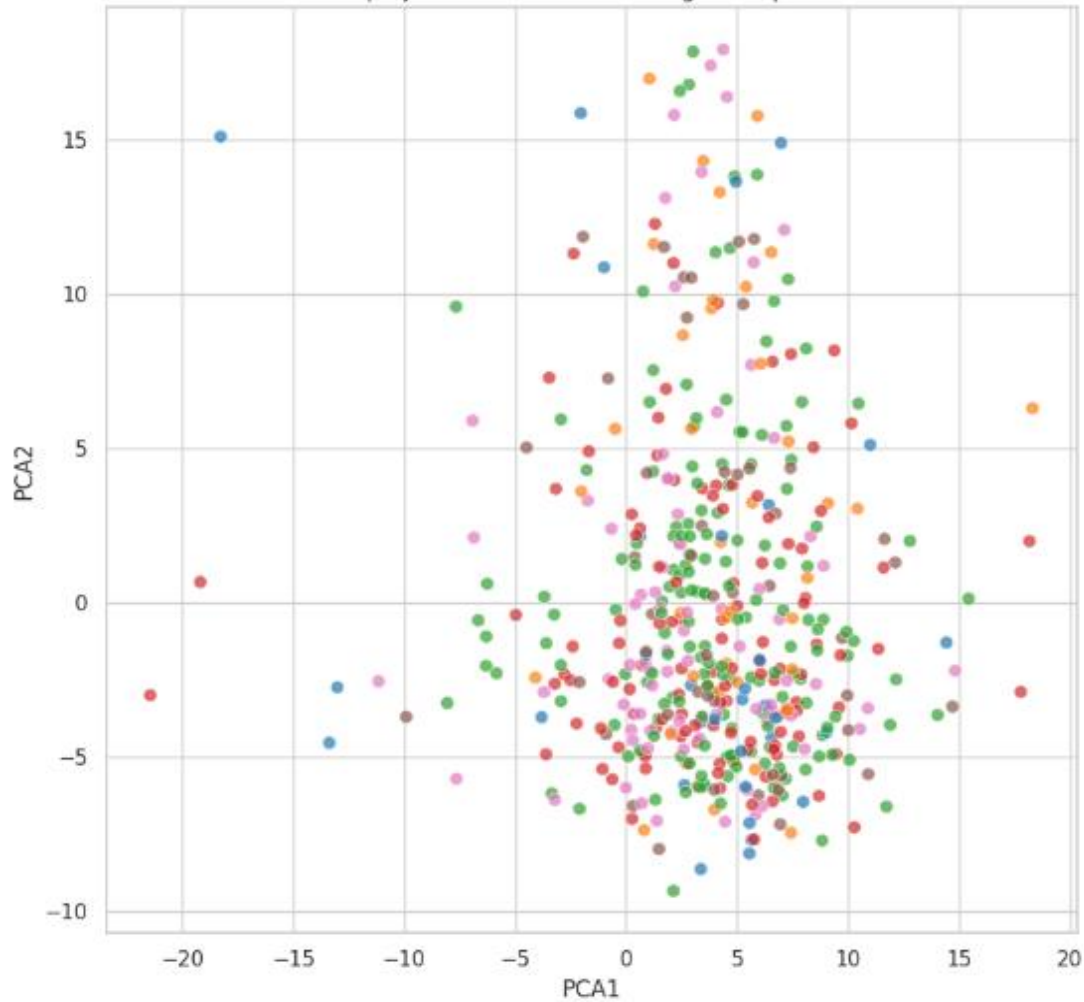
1. Split clinical and gene data
2. Visualized class distributions across clinical data attributes (via histograms & boxplots)
3. Pairplot of numerical columns of clinical data
4. Dropped NC class moving forward due to very low cases
5. Ran PCA & UMAP for dimensionality reduction
6. Ran K-Means & GMM clustering on UMAP
7. Selected top 50 features based on mutual information

Exploratory Data Analysis

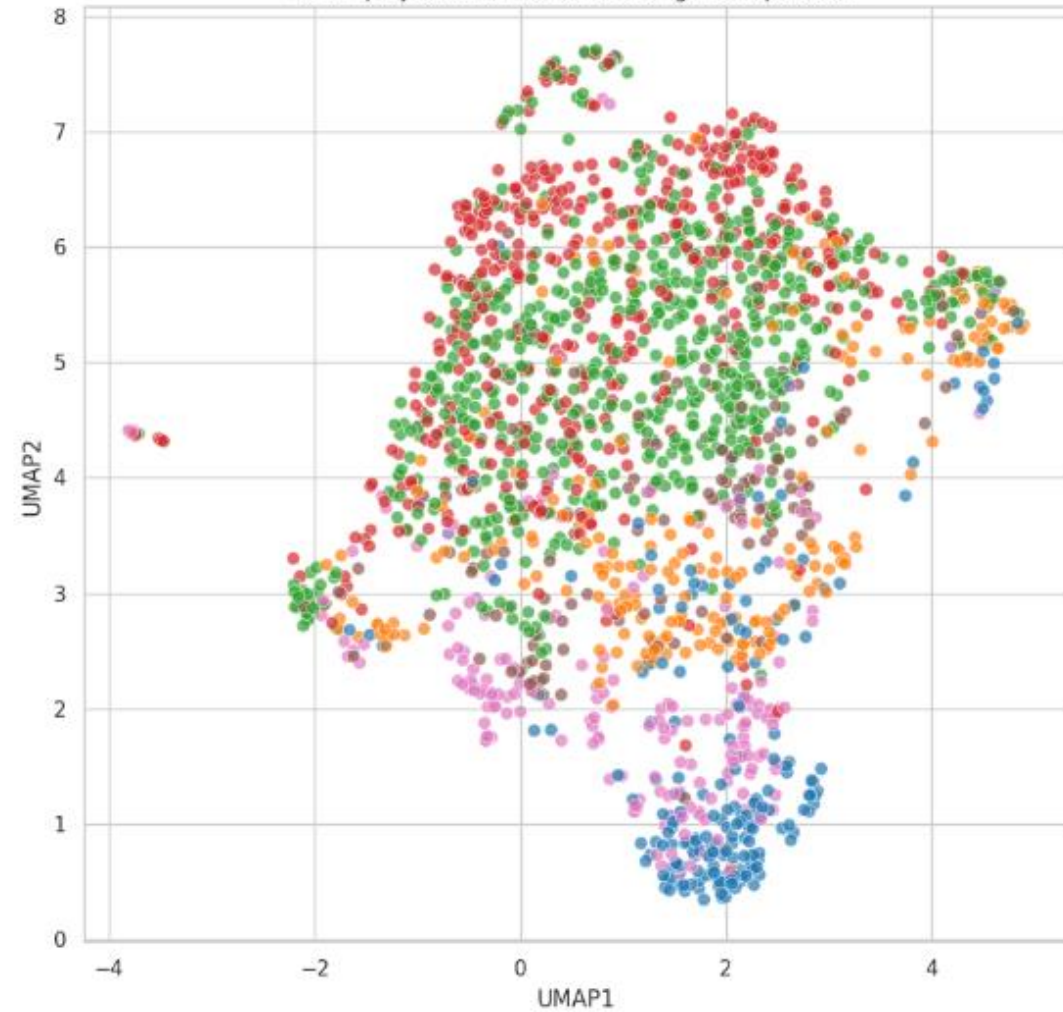


PCA & UMAP

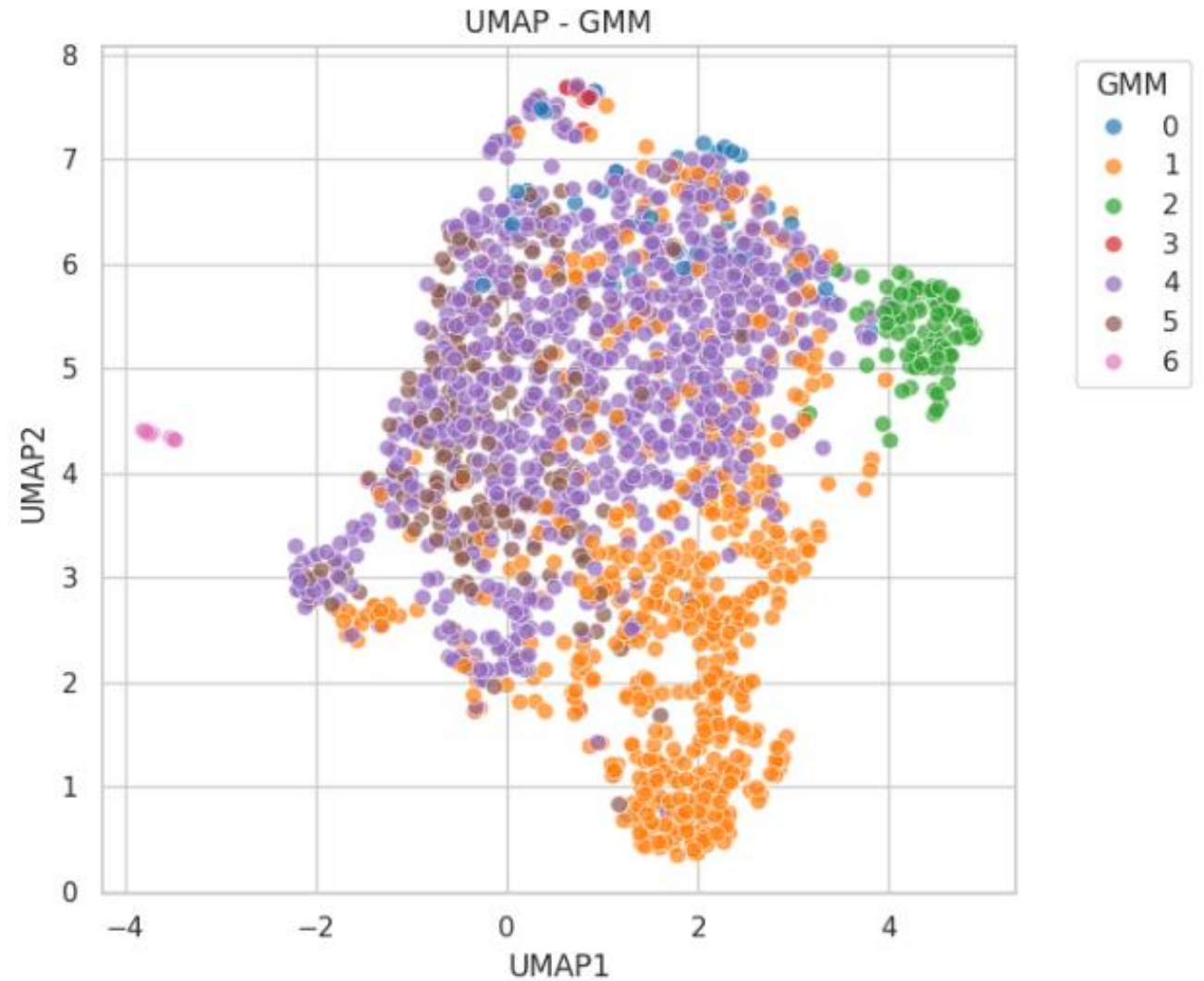
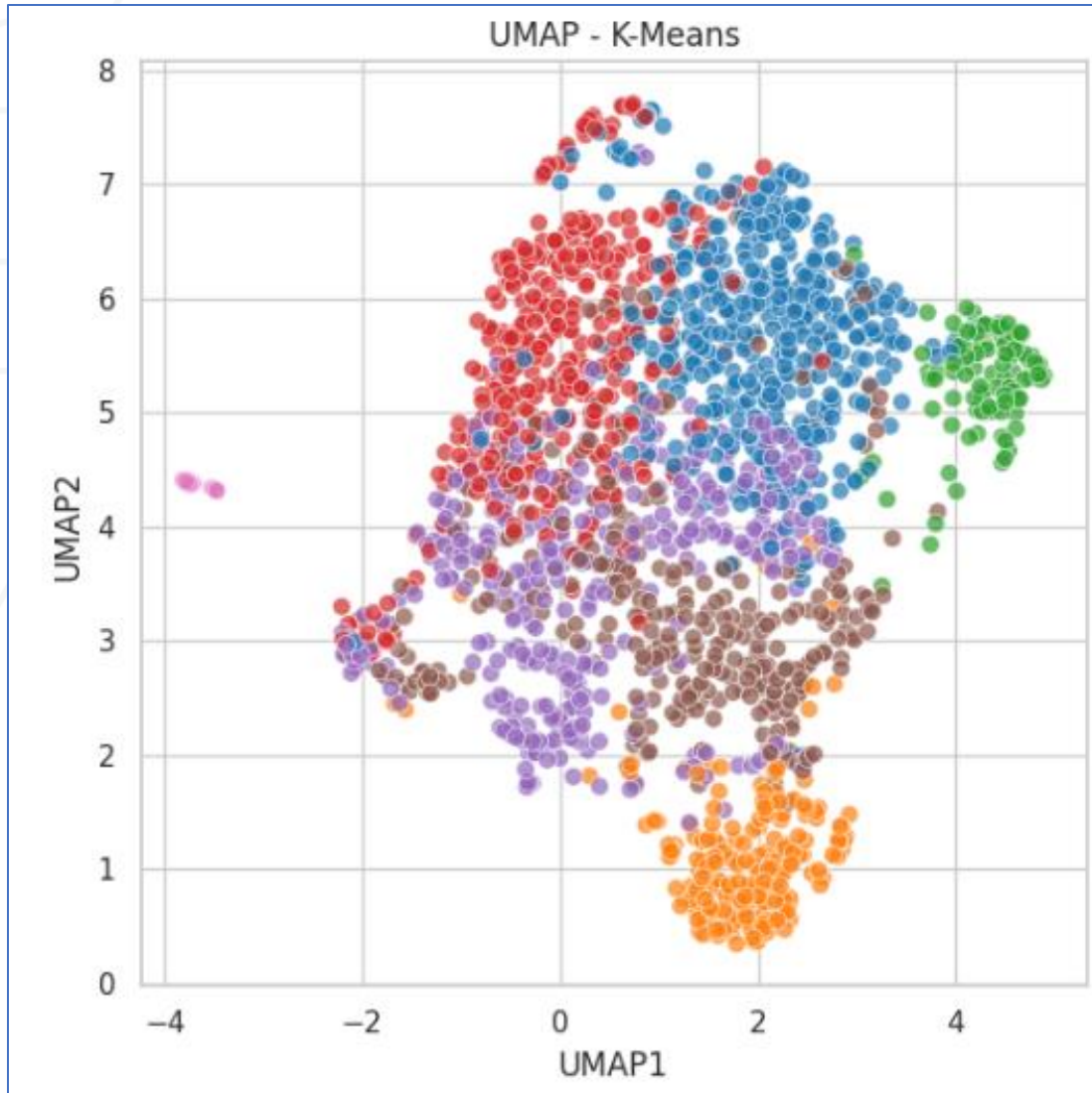
PCA projection of breast-cancer gene expression



UMAP projection of breast-cancer gene expression



Clustering: K-Means vs GMM



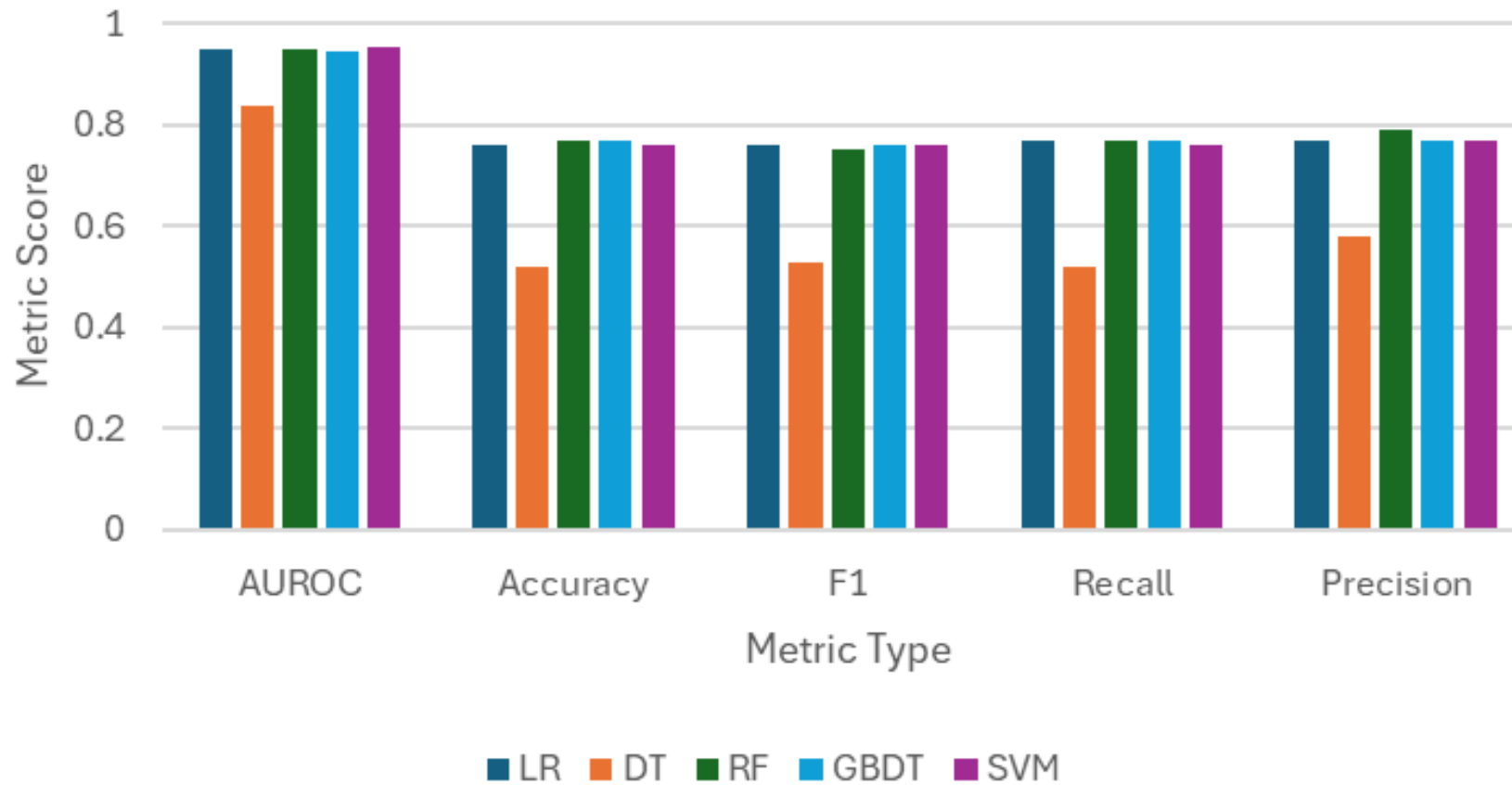
Data Preprocessing

- Train-Test Split (80% training, 20% testing)
- Clinical Data:
 - Imputation
 - Outlier treatment using IQR clipping
 - Scaling
 - Correlation Handling
 - Encoding Categorical Data
- Feature Selection
- No missing values in gene data, already in z-scale

Model Analysis

Aggregated Weighted Metrics Summary Across Models

Weighted Metrics Summary

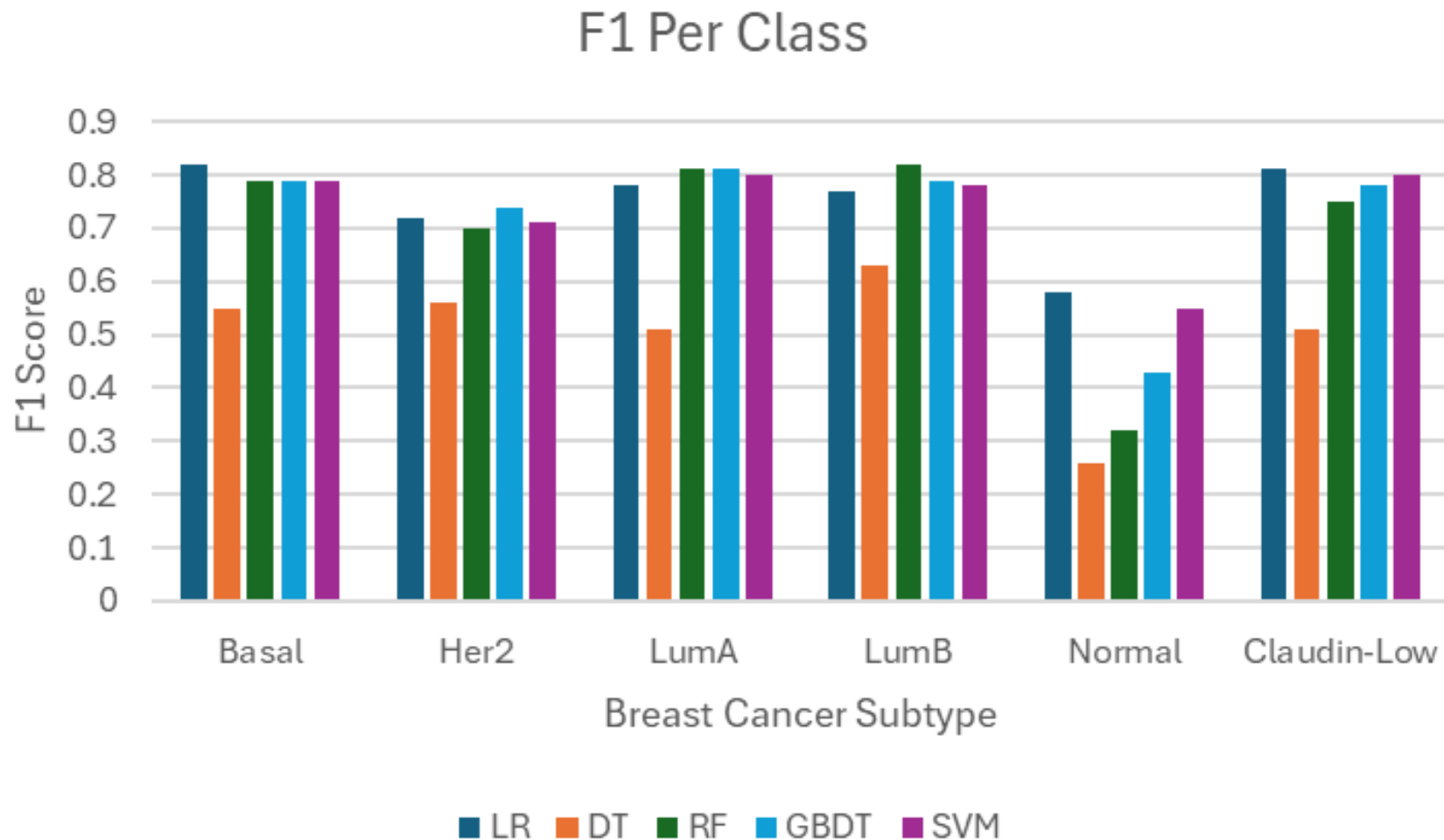


SVM: Highest AUROC
(0.9514)

LR is right behind (0.9475)

DT: Lowest metrics across
the board

F1 Per Class Across Models



LR wins on 3 classes:
Basal, Claudin-low &
Normal

RF wins LumB & tied w/
GBDT on LumA

GBDT wins Her2

SVM sits firmly in the
middle

DT lags far behind

Best Model: SVM

Hyperparameters:

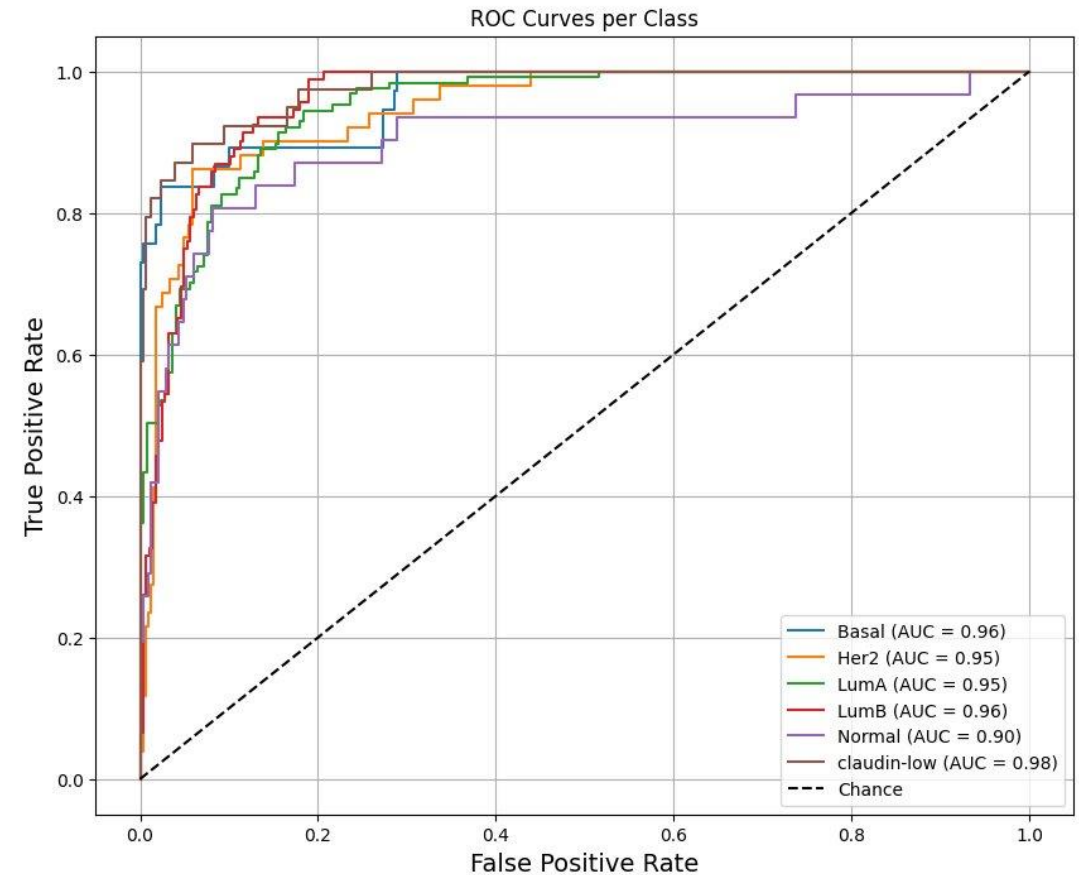
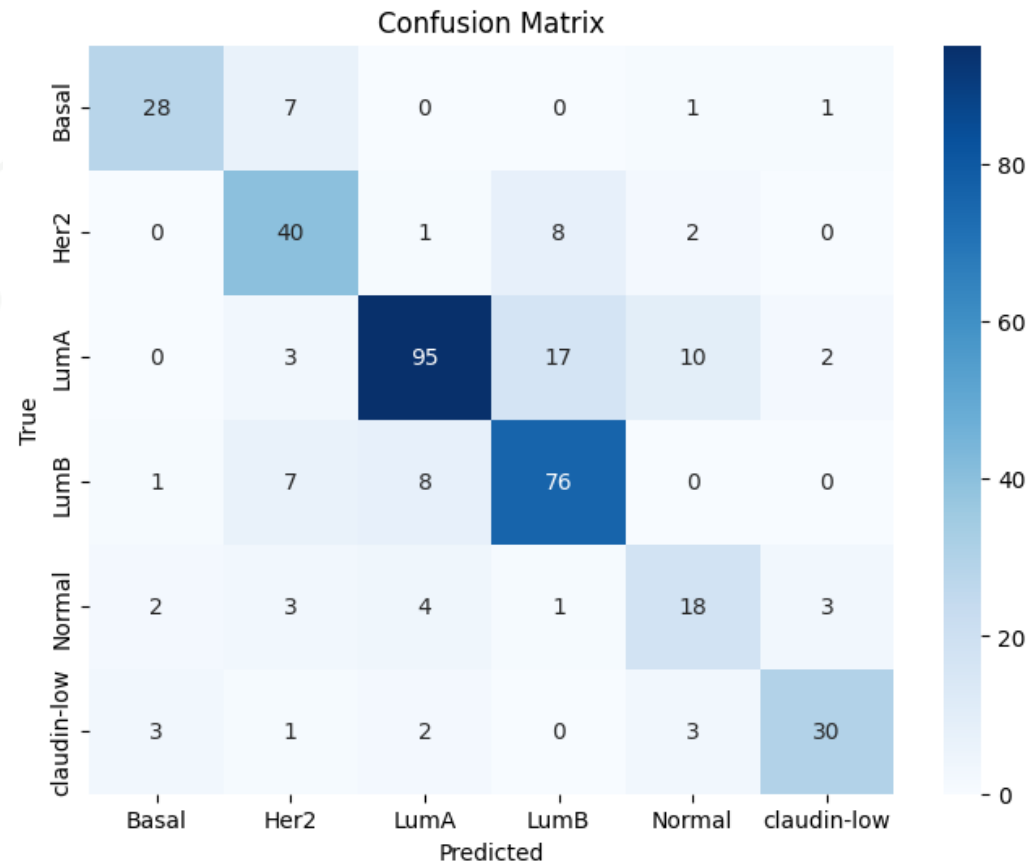
```
param_grid = {  
    'C': [0.001, 0.01, 0.1, 1, 10, 100], # Regularization parameter  
    'kernel': ['linear', 'rbf', 'poly', 'sigmoid'],  
    'gamma': ['scale', 'auto', 0.001, 0.01, 0.1, 1], # Kernel coefficient  
    'degree': [2, 3, 4, 5], # Only for 'poly' kernel  
    'coef0': [0.0, 0.1, 0.5, 1.0] # For 'poly' and 'sigmoid' kernels  
}
```

Hyperparameter	Chosen Value
Kernel	RBF
C	10
Gamma	0.001
Tolerance	0.001
Degree	2
Coef0	0.1
Probability	True

Classification Report (Weighted AUROC: 0.9475)

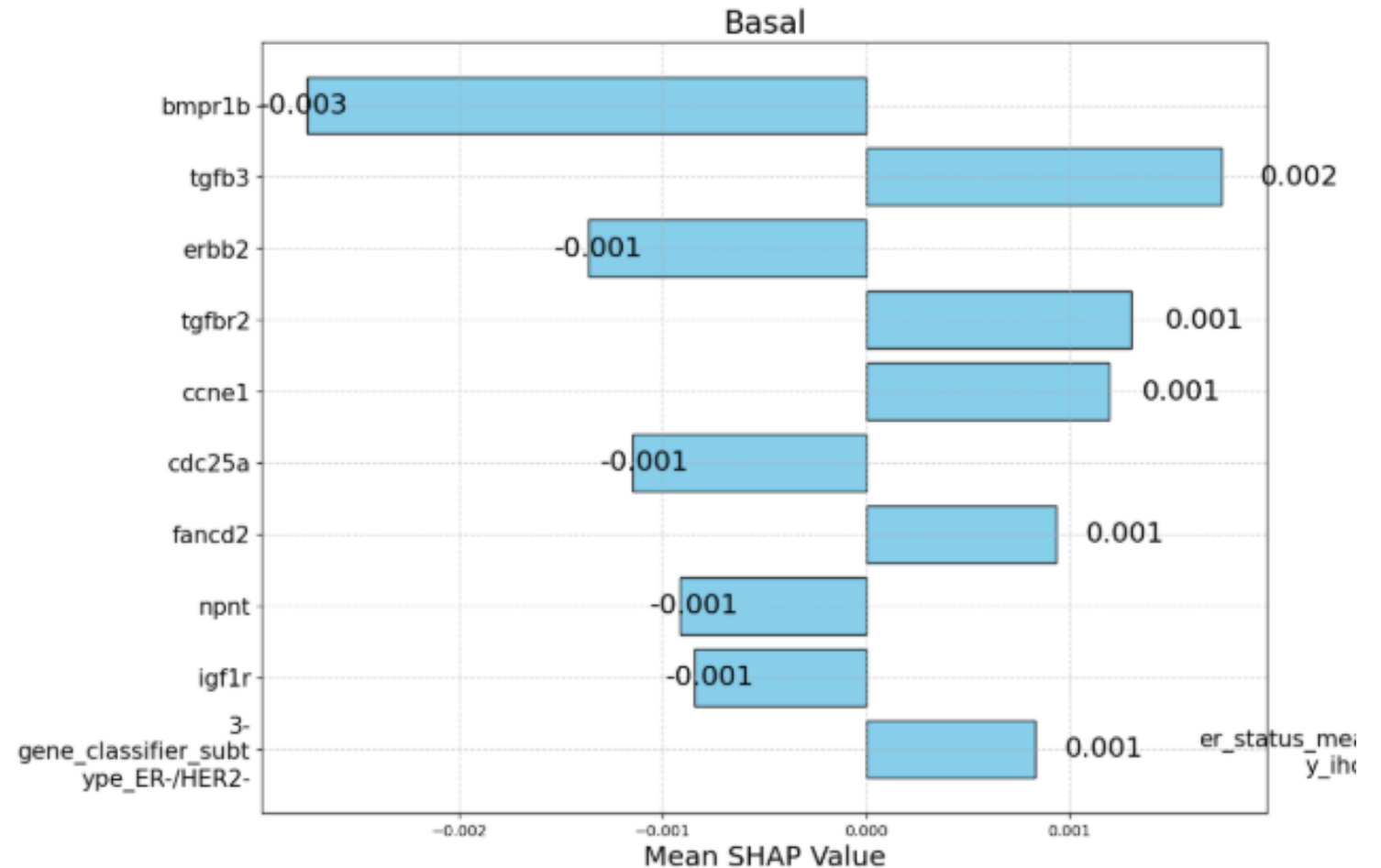
Class	Precision	Recall	F1-Score	Support
Basal	0.82	0.76	0.79	37
Her2	0.66	0.78	0.71	51
LumA	0.86	0.75	0.80	127
LumB	0.75	0.83	0.78	92
Normal	0.53	0.58	0.55	31
Claudin-low	0.83	0.77	0.80	39
Accuracy	0.76			377
Macro Avg	0.74	0.74	0.74	377
Weighted Avg	0.77	0.76	0.76	377

Best Model: SVM



SVM: Top 10 Most Informative Features

- Plotted these distributions per class
- Sticking w/ Basal class for SHAP analysis b/c it's the most aggressive subtype



SHAP Analysis: SVM

Gene	SHAP Value Sign	Coefficient's Sign Matches Clinical Literature?
bmpr1b	Negative	Yes , BMPR1B is significantly downregulated in basal-like breast cancers compared to other subtypes [4]
tgfb3	Positive	No . TGFB3 expression is documented to be lower in basal-like compared to HER-2 and LumA/B subtypes. In addition, our expression data shows the lowest expression for TGFB3 in basal-like [5].
erbb2	Negative	Yes , the sign of SHAP value and clinical literature's support that the gene is downregulated match [6] In addition, ERBB2 is the top feature for class HER2 according to the SHAP values obtained from the SVM model
tgfbr2	Positive	No , TGFBR2 is significantly downregulated across breast cancer subtypes, including basal-like ones, compared to normal tissue [7]
ccne1	Positive	Yes , CCNE1 amplification (copy-number gains) occurs in a subset (6–34%) of basal-like/TNBC cases [8]

Close Second: LR

Hyperparameters:

```
# Hyperparameter space
param_grid_l2 = {
    'C': np.logspace(-4, 2, 10),
    'penalty': ['l2'],
    'solver': ['lbfgs', 'newton-cg', 'sag'],
    'max_iter': [1000]
}
```

```
param_grid = param_grid_l2
logreg = LogisticRegression(multi_class='multinomial', random_state=42)
search = RandomizedSearchCV(
    estimator=logreg,
    param_distributions=param_grid,
    n_iter=30, # 20
    scoring=scoring,
    refit='roc_auc_ovr',
    cv=5,
    n_jobs=-1,
    random_state=42
)
```

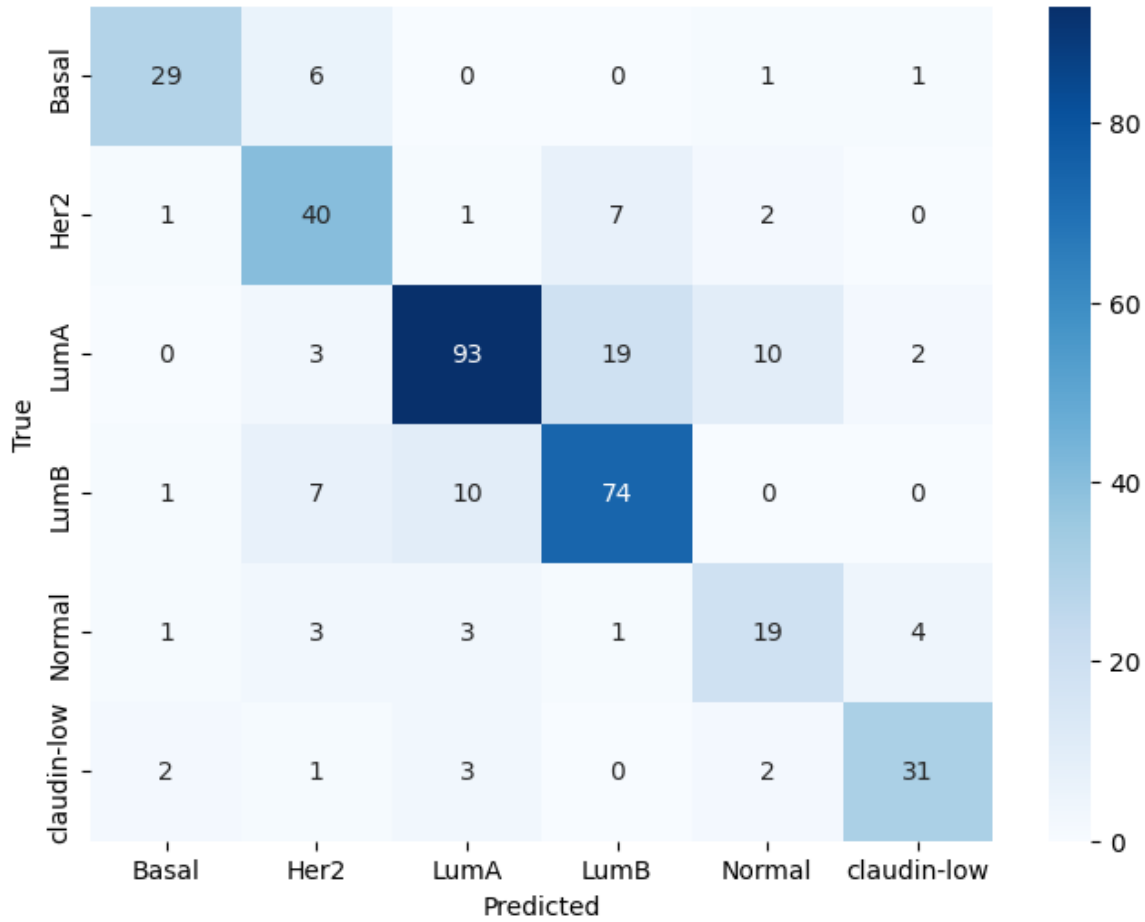
Hyperparameter	Chosen Value
Penalty	L2
Solver	Newton-cg
C	0.0464
Tolerance	0.0001
Fit Intercept	TRUE
Max Iter	1000
Multi-class	Multinomial

Classification Report (Weighted AUROC: 0.9475)

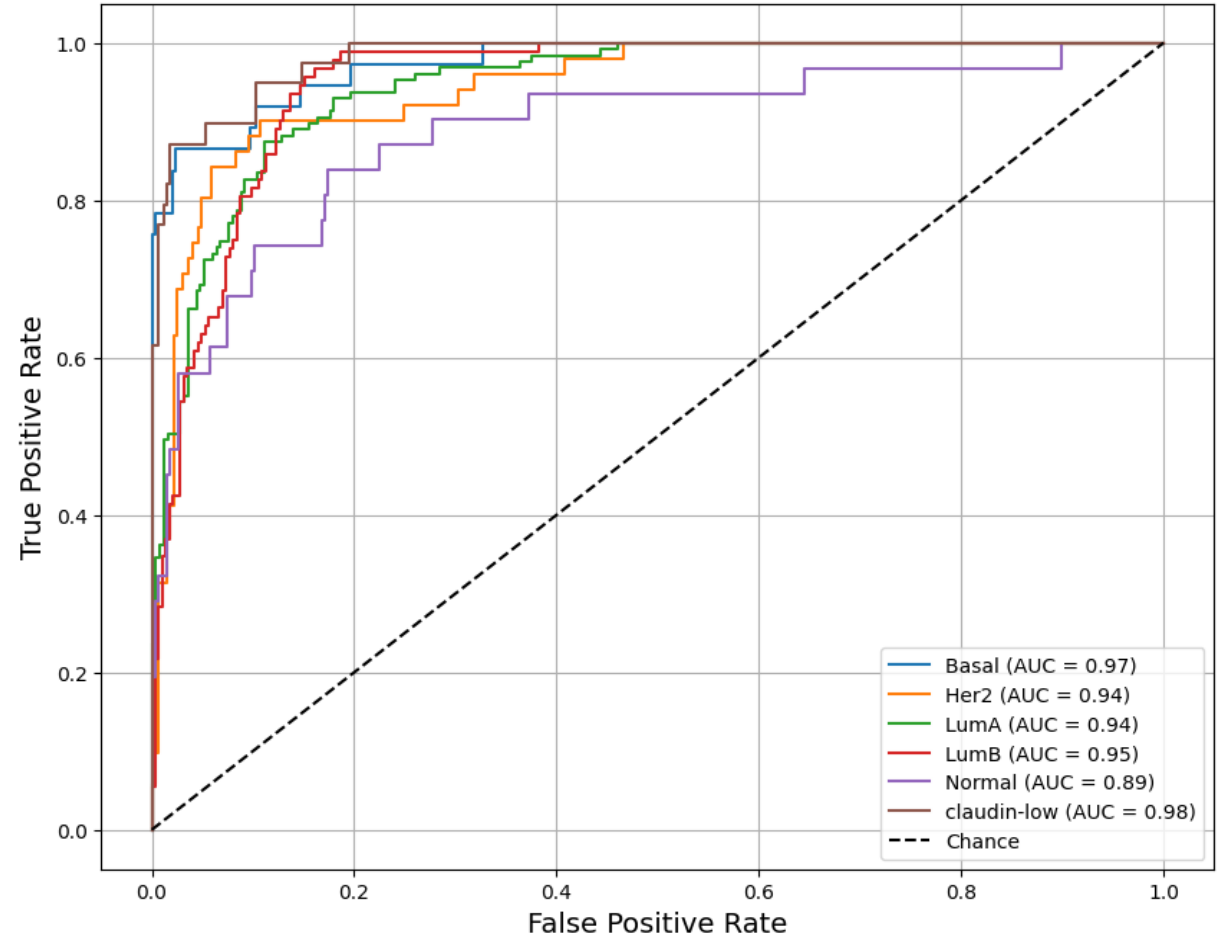
Class	Precision	Recall	F1-Score	Support
Basal	0.85	0.78	0.82	37
Her2	0.67	0.78	0.72	51
LumA	0.85	0.73	0.78	127
LumB	0.73	0.80	0.77	92
Normal	0.56	0.61	0.58	31
Claudin-low	0.82	0.79	0.81	39
Accuracy	0.76			377
Macro Avg	0.75	0.75	0.75	377
Weighted Avg	0.77	0.76	0.76	377

Close Second: LR

Confusion Matrix

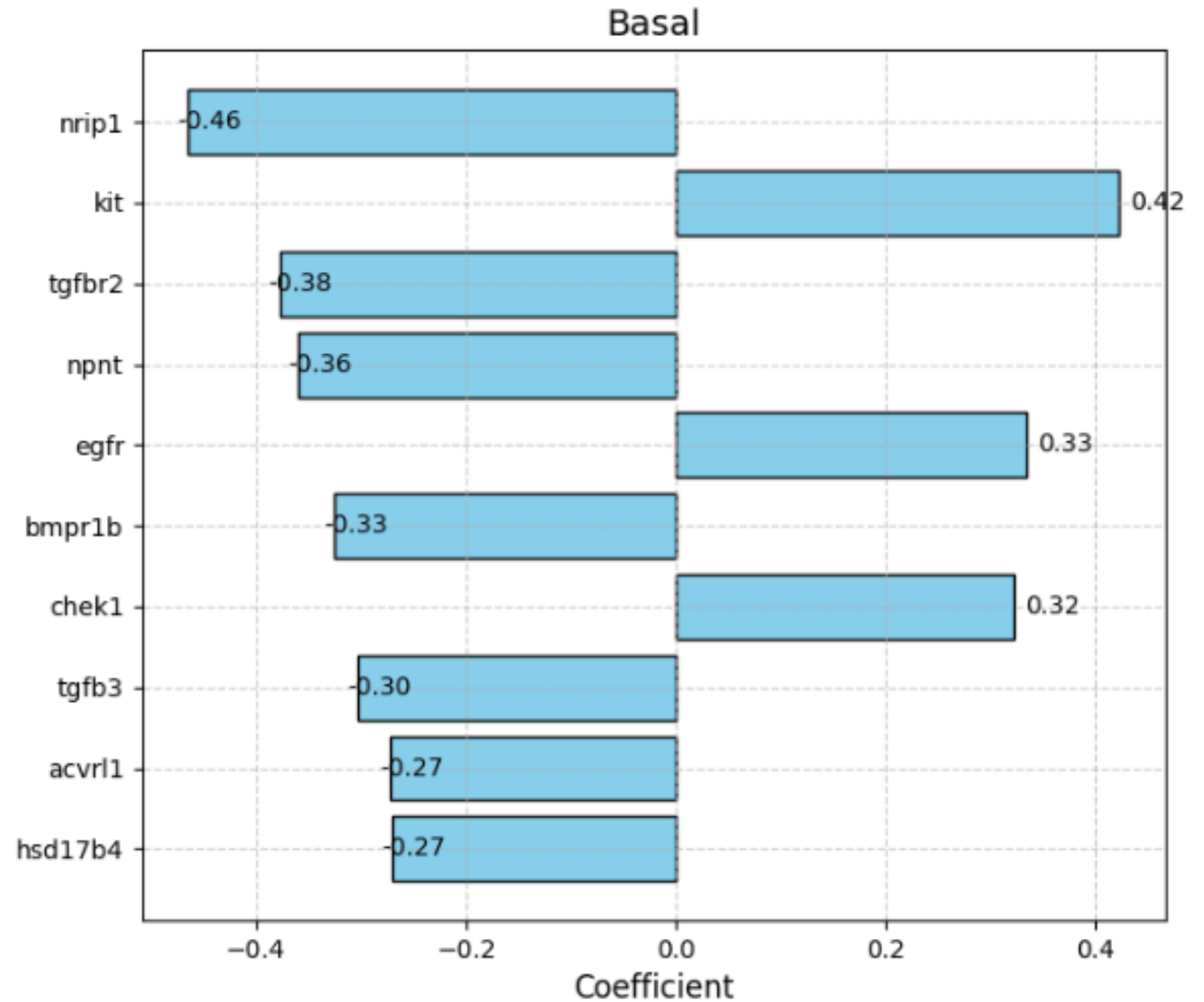


ROC Curves per Class



LR: Top 10 Most Informative Features

- Plotted these distributions per class
- Sticking w/ Basal class for SHAP analysis b/c it's the most aggressive subtype

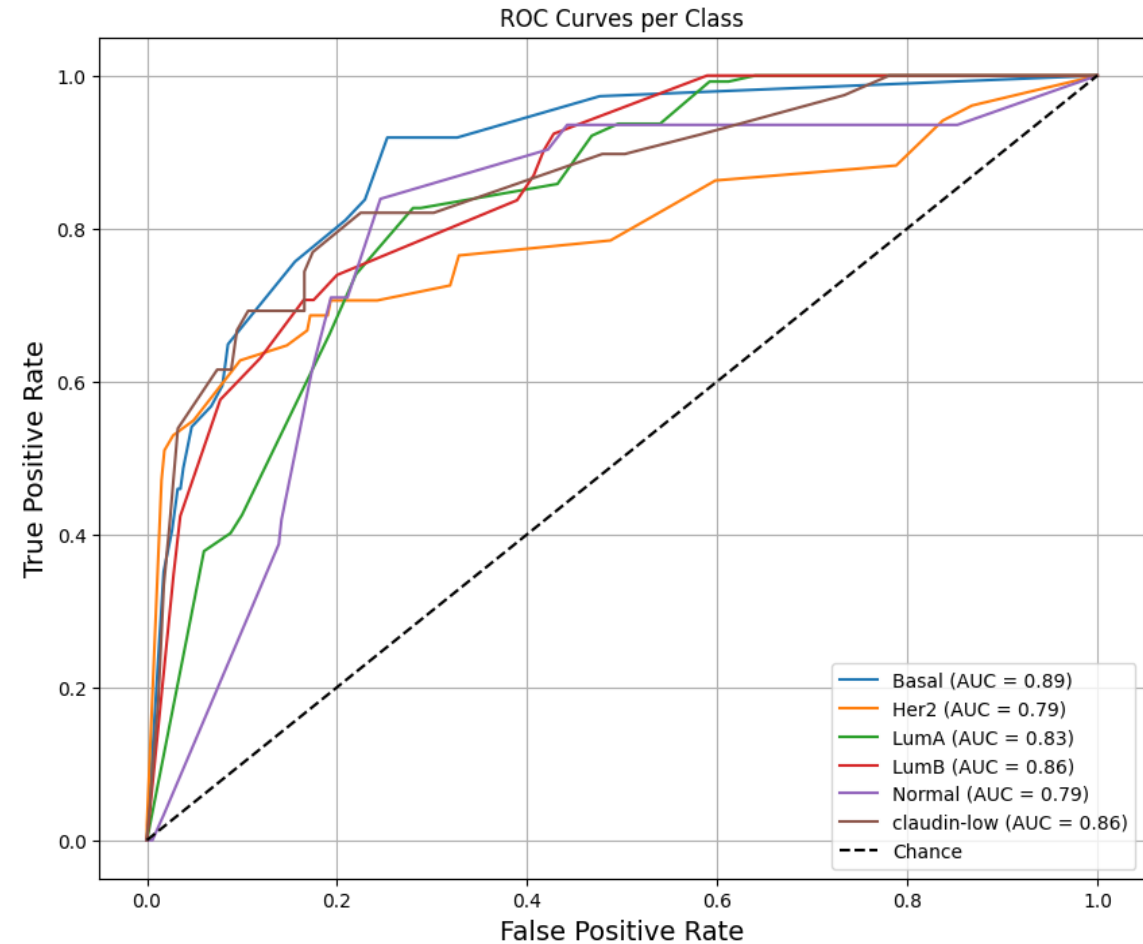
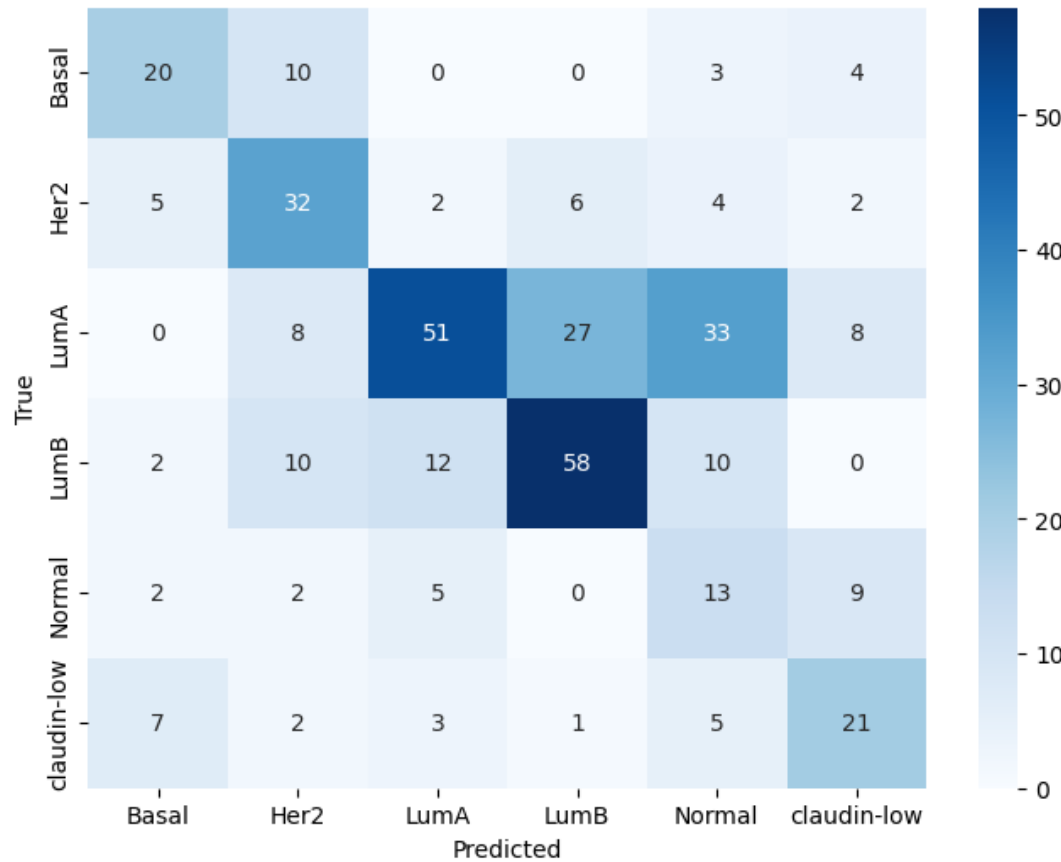


Coefficient Analysis: LR

Gene	Coefficient Value Sign	Coefficient's Sign Matches Clinical Literature?
nrip1	Negative	Yes , known to be lowly expressed in basal-like subtype [9].
kit	Positive	Yes , overexpressed in basal-like subtype [10].
tgfbr2	Negative	Yes , low expression in basal-like tissue compared to normal tissue [7].
npnt	Negative	Not known to be a marker but consistent with our expression data (box-plots)
egfr	Positive	Yes , known to be overexpressed [11].
bmpr1b	Negative	Yes . Known to be downregulated in basal-like subtype compared to other breast cancer subtypes [4].

Worst Model: DT

Confusion Matrix



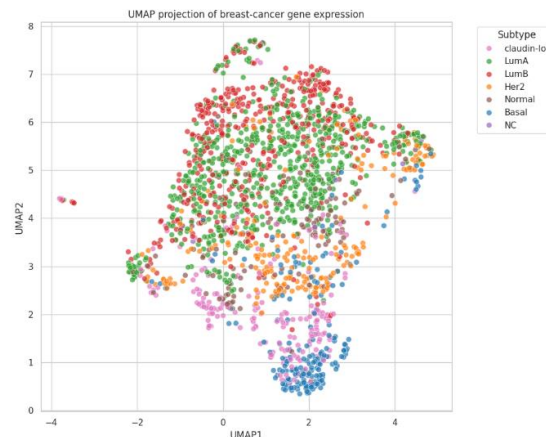
```
param_grid_dt = {
    'max_depth': [3, 5, 10, 15, 20, None],
    'min_samples_split': [2, 5, 10, 20],
    'min_samples_leaf': [1, 2, 5, 10],
    'max_features': ['sqrt', 'log2', None],
    'criterion': ['gini', 'entropy']
}
```

Worst Model: DT

- Shallow depth leading to high bias and underfitting as each split is only going to use a handful of the most obvious features
- Random Forest & GBDT also used depth ≤ 5 , but they aggregated hundreds of such trees, letting later trees (or other bootstrap samples) pick different features and interactions
- Random Forest's bootstrap sampling and feature bagging smooth out noise; GBDT turns variance into bias-correction

Conclusion

- Our data is linearly separable (UMAP supports)
- SVM is the best model, followed by LR (Based on weighted test AUROC)
- For both SVM & LR, confusion matrices show high mismatch for LumA & LumB. LumA samples were also mostly misclassified as normal in both LR and SVM
- DT is the weakest model
- Optimized the ML pipeline for speed and low-resource systems, balancing simplicity with potential performance trade-offs.



References

1. <https://www.nature.com/articles/s41591-025-03502-3>
2. <https://pmc.ncbi.nlm.nih.gov/articles/PMC3195489/>
3. [METABRIC Dataset](#)
4. <https://pmc.ncbi.nlm.nih.gov/articles/PMC10331528/>
5. <https://pmc.ncbi.nlm.nih.gov/articles/PMC4889288/>
6. <https://pmc.ncbi.nlm.nih.gov/articles/PMC6293553/>
7. <https://pmc.ncbi.nlm.nih.gov/articles/PMC2693232/>
8. <https://pubmed.ncbi.nlm.nih.gov/35384378/>
9. <https://pmc.ncbi.nlm.nih.gov/articles/PMC4741857/>
10. <https://pubmed.ncbi.nlm.nih.gov/18754326/>
11. [https://www.modernpathology.org/article/S0893-3952\(22\)03566-9/fulltext](https://www.modernpathology.org/article/S0893-3952(22)03566-9/fulltext)