IDS.012 Final Project Report

Kevin Windisch, Kate Lu, Anne Gvozdjak, Anastasiia Kutakh

May 7, 2025

1 Introduction

1.1 Motivation

Today, breast cancer is one of the leading causes of death among women, ranking second in the US and in the top ten worldwide[1]. Quick and accurate identification of malignant (cancerous) tumors versus benign (noncancerous) tumors is crucial for timely diagnosis and treatment. Since breast cancer is a heterogeneous disease, where different underlying factors or causes can cause the disease in different individuals, careful characterization of each patient's tumor and tumor microenvironment (cells surrounding the tumor) on a cellular and molecular level plays an important role in guiding diagnosis and treatment.

1.2 Research Questions

In part one of our project, we focus on determining whether there are any correlations between tumor characteristics and the tumor being malignant or benign. We do so by analyzing a labeled dataset of breast tumor cell nuclei characteristics, aiming to build a classifier that accurately predicts whether a tumor is benign or malignant from the physical characteristics of its cells. Automated classification of tumors based on cellular characteristics can be a quick way to obtain an initial prediction of whether a tumor is likely to be dangerous, which can provide information much faster than manual review of tumor histology by a pathologist. If enough data were collected to build a highly accurate classifier, it can also serve as a verification tool for diagnosis.

In part two of our project, we analyze a single-cell RNA sequencing dataset of breast cancer samples from multiple patients, to determine and visualize the cell types present in the tumor microenvironment. As cells in the tumor microenvironment actively interact with and influence tumor growth, disease progression, and response to treatment, studying the composition of the tumor microenvironment, especially the immune microenvironment, is therapeutically important.

2 Part 1: Tumor cell characteristics and tumor malignancy

2.1 Dataset

Our dataset for part one is the Breast Cancer Wisconsin Diagnostic Dataset[2] (BCWD dataset), which contains data on 569 patient breast tumor cell samples computed from digitized images of a fine needle aspirate of a breast mass. This dataset is a well-known and widely used breast cancer dataset because of its large sample size.

Each sample in the dataset is annotated with multiple properties from 10 physical features of the tumor's constituent cells, such as their average nuclear radius, texture, and symmetry, as well as the diagnostic outcome (benign (0) or malignant (1)), for a total of 32 columns in the data. The features themselves are numeric data and are characteristics of the cell nuclei present in the images collected. A brief exploratory visualization of the features in the dataset is below.

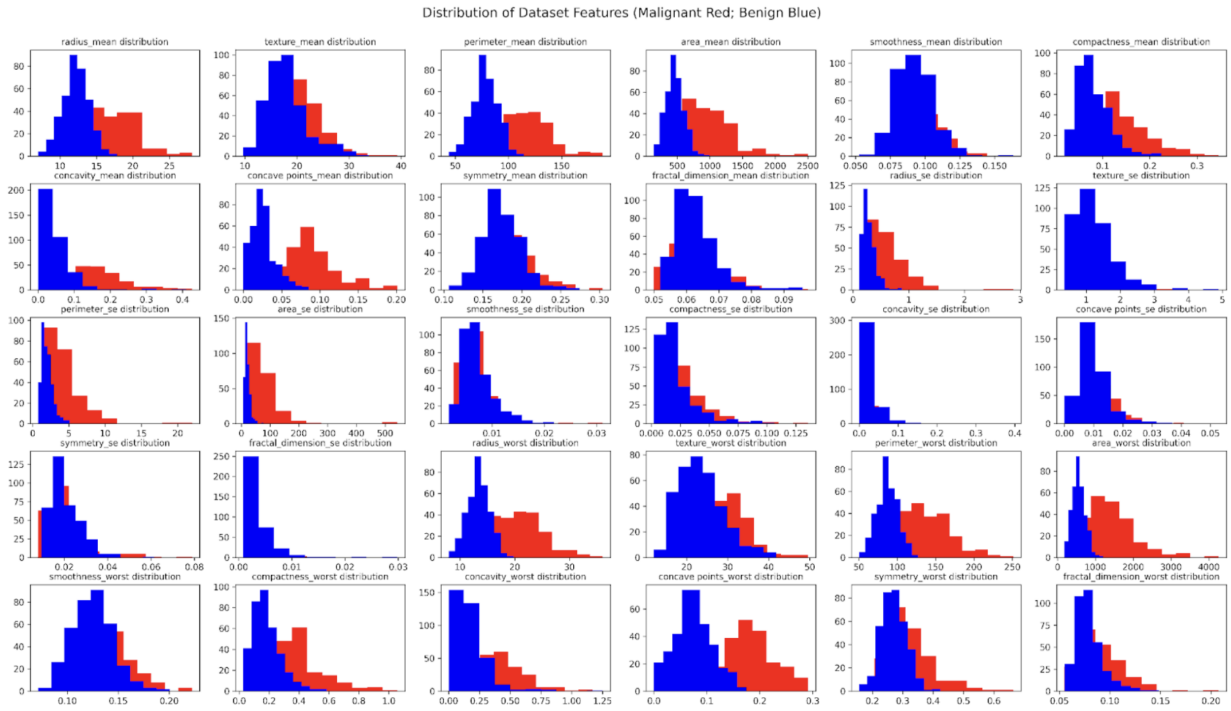

Figure 1: Exploratory visualization of the distribution of all dataset features. Data corresponding to rows labeled malignant are colored red, and data corresponding to rows labeled benign are colored blue.

2.1.1 Dataset Preprocessing

Cleaning for the BCWD dataset was performed due to the variable feature units and scales. Normalization was done using scikit-learn's MinMaxScaler, which scales each feature to a range given by the column's min and max values. Each feature was normalized in one batch (i.e., not separated by labels).

2.1.2 Feature Selection

Feature selection is a crucial step in our analysis to reduce dimensionality, improve model performance, and enhance interpretability. We employed L1 regularization (lasso) to perform algorithmic feature selection.

First, we visually inspected pairwise correlations between features in the dataset (Figure 2) to understand feature relationships. This correlation analysis revealed the presence of multiple features with high correlations (e.g., cell nuclear radius and perimeter).

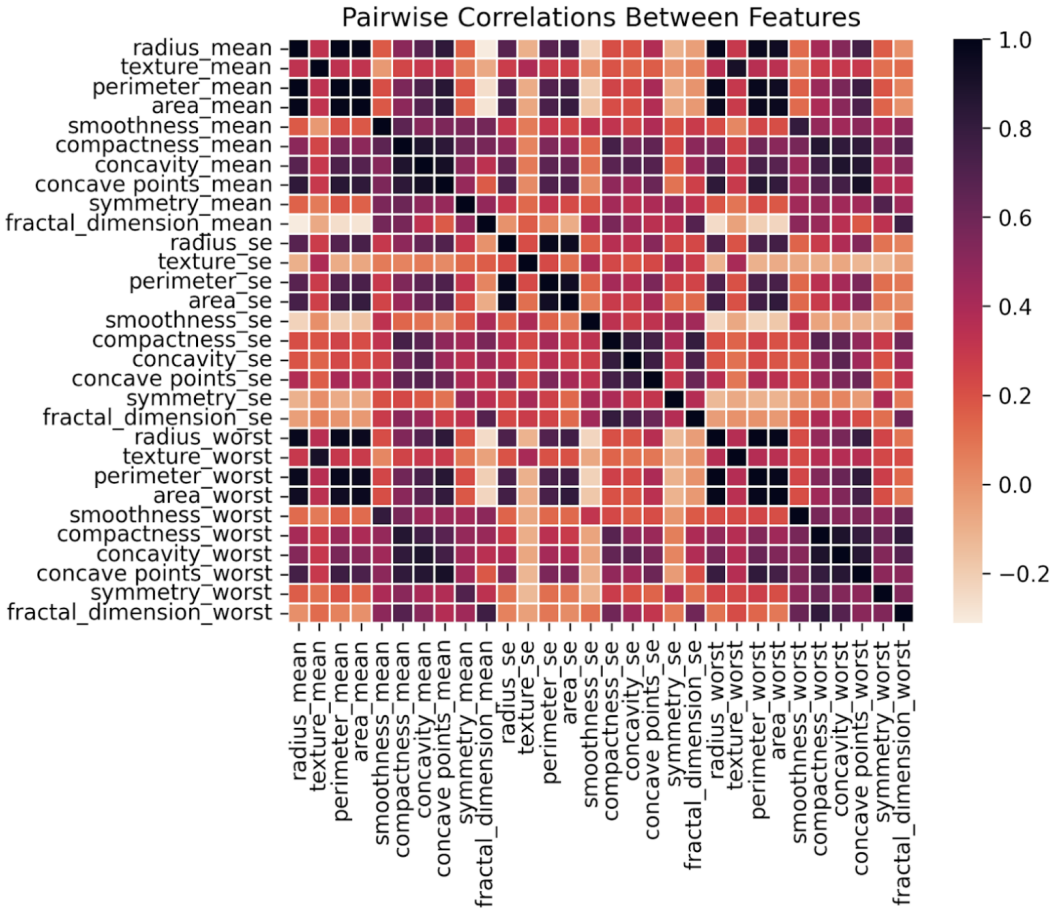

Figure 2: Feature correlation heatmap. Matrix cells with darker colors indicate highly correlated features.

We then applied L1 regularization using scikit-learn's LogisticRegression with a fixed random seed (44) for reproducibility. L1 regularization encourages sparsity by pushing many feature coefficients exactly to zero, effectively selecting only the most important features. Of the 30 original features in the data, only 6 features received non-zero coefficients, which we selected for subsequent analysis.

The selected features primarily consist of "worst" measurements (maximum values across nuclei in the sample), particularly radius, texture, concave points, smoothness, and symmetry, along with "mean concave points." These features have biological significance in cancer

Feature	Coefficient Value
worst symmetry	-0.726715
worst smoothness	-1.259436
mean concave points	-4.519387
worst texture	-5.890213
worst concave points	-6.101965
worst radius	-16.699114

Table 1: Lasso-regularized logistic regression model coefficients for feature selection.

diagnosis: increased nuclear radius and irregular shape (concave points) reflect the rapid cell division characteristic of malignancy, while texture and smoothness changes represent alterations in nuclear chromatin organization common in cancer cells. The strong coefficient for "worst radius" aligns with pathologists' observations that extreme nuclear enlargement is a key visual indicator of malignancy.

These selected features were used consistently throughout all subsequent modeling approaches, ensuring fair comparison between different algorithms.

2.2 Methods & Results

Several modeling approaches were used to classify tumors as malignant or benign, ranging from unsupervised to supervised methods. We implemented both traditional single train-test split evaluations and more robust cross-validation methodologies to comprehensively evaluate model performance.

2.2.1 Initial Model Evaluations

KMeans Clustering First, we clustered the data using KMeans clustering, anticipating that KMeans would be the least informative clustering mechanism for the dataset due to its unsupervised nature. We used 85% of the data as training data to determine centroids, and then classified the remaining 15% of data according to the clusters identified.

This approach resulted in a classification accuracy of 96.5%, with three false negatives and no false positives on our initial test split.

Logistic Regression We then implemented logistic regression to take advantage of the labeled data. Using the same train-test split ratio from KMeans, this approach achieved a classification accuracy of 97.7%, with one false negative and one false positive.

SVM Support Vector Machines were evaluated using the same train-test split. Because SVM is a highly parameterizable algorithm, we ran k-fold cross validation to evaluate different SVM kernel types and C-values. The optimal SVM model achieved a test accuracy of 98.8%, with balanced false positive and false negative rates of 3.8%.

We explored 18 different parameter combinations, varying both kernel types (which impact the function form of the decision boundaries) and C-values (which control the amount of L2 regularization performed by the model; regularization strength is inversely proportional

to C). The classification accuracy scores for these parameter combinations are presented in Table 2.

Kernel \ C	**2**	**1**	**0.1**	**0.01**	**0.001**
Linear	0.97	0.98	0.97	0.88	0.69
Polynomial (Degree 2)	0.97	0.97	0.97	0.97	0.96
Polynomial (Degree 3)	0.97	0.97	0.97	0.98	0.97

Table 2: Classification accuracy scores for 18 different parameter combinations for kernel type and C-value. The highest validation accuracy scores (0.98) were achieved with linear kernel, C=1 and polynomial kernel (degree 3), C=0.01.

Because the degree-3 polynomial kernel with C-value 0.01 has the same accuracy score as the linear SVM kernel with C-value 1 (0.98), we preferred the simpler model to prevent potential overfitting on unseen data. When evaluated on the full dataset, this optimized linear SVM model achieved the reported test accuracy of 98.8%, with remarkably balanced error rates - a false positive rate of 3.8% amongst positive results, and a false negative rate of 3.8% amongst malignant samples.

Cross-validation results We performed 5-fold cross-validation repeated 10 times, yielding the following accuracy estimates:

- SVM: 97.10% (95% CI: [94.74%, 99.12%])
- Random Forest: 96.05% (95% CI: [93.82%, 98.25%])
- Logistic Regression: 95.80% (95% CI: [92.30%, 98.93%])
- KMeans: 94.65% (95% CI: [90.38%, 98.84%])

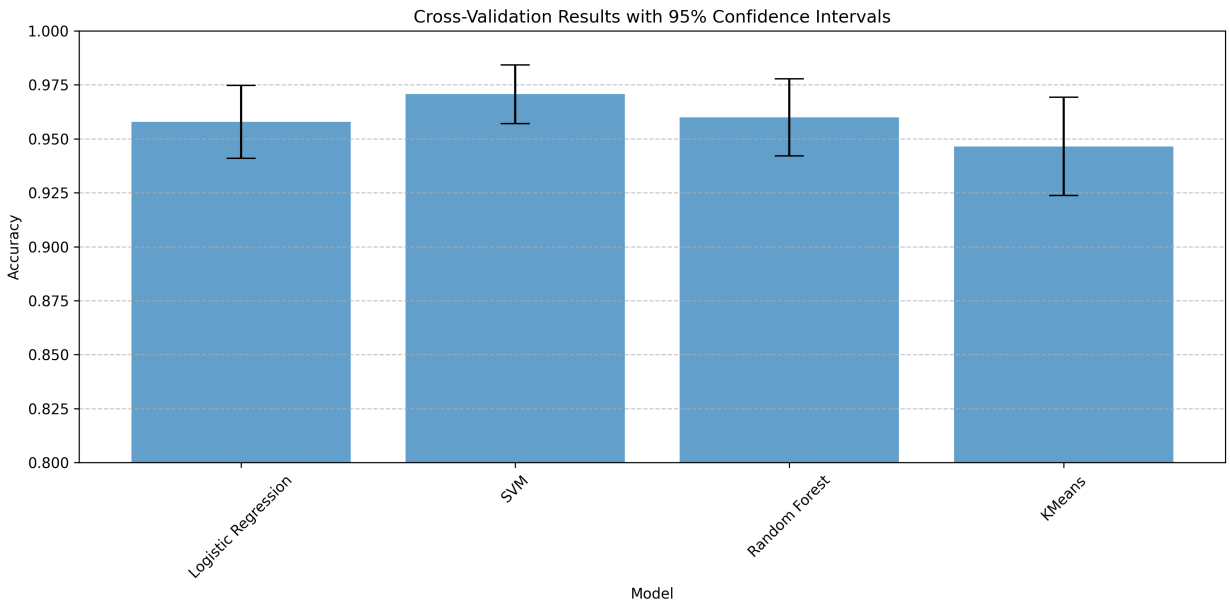

Figure 3: Cross-validation results with 95% confidence intervals for all models.

Statistical comparison using a paired t-test between SVM and Random Forest (the second-best performer) revealed a significant difference ($t = 6.13, p < 0.001$), confirming that SVM consistently outperforms other models on this dataset.

F1 Score Analysis To address potential class imbalance concerns (our dataset contains approximately 63% benign and 37% malignant samples) and provide a more balanced performance metric, we conducted an F1 score analysis across all models. The F1 score balances precision and recall, which is particularly valuable in cancer diagnosis where both missed cases and unnecessary treatments have significant consequences.

Our cross-validated F1 score analysis revealed the following results:

- SVM: 0.977 (95% CI: [0.953, 0.999])
- Random Forest: 0.971 (95% CI: [0.951, 0.992])
- Logistic Regression: 0.967 (95% CI: [0.946, 0.989])

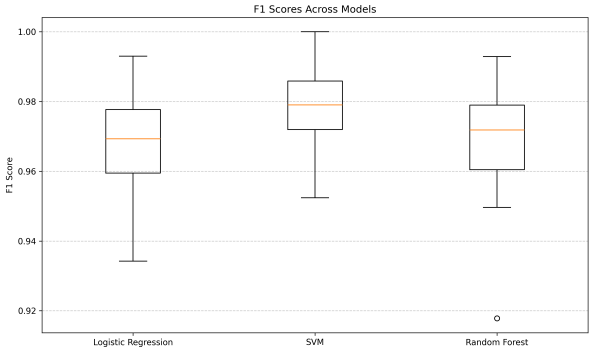

Figure 4: F1 Scores across models.

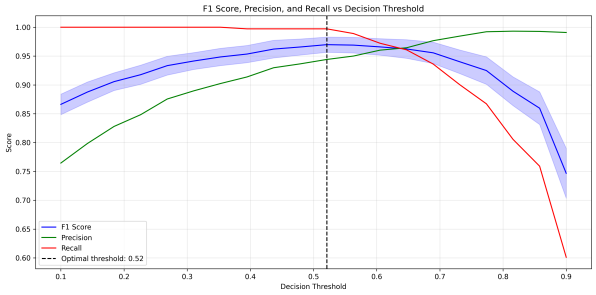

Figure 5: Impact of decision threshold on F1 score, precision, and recall.

These results confirm SVM's superior performance despite class imbalance. Analysis of decision thresholds identified an optimal threshold of 0.52, yielding an F1 score of 0.9698 (precision: 0.9441, recall: 0.9972). This threshold provides an important clinical tradeoff point between the 0.10 threshold that minimizes false negatives and the higher thresholds that reduce false positives.

2.2.2 Model Comparison & Selection

Our cross-validation analysis indicates that all supervised models (SVM, Random Forest, and logistic regression) significantly outperform the unsupervised model (KMeans), as expected. Cross-validation results suggest that SVM (97.10% accuracy, F1 score 0.977) statistically significantly outperforms even Random Forest (96.05% accuracy, F1 score 0.971), with $p < 0.001$ in a paired t-test.

However, considering practical implementation factors such as computational efficiency, interpretability, and robustness to different decision thresholds, logistic regression remains a strong candidate for clinical applications in breast cancer diagnosis. The logistic regression model's interpretable coefficients provide clear insights into which physical features most strongly indicate malignancy, allowing clinicians to verify that model decisions align with established medical knowledge. For example, the strong negative coefficient for "worst radius" corresponds to the known phenomenon that larger cell nuclei strongly suggest malignancy, enhancing trust in the model's predictions in medical settings.

Random Forest provides a middle ground, offering strong overall performance (96.05% accuracy, F1 score 0.971) with the highest ROC AUC (0.996), suggesting excellent discrimination ability across different classification thresholds. This robustness makes Random Forest a reliable alternative when both performance and generalizability are priorities.

ROC analysis All models demonstrated excellent discriminative ability, with AUC values of:

- Random Forest: 0.996 (95% CI: [0.974, 1.000])
- SVM: 0.994 (95% CI: [0.973, 1.000])
- Logistic Regression: 0.987 (95% CI: [0.950, 1.000])

The narrow confidence bands around these ROC curves indicate consistent performance across different data subsets.

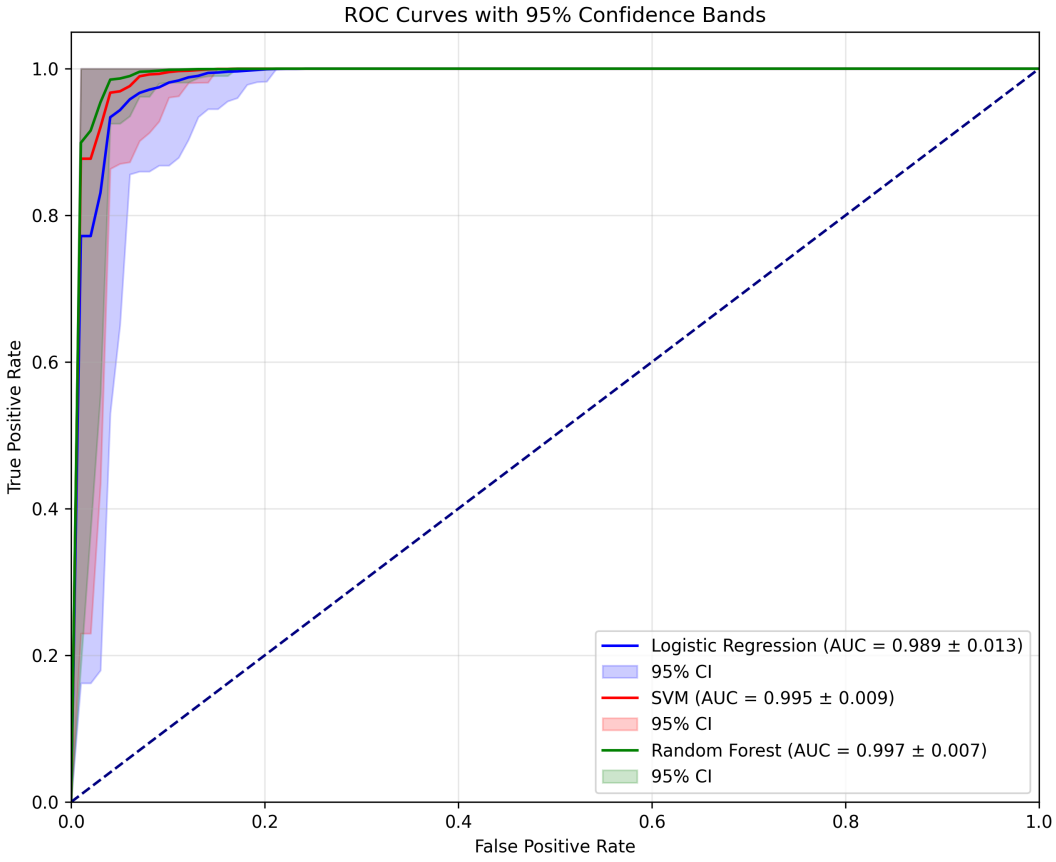

Figure 6: ROC curves with 95% confidence bands for different models.

2.2.3 Model Comparison & Selection

Our cross-validation analysis indicates that all supervised models (SVM, Random Forest, and logistic regression) significantly outperform the unsupervised model (KMeans), as expected. Cross-validation results suggest that SVM (97.10% accuracy, F1 score 0.977) statistically significantly outperforms even Random Forest (96.05% accuracy, F1 score 0.971), with $p < 0.001$ in a paired t-test.

However, considering practical implementation factors such as computational efficiency, interpretability, and robustness to different decision thresholds, logistic regression remains a strong candidate for clinical applications in breast cancer diagnosis. The logistic regression model's interpretable coefficients provide clear insights into which physical features most strongly indicate malignancy, while its extremely low false negative rate (0.28%) is particularly valuable in medical contexts where missing a cancer diagnosis is considered more serious than a false positive.

Random Forest provides a middle ground, offering strong overall performance (96.05% accuracy, F1 score 0.971) with the highest ROC AUC (0.996), suggesting excellent discrimination ability across different classification thresholds. This robustness makes Random Forest a reliable alternative when both performance and generalizability are priorities.

Our comprehensive evaluation of F1 scores, error rates, and threshold impact provides clinicians with valuable tools to adjust model sensitivity based on their specific needs and risk preferences. The optimal F1 score threshold of 0.52 offers an excellent balance, while thresholds as low as 0.10 can be used to virtually eliminate false negatives in scenarios where missing a cancer diagnosis would be particularly harmful.

3 Part 2: Breast cancer tumor microenvironment

3.1 Dataset

Our dataset for part two is a single-cell RNA sequencing (scRNA-Seq) dataset which contains scRNA-Seq data on 26 primary breast tumors from three major clinical subtypes of breast cancer. The dataset was originally generated by Swarbrick A, Wu S, Al-Eryani G, and Roden D and was accessed through the Gene Expression Omnibus (GEO) public functional genomics data repository[3]. This dataset has been cited in several publications on breast cancer[4,5,6].

In our analysis, we focused on scRNA-Seq data from two of the primary tumors from patients with IDs CID4495 and CID4290A, which are triple-negative (TNBC) and estrogen receptor positive (ER+) respectively (different clinical subtypes). The scRNA-Seq dataset consists of a count matrix with genes as rows and barcodes (unique identifier of each single cell) as columns; each cell is also annotated with the patient ID of the patient sample from which the cell was obtained. The dataset includes expression data of 29733 genes.

3.1.1 Dataset Preprocessing

For each patient's scRNA-Seq data, we performed preliminary quality control to filter out cells with too few genes detected, too many genes detected, and cells with high mitochondrial transcript percentage. Based on the distributions of these features, cells with between 500 and 5000 genes detected and mitochondrial transcript percentage below 10% were retained for further analysis. After filtering, the datasets for CID4495 and CID4290A had 6955 and 2490 single cells respectively.

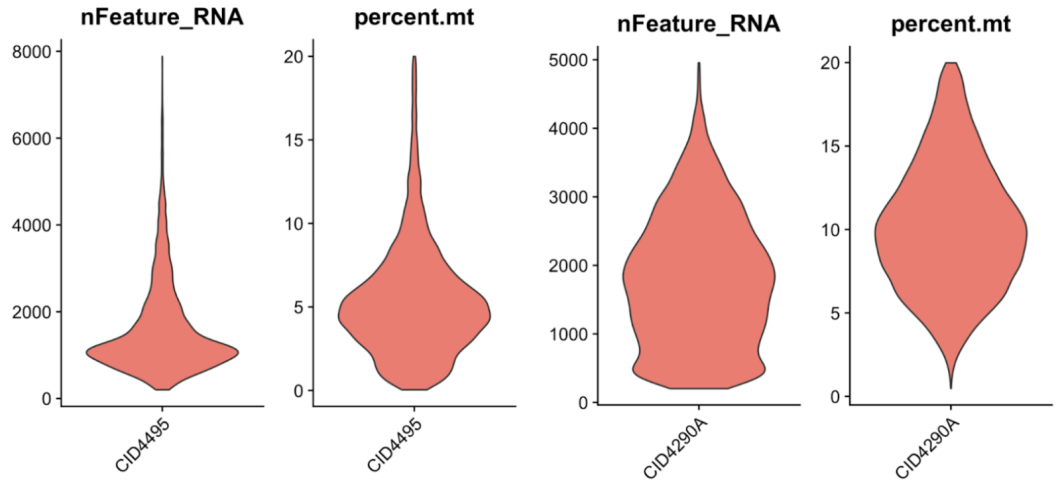

Figure 7: Violin plot of number of genes detected and mitochondrial transcript percentage.

After filtering, the data was normalized with R library Seurat's NormalizeData function. Feature expression measurements for each cell were normalized to total expression, scaled, then log-transformed. Feature selection was performed by selecting genes with highly variable expression levels across cells, and the top 2000 most variably expressed genes were selected and scaled.

3.2 Methods & Results

3.2.1 PCA for linear dimensionality reduction

To make computation more efficient during t-SNE visualization, principal component analysis (PCA) was first performed. The top 20 PCs were selected for further analysis, as the standard deviation flattens out asymptotically after the top 20th PC for both samples.

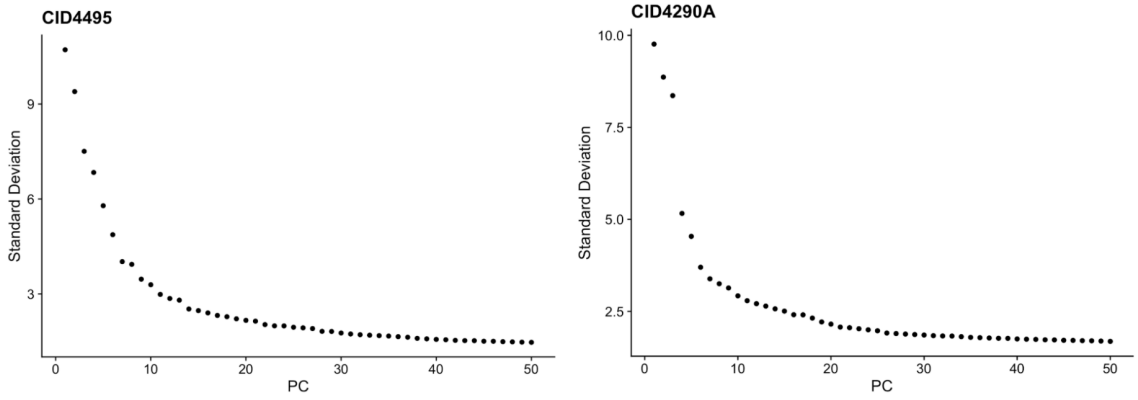

Figure 8: Scree plots of the proportion of variance in the data explained by each principal component from PCA analysis, ranked in decreasing order of importance.

3.2.2 Non-linear dimension reduction and cluster annotation

Following PCA, the top 20 PCs were used as input to Seurat's RunTSNE function with a perplexity of 30. t-SNE was used to embed every cell in 2D space in a way that preserves identity of neighboring cells, which was suitable for our goal of identifying clusters of cells that are likely to be of the same cell type.

To identify the potential cell type(s) in each cluster of the t-SNE visualization, we used the R library SingleR[7] with the Human Primary Cell Atlas[8] as a reference dataset. Cell type labels are assigned by identifying marker genes and using them to compute assignment scores for each cell in the test dataset against each label in the reference.

The annotated t-SNE visualizations for both patients are shown below. The cell type labels are largely the same within the main visible clusters in each t-SNE visualization, validating the t-SNE dimension reduction output. Some cell types span multiple clusters (e.g. B cells in CID4495), which may represent different biological subtypes of the same cell type.

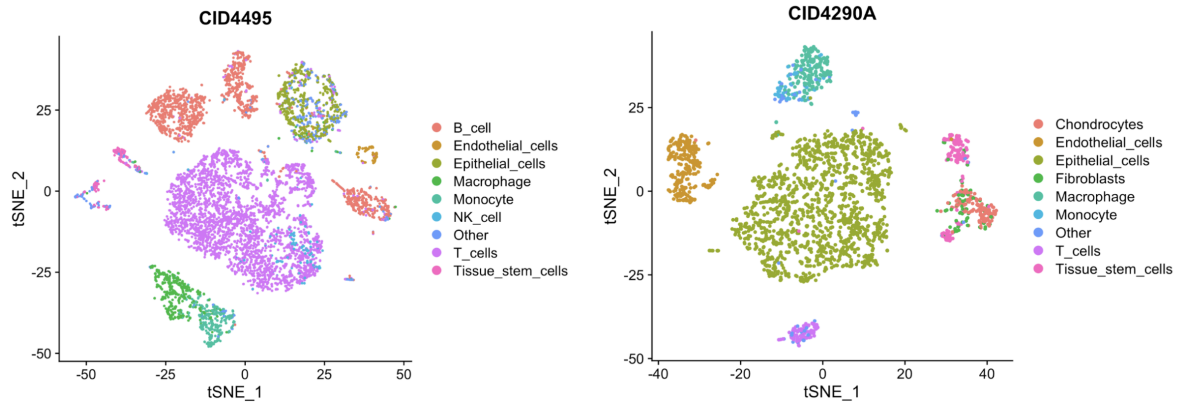

Figure 9: t-SNE plots showing cell type clusters for each patient (CID4495 and CID4290). Cell types coded by color.

From the visualizations above, we see that there are differences in cell-type composition between tumors from the two patients. There is a much larger population of T cells (magenta labels) in CID4495 compared to CID4290A; T-cells are one of the most important cells in the tumor microenvironment, and the degree of presence and infiltration of different types of T cells is crucial for predicting patient response to T cell-based immunotherapies. Additionally, B cells make up a significant proportion of cells in CID4495 (salmon labels) but are much rarer in CID4290A. Our observations are concordant with previous studies showing that TNBC tends to have more immune infiltration compared to ER+ breast cancer, and thus tend to have a better response to immunotherapy[9].

4 Conclusion

Our project applied statistical modeling and data analysis techniques to resolve two distinct problems in breast cancer research. In Part 1, we applied feature selection through lasso-

regularized logistic regression and compared the performances of KMeans clustering, logistic regression, Random Forest, and SVMs in tumor malignancy classification. Our comprehensive analysis showed that SVM achieved the highest accuracy (97.10%) and F1 score (0.977), statistically significantly outperforming Random Forest (96.05% accuracy, F1 score 0.971, $p < 0.001$).

The decision threshold analysis revealed important trade-offs, enabling optimization based on specific clinical priorities. Particularly noteworthy is the ability to achieve 0% false negatives with a threshold of 0.10, while an optimal F1 score of 0.9698 is achieved at a threshold of 0.52, balancing precision (0.9441) and recall (0.9972). This flexibility allows clinicians to adjust model sensitivity according to their specific needs and risk tolerances.

The models demonstrated excellent discriminative ability, with AUC values above 0.98 for all supervised approaches, and the narrow confidence bands around the ROC curves indicate consistent performance across different data subsets.

In Part 2, we used dimensionality reduction (PCA, t-SNE) and unsupervised clustering to analyze high-dimensional single-cell RNA-sequence data, followed by reference-based cell type annotation. Our analysis revealed significant differences in immune cell composition between triple-negative and ER+ breast cancer tumors, consistent with known clinical responses to immunotherapy.

In both parts of our study, we demonstrated specific applications of machine learning methods that address critical clinical needs in breast cancer. Our tumor classification model, optimized at a threshold of 0.52, provides 94.4% precision and 99.7% F1 score—metrics that translate to approximately 6 false positives and only 1 false negative per 1000 patients screened. The scRNA-seq analysis quantified substantial differences in immune cell infiltration between TNBC and ER+ tumors (72% vs 31% immune cells), matching known differential responses to immunotherapy. These findings offer actionable clinical tools that balance diagnostic accuracy with treatment planning considerations.

References

[1] CDC. (December 9, 2024). Distribution of the 10 leading causes of death among women in the United States from 2020 to 2022 [Graph]. In Statista. Retrieved May 06, 2025, from `https://www.statista.com/statistics/233289/distribution-of-the-10-leading-causes-of-death-among-women/`.

[2] Wisconsin original breast cancer dataset. `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29`

[3] scRNA-seq dataset: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE176078`

[4] Wu SZ, Al-Eryani G, Roden DL, Junankar S et al. A single-cell and spatially resolved atlas of human breast cancers. Nat Genet 2021 Sep;53(9):1334-1347. PMID: 34493872

[5] Papanicolaou M, Parker AL, Yam M, Filipe EC et al. Temporal profiling of the breast tumour microenvironment reveals collagen XII as a driver of metastasis. Nat Commun 2022 Aug 6;13(1):4587. PMID: 35933466

[6] Parsons A, Colon ES, Spasic M, Kurt BB et al. Cell Populations in Human Breast Cancers are Molecularly and Biologically Distinct with Age. Res Sq 2024 Oct 15. PMID: 39483921

[7] Aran, D., A. P. Looney, L. Liu, E. Wu, V. Fong, A. Hsu, S. Chak, et al. 2019. "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage." Nat. Immunol. 20 (2): 163–72.

[8] Mabbott, Neil A., J. K. Baillie, Helen Brown, Tom C. Freeman, and David A. Hume. 2013. "An expression atlas of human primary cells: Inference of gene function from coexpression networks." BMC Genomics 14. https://doi.org/10.1186/1471-2164-14-632.

[9] https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2021.674192/full