BT3041 AIBD Project Report
Group 6

Tumor vs Non-Tumor Classification Using scRNA-seq Data

1. Introduction

Cancer is a highly complex disease, largely due to the diverse and unpredictable nature of its cells. Traditional bulk RNA sequencing methods, which measure average gene expression across many cells, often miss critical variations that exist within tumors. These hidden differences-known as cellular heterogeneity, are a major reason why cancer treatments can fail. Single-cell RNA sequencing (scRNA-seq) offers a powerful alternative by allowing researchers to examine gene expression at the level of individual cells. This approach uncovers subtle differences and rare cell types that bulk methods cannot detect. Although scRNA-seq presents challenges such as technical noise and data sparsity, its ability to provide detailed insights into tumor composition makes it especially valuable. By combining scRNA-seq data with machine learning techniques, we can more accurately classify cancer cells and identify key features for diagnosis and therapy. For these reasons, scRNA-seq was chosen as the focus of our study.

In this project, we focused on leveraging single-cell RNA sequencing (scRNA-seq) to identify genes that play a significant role in distinguishing between cancerous and non-cancerous states. Our aim was to pinpoint which genes are most influential in determining the presence or absence of cancer, as well as to explore which specific cell types provide the most valuable information for cancer diagnosis. By analyzing gene expression at the single-cell level, we sought to uncover subtle differences that might be overlooked by traditional methods. As we move forward in this report, we will detail the steps involved in extracting and processing scRNA-seq data, followed by a comprehensive analysis that enables accurate classification of cancer cells.

2. Data Acquisition and Preprocessing

We explored the literature and chose a paper that suited the dataset and our main objective.
From the paper "*PanClassif: Improving Pan Cancer Classification of Single Cell RNA-seq Gene Expression Data Using Machine Learning,*" we focused on the classification of tumor versus non-tumor in breast cancer.

The dataset used in our project was obtained from the **Gene Expression Omnibus (GEO)**, it is a public online database that provides access to many biological datasets. The specific dataset is identified with the **GEO accession number GSE75688**, and it contains information related to **breast cancer** at the **single-cell RNA expression levels**. This type of data shows how active each gene is in individual cells.

In this dataset, each sample represents a single cell, and for each cell, the expression level of **57,915 genes** was recorded. There are **515 total cell samples** in the dataset.

Each sample is labeled as **Tumor** or **Non-Tumor**, meaning we can treat this as a **binary classification problem**, where our goal is to train a model to predict whether a cell is cancerous or not.

Here is how the samples are distributed:

- **198 samples** are labeled as **Non-Tumor**
- **317 samples** are labeled as **Tumor**

This shows that there are more tumor samples than non-tumor samples, and this creates an **imbalance in the class distribution**. This imbalance can cause some problems in training machine learning models, as the model might learn to favor the majority class (tumor) and perform poorly on the non-tumor class.

To deal with this class imbalance problem, a **class balancing strategy** was used. An **equal number of samples** from both classes was **randomly selected**, so that the number of tumor and non-tumor samples would be the same, with 50 tumor samples and 50 non-tumor samples. This helps to ensure that our model learns patterns from both classes equally and reduces the risk of bias.

After balancing the classes, the data was prepared for training and testing the model. A **train-test split** was performed:

- **80%** of the **balanced data** was used for **training**
- **20%** was used for **testing**

This split allows the model to learn from most of the data (80%), while still having enough data left to test how well it performs on the new, unseen samples.

3. Principal Component Analysis (PCA)

Purpose of PCA in the Project

In our project, Principal Component Analysis (PCA) was employed as a dimensionality reduction technique to address the high dimensionality of single-cell RNA sequencing (scRNA-seq) data. The original dataset consisted of 57,915 gene expression values per cell, which makes direct analysis computationally intensive and may obscure meaningful biological patterns due to noise and redundancy.

PCA helps by transforming the original high-dimensional data into a lower-dimensional space, where each dimension (principal component) captures as much variance in the data as possible. This allows for more efficient processing and visualization, and it improves the performance of downstream tasks like clustering and classification.

Implementation Steps of PCA

1. Data Standardization: Before applying PCA, the gene expression data were standardized. Standardization ensures that each gene has zero mean and unit variance, so that genes with inherently larger expression values don't dominate the analysis.
2. Application of PCA: PCA was applied to the entire dataset to extract the principal components—linear combinations of the original features (genes). The top 2 principal components were selected, which captured the majority of the variance in the data. This dimensionality reduction step made the data more manageable and allowed focus on the most informative patterns.
3. Use in Feature Selection: The top 500 genes with the highest absolute loading values (i.e., those that contributed most to the top principal components) were identified.

Visualization of PCA Results

2D PCA plot

Figure 1: 2D PCA projection of tumor and non-tumor samples.

- The x-axis represents PC1, and the y-axis represents PC2.
- Points are color-coded based on labels: Tumor (red) vs. Non-Tumor (blue).
- The plot shows a clear separation between the two classes, indicating that PCA successfully captured biologically meaningful variance that distinguishes tumor from non-tumor cells. This separation supports the effectiveness of PCA in compressing high-dimensional gene data into low-dimensional components that still retain discriminatory power.

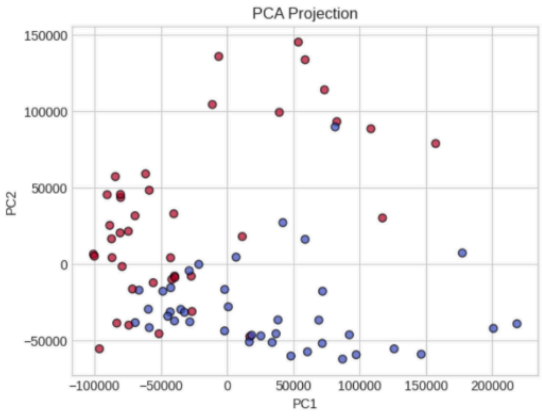

Figure 1

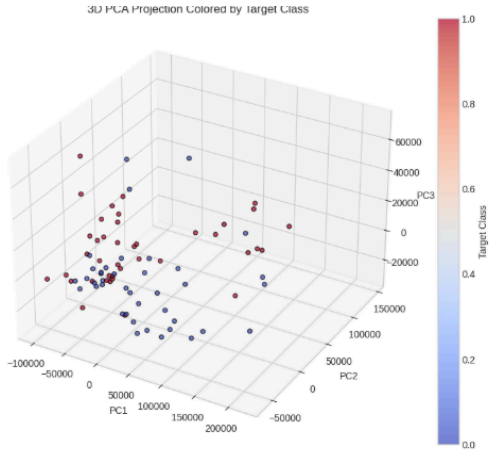

Figure 2

3D PCA Plot

Figure 2: 3D PCA projection of tumor and non-tumor samples.

- The three axes correspond to PC1, PC2, and PC3.
- This 3D visualization further enhances class separation and confirms that tumor and non-tumor cells exhibit distinct expression patterns.
- It visually demonstrates how PCA uncovers latent structures within the dataset

4. Reducing Dimensionality Further Using t-SNE

t-SNE (t-distributed Stochastic Neighbor Embedding) is a machine learning algorithm used for dimensionality reduction and data visualization. It helps to visualize high-dimensional data in a lower-dimensional space (typically 2D or 3D) while preserving the underlying structure and patterns. Now, we will be using this to view the data clusters even more clearly.
Firstly, we will perform t-SNE without PCA preprocessing, then perform the same with PCA preprocessing, and compare the results obtained from the two flows.

t-SNE Without PCA Preprocessing

To capture nonlinear relationships in the high-dimensional gene expression space, the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm was applied directly to the standardized data. Parameters included:

- n_components = 2: to enable 2D visualization,
- perplexity = 30: balancing local and global aspects of the data,
- random_state =42: to ensure reproducibility.

The resulting 2D embedding was visualized using Seaborn scatter plots with color-coded labels based on the tumor/non-tumor classification.

t-SNE With PCA-Based Feature Selection

To improve the interpretability and reduce computational complexity, a two-stage dimensionality reduction was performed:

1. **Principal Component Analysis (PCA)** was applied to project the data into a lower-dimensional subspace (50 components), capturing the majority of the variance.
2. The top two principal components were used to identify genes with the highest absolute loadings (top 500 features) — presumed to be the most informative for discrimination.
3. These selected features formed a reduced dataset, which was subjected again to t-SNE, using the same parameters as the earlier embedding.

This combined PCA–t–SNE approach aimed to focus the embedding on the most discriminative genes and further highlight separability in the tumor and non-tumor profiles.

Output:
Two key t-SNE visualizations were produced:

- One from raw standardized data,
- Another from PCA-informed feature selection.

The latter was found to provide a clearer cluster structure, suggesting the advantage of incorporating prior variance-based filtering. The most informative gene features were exported to an Excel file (`Important_Features.xlsx`) for downstream biological interpretation.

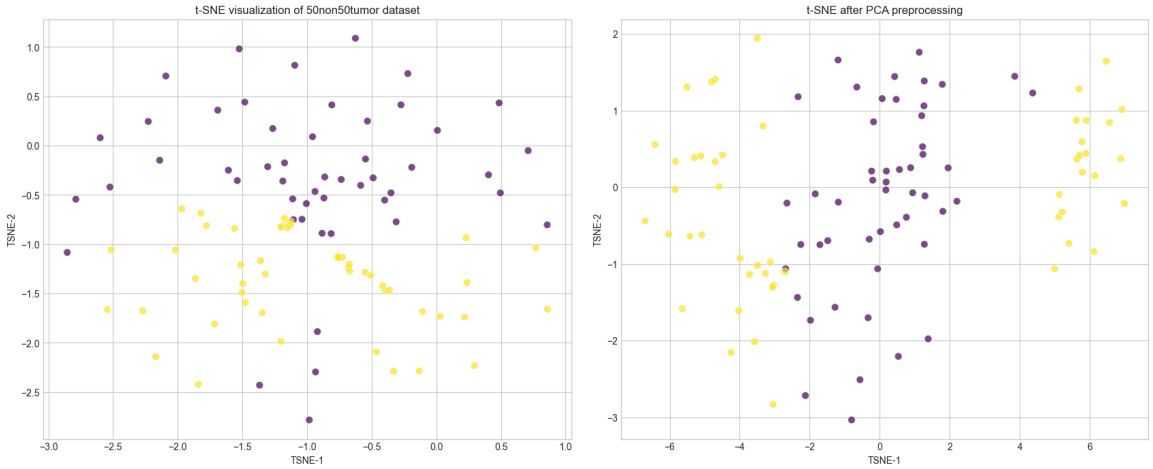

Fig: Left - t-SNE visualization of the tumor cells dataset without PCA preprocessing, right - the same process with PCA preprocessing. Notice the major difference in the distribution of the clusters between the two visualizations.

5. Model Preparation:

Objective

To develop and evaluate machine learning pipelines for accurately classifying single cells as cancerous (tumorous) or non-cancerous using single-cell RNA sequencing (scRNA-seq) data from the Gene Expression Omnibus (GEO). We implemented and compared two different pipelines focused on dimensionality reduction and gene feature selection strategies.

Pipeline Overview

Pipeline 1: PCA-Based Projection Approach

- **Steps:**
  - Data Acquisition from GEO
  - Data Preprocessing (normalization, filtering, etc.)
  - Principal Component Analysis (PCA)
  - Use of the first two principal components (PC1 and PC2)
  - Random Forest Classifier

- **Test Set Performance:**
  - **Accuracy:** 85%

Classification Report			
Class	**Precision**	**Recall**	**F1-Score**
Non-Tumor (0)	0.82	0.90	0.86
Tumor (1)	0.89	0.80	0.84

Pipeline 2: Gene-Level Feature Selection Approach

- **Steps:**
  - Data Acquisition from GEO
  - Data Preprocessing
  - PCA for Loadings
  - Select Top 500 Genes Contributing to Variance
  - Train a Random Forest Classifier on these genes
  - Extract the Top 100 Most Important Genes from the Classifier
  - Retrain the Classifier using only the expression values of these 100 genes

- **5-Fold Cross-Validation Performance:**
  - **Mean Accuracy:** 97%
  - **Standard Deviation:** 0.04

Classification Report			
Class	**Precision**	**Recall**	**F1-Score**
Non-Tumor (0)	0.98	0.96	0.97
Tumor (1)	0.96	0.98	0.97

Biological Validation

A literature review of the top 100 genes used in Pipeline 2 revealed that **at least 20 genes** are already documented in the scientific literature as being **overexpressed in breast cancer cells**. Notable examples include:

- **ARHGDIB**
- **EPCAM**
- **CYB561**
- **MAL2**

- **TFAP2A**
- **TFAP2C**

This alignment with biological findings provides further confidence in the relevance and interpretability of the model's outputs.

Conclusion

- **Pipeline 2 significantly outperformed Pipeline 1**, achieving a mean accuracy of 97% and strong class-wise precision, recall, and F1-scores.
- The use of gene-level features not only improved performance but also allowed biological interpretability by identifying genes associated with cancer phenotypes.
- This result suggests that gene expression-based feature selection combined with machine learning is a robust approach for single-cell classification in cancer research.

6. Cell-Type-Specific Classification for Tumor Prediction

Objective

The primary objective of this experiment was to assess whether certain subtypes of non-tumor cells exhibit distinctive gene expression profiles compared to tumor cells, thereby acting as informative features for classification. Instead of analyzing all non-tumor cells as a single lumped group, we focused on specific immune and stromal cell types to identify which ones contribute most to distinguishing between cancerous and non-cancerous states.

Dataset and Methodology

The dataset includes RNA-seq gene expression data from tumor and non-tumor cells. While tumor cells lack subtype annotations, non-tumor cells are classified into the following subtypes:

- B cells
- T cells
- Myeloid cells
- Stromal cells

For each of the four subtypes above, we created separate binary classification datasets pairing the subtype-specific non-tumor cells with the tumor cells. This resulted in four distinct dataframes:

1. B cell RNA-seq data vs. Tumor data
2. T cell RNA-seq data vs. Tumor data
3. Myeloid cell RNA-seq data vs. Tumor data
4. Stromal cell RNA-seq data vs. Tumor data

Each data frame was preprocessed and subjected to **Pipeline 1**:

- Principal Component Analysis (PCA)
- Feature extraction using PC1 and PC2
- Random Forest Classification

Each dataset was split into training and testing sets with a test size of 0.2.

Results:

T cell vs tumor:

```
Accuracy: 0.9375
Confusion Matrix:
 [[7 1]
 [0 8]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.88      0.93         8
           1       0.89      1.00      0.94         8

    accuracy                           0.94        16
   macro avg       0.94      0.94      0.94        16
weighted avg       0.94      0.94      0.94        16
```

Good performance, suggesting informative differences

B cell vs tumor:

```
Accuracy: 0.875
Confusion Matrix:
 [[8 0]
 [2 6]]
Classification Report:
              precision    recall  f1-score   support

           0       0.80      1.00      0.89         8
           1       1.00      0.75      0.86         8

    accuracy                           0.88        16
   macro avg       0.90      0.88      0.87        16
weighted avg       0.90      0.88      0.87        16
```

Reasonably high accuracy

Myeloid cells vs tumor

```
Accuracy: 1.0
Confusion Matrix:
 [[5 0]
 [0 5]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         5
           1       1.00      1.00      1.00         5

    accuracy                           1.00        10
   macro avg       1.00      1.00      1.00        10
weighted avg       1.00      1.00      1.00        10
```

High accuracy may be due to small test set size; possible overfitting

Stromal cells vs tumor:

```
Accuracy: 1.0
Confusion Matrix:
 [[8 0]
 [0 8]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         8
           1       1.00      1.00      1.00         8

    accuracy                           1.00        16
   macro avg       1.00      1.00      1.00        16
weighted avg       1.00      1.00      1.00        16
```

The dataset had only 28 data points per class; the test set had ≈ 5 per class

Interpretation

The results indicate that each non-tumor cell subtype, when compared individually to tumor cells, contains transcriptomic signals that allow for effective classification. T cells and B cells demonstrated robust accuracy, indicating meaningful differences from tumor cells.

However, the perfect classification observed for myeloid and stromal cells may not reflect actual performance, as these datasets had relatively few test samples. For example, the stromal cell dataset had only ~28 samples per class, leading to test sets of only ~5 samples/class. This small sample size likely inflated performance metrics and suggests a need for caution in interpretation.

Conclusion

This analysis demonstrates that considering specific non-tumor cell subtypes individually can provide a more nuanced understanding of how tumor cells differ at the transcriptomic level. High classification accuracy, especially for T and B cells, suggests that immune cell subtypes play a significant role in distinguishing tumor from non-tumor profiles. Based on our analysis, we are not able to find which subtype of cell has more distinguishing capability. Further validation on larger and more balanced datasets is necessary to confirm these findings, particularly for stromal and myeloid cells.

References:

https://doi.org/10.1016/j.ygeno.2022.01.001

https://doi.org/10.1200/jco.2015.33.7_suppl.474

https://doi.org/10.3892/ol.2024.14857

Colab notebook links:

Notebook_1

Notebook_2

Notebook_3

Group Members:

Roshan (BS210B19)
Mohith (BS21B020)
Sishir (BS21B034)
Suraj Kiran (BS21B036)
Naga Keerthana (BS21B038)
Omkar (BE22B014)