Statistical Recognition Methods for Cell Type Identification in Single-Cell RNA-seq

Metody statistického rozpoznávání pro identifikaci buněčných typů v single-cell RNA-seq

UNIVERSITY OF CHEMISTRY AND TECHNOLOGY PRAGUE

University of Chemistry and Technology, Prague

Faculty of Chemistry and Technology

Bioinformatics and Chemistry informatics

Bc. Jiří Vlasák

Prague, 2025

Contents

Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized modern biology by enabling the transcriptomic profiling of thousands of individual cells simultaneously . Unlike traditional bulk sequencing, which averages gene expression across a tissue, scRNA-seq reveals the heterogeneity of complex biological systems at cellular resolution. However, this granularity comes at a cost. These datasets are characterized by high dimensionality, often exceeding 20,000 features (genes), and extreme sparsity, where a high percentage of values are zero due to "dropout" events. Consequently, the automated identification of cell types from such complex data represents a fundamental problem in statistical pattern recognition that requires robust computational strategies [1-2].

The primary objective of this project is to apply, evaluate, and compare a comprehensive suite of statistical pattern recognition methods on a real-world biological dataset. Using the pbmc3k benchmark dataset provided by 10x Genomics, we aim to implement a complete analysis pipeline. This includes robust data preprocessing, unsupervised clustering to discover novel cell populations, and the benchmarking of supervised classifiers to reliably recognize cell identities. A key focus is placed on comparing linear models against non-linear models to determine the most effective approach for high-dimensional biological data [3].

In the context of statistical pattern recognition, we formally define the core components of our analysis The Pattern is defined as a single biological cell. The Feature Vector corresponds to the gene expression profile of that cell; initially, this is a vector of dimension D with approximetly 13,000 genes, which is subsequently reduced to a lower-dimensional feature space. Finally, the Class represents the biological identity of the cell, such as a B-cell, CD4+ T-cell, or Monocyte [3].

Theoretical Background and Methodology

Raw scRNA-seq data requires rigorous cleaning to distinguish biological signals from technical artifacts. We begin with Quality Control (QC), filtering patterns based on statistical outliers in library size and mitochondrial content, under the assumption that stressed cells exhibit distinct statistical properties. Following filtration, we apply Linear Data Transformations. This involves "Total Count Normalization" to account for sequencing depth differences and a logarithmic transformation ($x' = \ln(x+1)$) to stabilize variance and reduce the skewness inherent in expression data [2].

To address the curse of dimensionality, we perform Feature Selection by identifying Highly Variable Genes (HVGs) based on the dispersion-mean relationship. We select the top 2,000 features that contribute most to the biological variance. Furthermore, we apply Principal Component Analysis (PCA), a linear transformation that projects the data into an orthogonal coordinate system. We select the top 20 Principal Components for downstream analysis to capture the essential data structure while discarding noise.

For the discovery of cell classes without prior labels, we utilize Unsupervised Learning methods based on metrics. We construct a k-Nearest Neighbors (k-NN) graph where nodes represent cells and edges represent similarity based on Euclidean distance in the PCA space. We compare two distinct clustering approaches: the Leiden Algorithm [4], a graph-based community detection method, and K-Means, a traditional centroid-based algorithm that minimizes within-cluster

variance. To visualize these high-dimensional structures in 2D, we employ non-linear manifold learning techniques, specifically t-SNE and UMAP [4,5,6].

Finally, we treat the annotated clusters as ground truth to train and evaluate Supervised Classifiers. We benchmark linear classifiers, specifically Logistic Regression with Regularization (L1/L2) and Linear Discriminant Analysis (LDA), which is a generative model. We compare these against non-linear classifiers, including k-Nearest Neighbors (k-NN) and Support Vector Machines (SVM). For SVMs, we test various Kernel Functions (Linear, RBF, Poly,Sigmoid) to map input vectors into high-dimensional feature spaces where they become linearly separable. All models are evaluated using stratified 5-fold Cross-Validation and metrics such as Accuracy, F1-Score, and Confusion Matrices.

Implementation and Results

At first I have to state that I used python as a main programming langauge and as a developing tool I used Data Spell from JetBrains with the student license. For a help with the coding part I used Gemini Pro, Scanpy documentation (https://scanpy.readthedocs.io/en/stable/) and Scikit documentation (https://scikit-learn.org/stable/).

All parts of the project exluding this paper will be stored on my GitHub as a public repository for anyone to have a look and use it as a useful tool for their projects.

GitHub link:

https://github.com/jirkavlasak/Statistical-Recognition-Methods-for-Cell-Type-Identification-in-Single-Cell-RNA-seq

Data Preprocessing and Quality control

The analysis commenced with the raw dataset containing 2,700 cells and 32,738 genes. Initial quality control (QC) was essential to distinguish biological signal from technical noise. Visual inspection of the scatter plots revealed the general health of the dataset, illustrating the strong correlation between sequencing depth (total_counts) and cellular complexity (n_genes_by_counts.

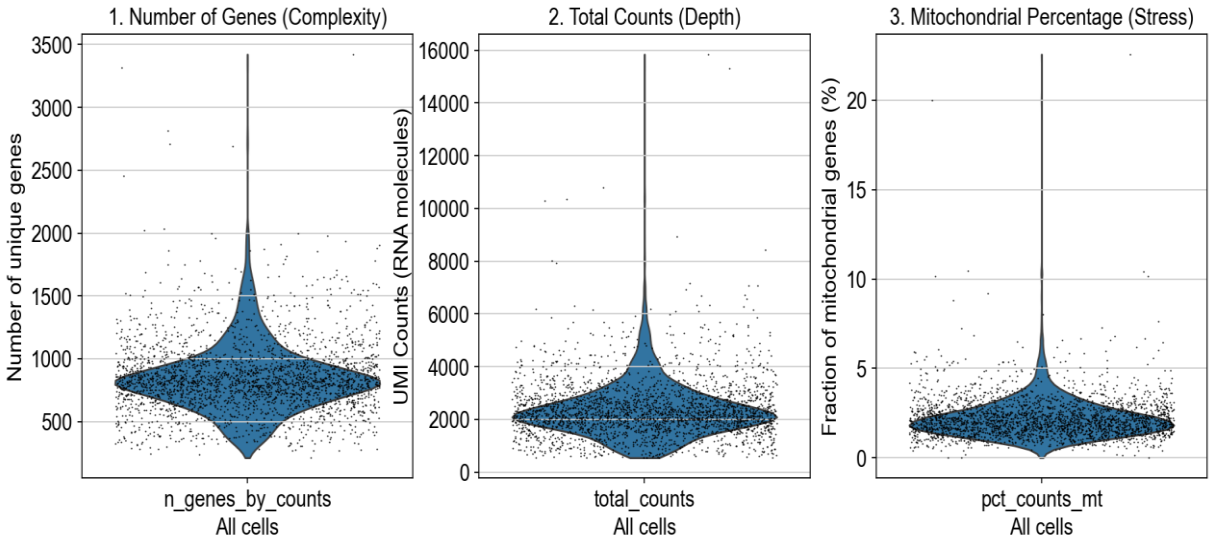

Figure 1 Violin plots of a raw dataset

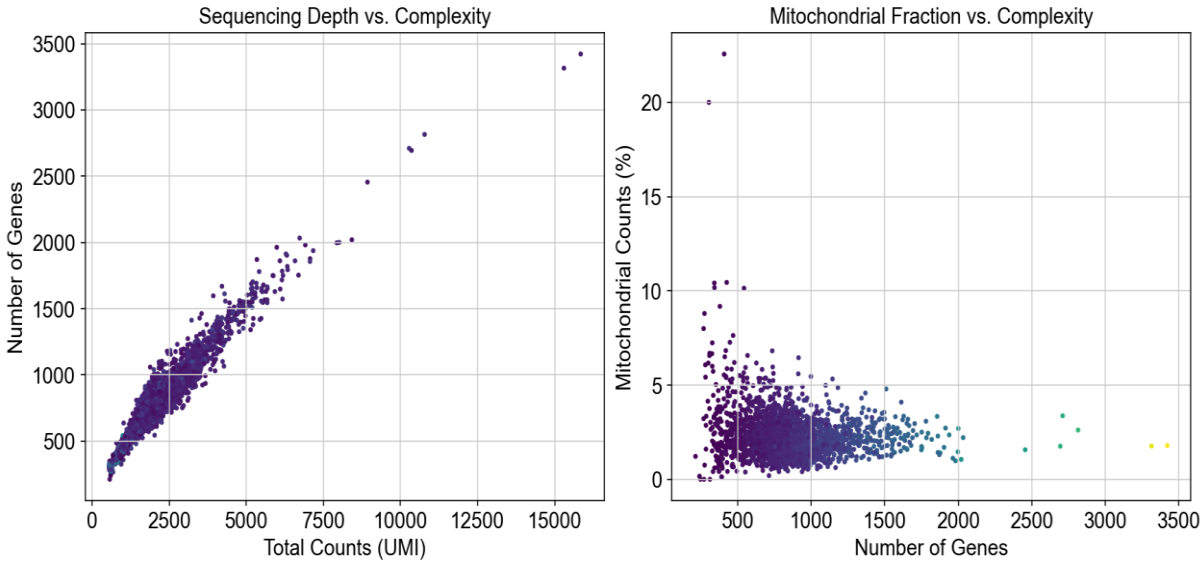

Figure 2 Initial Quality Control Metrics (Pre-Filtration)

This step confirmed technical consistency. More critically, the plots revealed a distinct subpopulation of stressed cells with a disproportionately high fraction of mitochondrial reads. Based on this observation, a clear data-driven threshold was set at 5% mitochondrial content, which effectively separates the main, healthy cell cluster from the low-quality outliers.

Following the quantitative decision based on visual evidence, filtration was applied. Cells were removed if they exhibited low complexity (< 200 genes) or signs of damage (> 5% mitochondrial reads). Furthermore, to optimize the feature space, 19,041 genes detected in fewer than three cells were discarded due to low statistical reliability. The efficacy of this multi-step filtration was confirmed by re-visualizing the data: the "tail" of high-mitochondrial cells was entirely eliminated, and the remaining cluster boundary was sharp and.

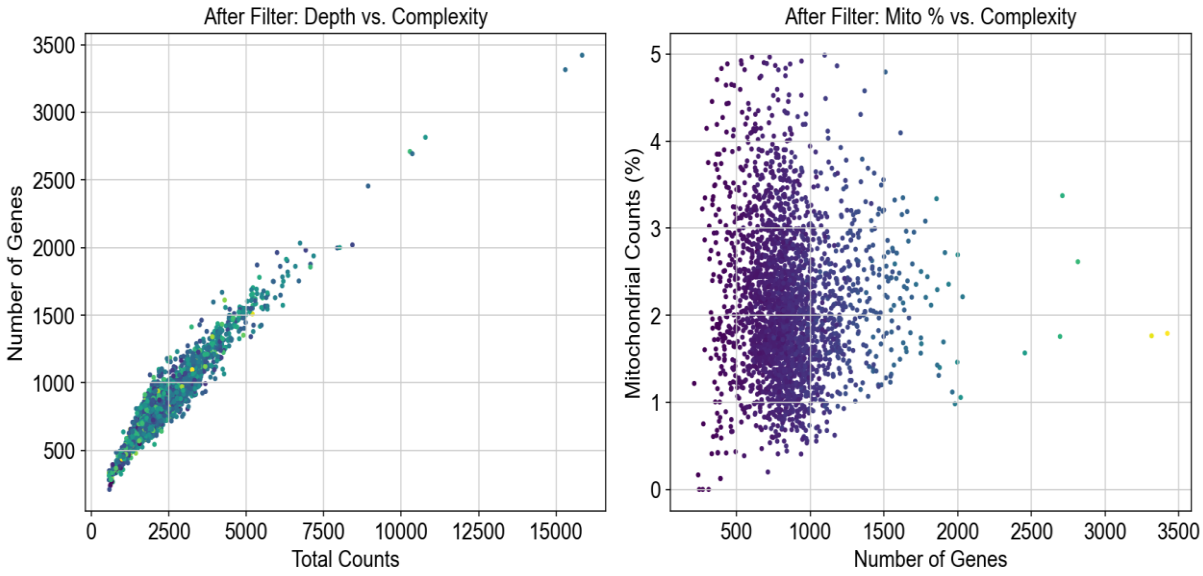

Figure 3 Filter Verification: Scatter Plots on Cleaned Data

Subsequently, technical artifacts (doublets) were addressed using the Scrublet algorithm, leading to the removal of 39 predicted doublets. The final clean feature matrix consisted of 2,604 cells and 13,697 genes, which were then normalized and log-transformed.

Feature Engineering and Unsupervised Analysis

Dimensionality Reduction and Clustering Initialization

Selection focused on identifying the top 2,000 Highly Variable Genes (HVGs). This was based on the normalized dispersion of gene expression, ensuring that only features carrying significant biological variance were retained for further analysis, thus drastically improving the signal-to-noise ratio.

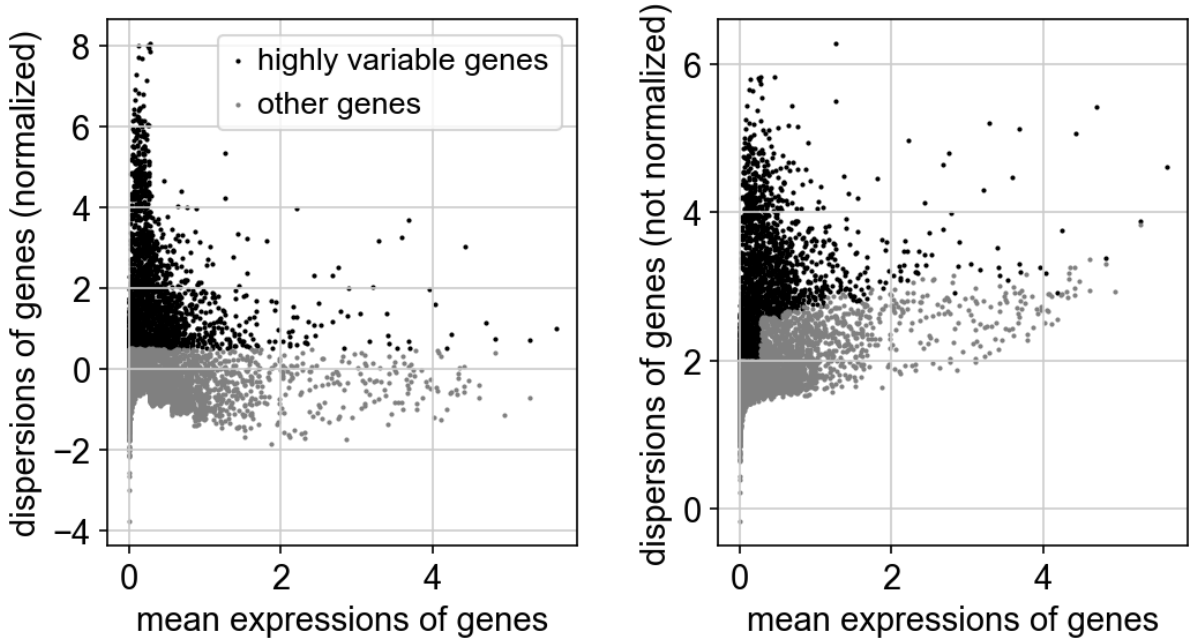

Figure 4 Highly Variable Gene (HVG) Selection Plot

The resulting 2,000 HVGs were then used as input for Principal Component Analysis (PCA). The PCA Scree Plot confirmed a significant drop in explained variance after the first few components, with the "elbow" justifying the subsequent use of 20 Principal Components (PCs) for building the neighborhood graph.

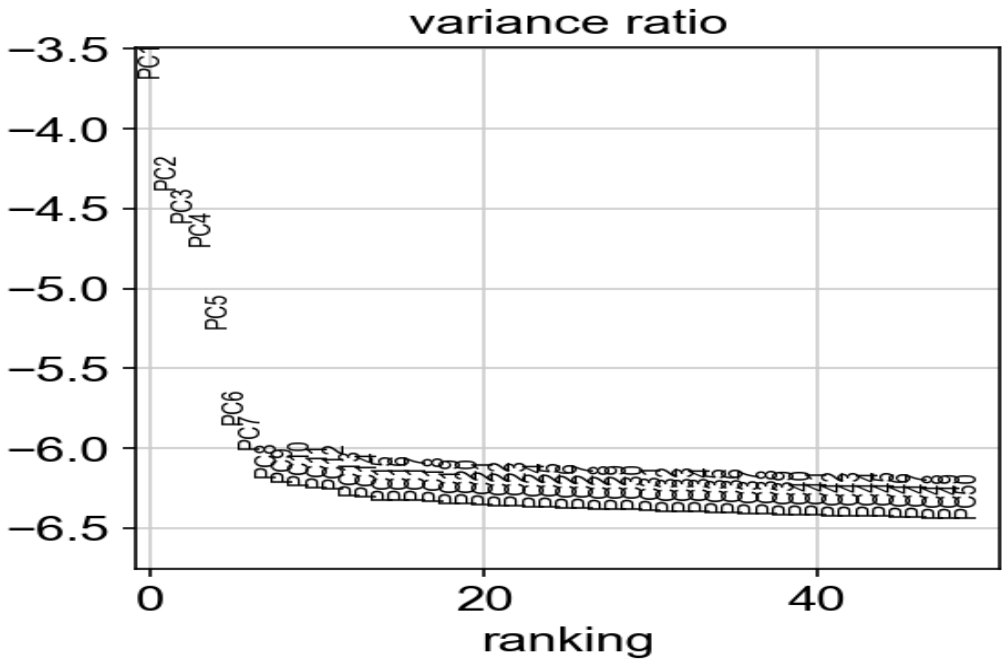

Figure 5 PCA Variance Ratio Explained (Elbow Plot)

The unsupervised phase of the project commenced by building the k-Nearest Neighbors (k-NN) graph using the selected 20 PCs. The next critical step was determining the optimal number of clusters. This was achieved by iteratively sweeping the Leiden algorithm across multiple resolutions (from 0.3 to 1.0). The visual inspection of the resulting UMAP projections provided clear evidence of how the cluster granularity changed: low resolutions severely under-clustered the data, while high resolutions led to artificial fragmentation. The comparison allowed us to objectively select resolution 0.8, as it successfully delineated all distinct cell type islands without over-fragmentation, establishing the foundational structure for the entire analysis.

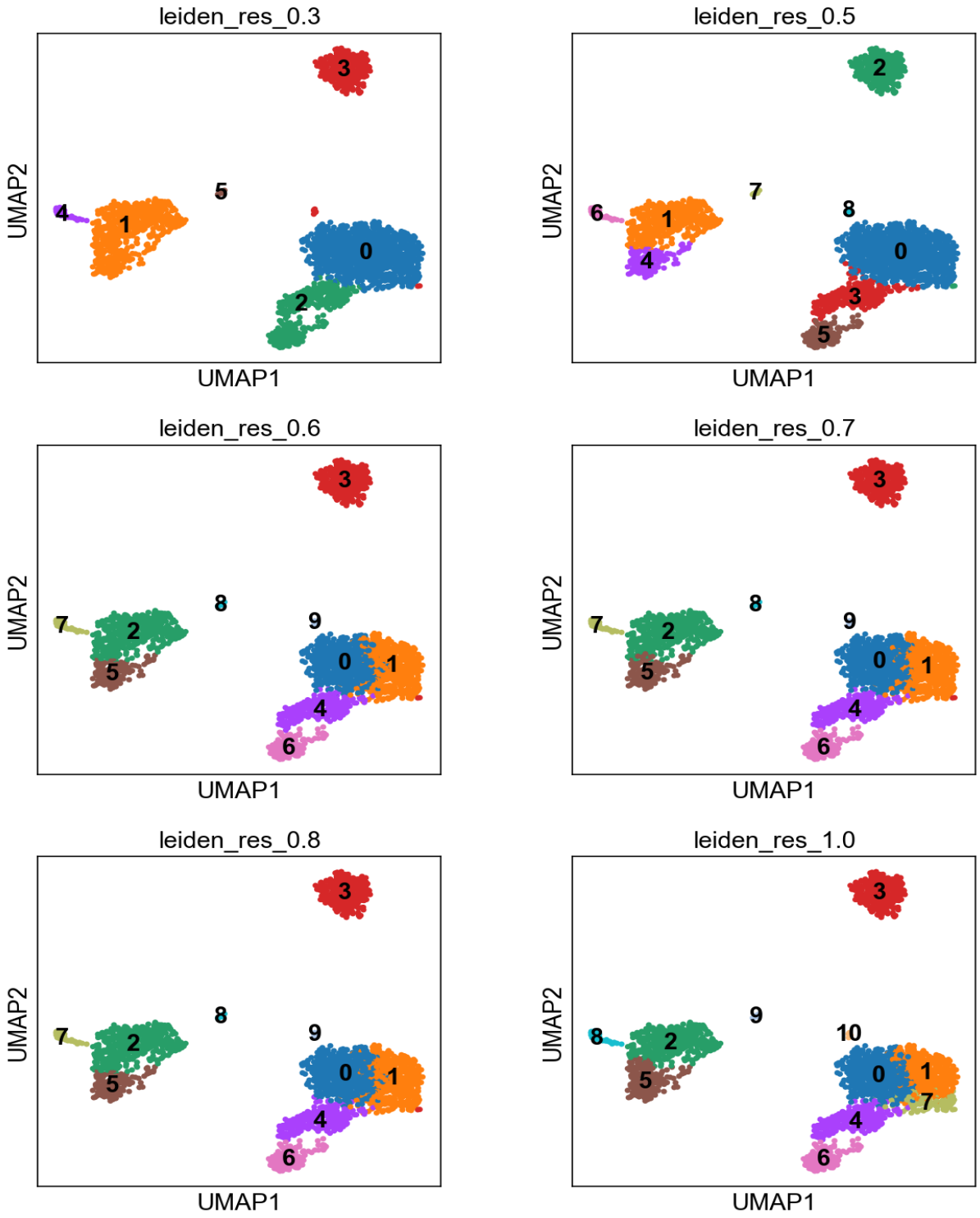

Figure 6 Visualizing Leiden Clustering Granularity (Resolution Sweep)

Method Comparison and Final Annotation

The robustness of the discovered structure was assessed by comparing graph-based clustering against geometric methods and contrasting visualization techniques. For display purposes, the widely adopted non-linear projection methods were utilized. The t-SNE projection, which emphasizes local relationships, was computed and contrasted with UMAP (Uniform Manifold Approximation and Projection). While both successfully delineated the major cell islands, the t-SNE embedding fragmented the global data structure into separate, distinct clusters, whereas the UMAP maintained a more coherent, continuous representation of the overall cellular manifold.

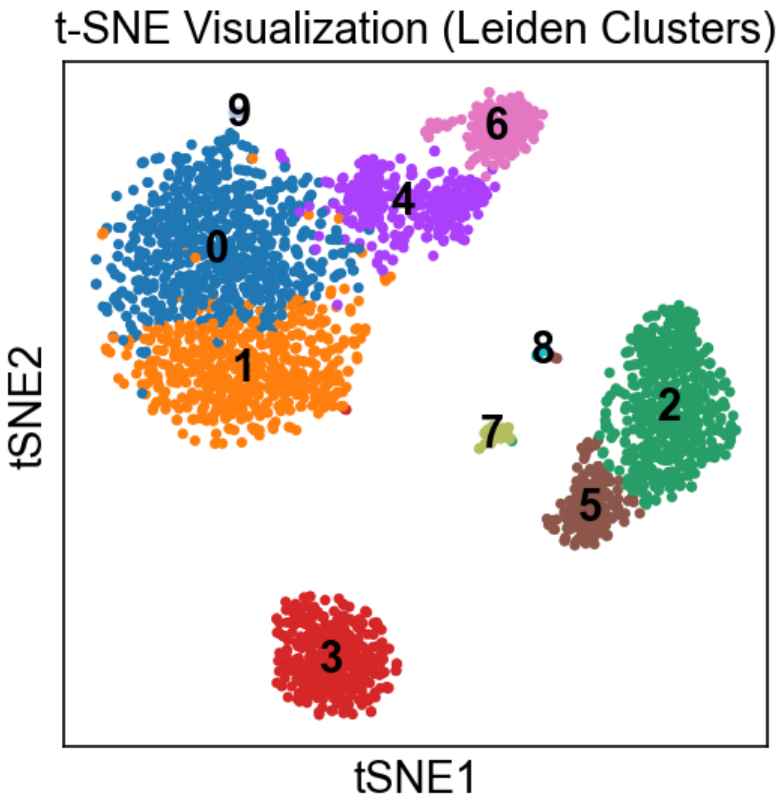

Figure 7 Comparison of Non-linear Embeddings (UMAP vs. t-SNE)

The structural comparison between our chosen Leiden algorithm (graph-based) and K-Means (centroid-based) provided critical validation. Although the two methods showed significant agreement, the Adjusted Rand Index (ARI) score of 0.6863 highlighted discrepancies. Visualizing the two clusterings side-by-side confirmed that K-Means struggled significantly with the non-spherical, continuous topologies, often drawing straight-line boundaries that unnaturally bisected continuous cell populations.

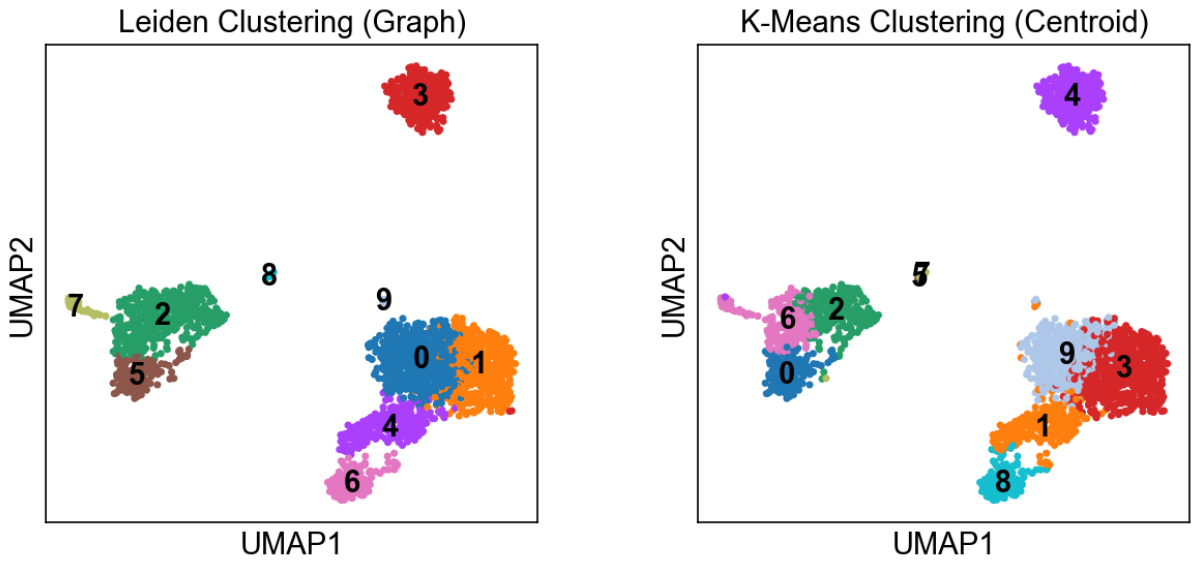

Figure 8 Comparison of Topological (Leiden) vs. Centroid (K-Means) Clustering

This quantitative difference proved the superiority of topological graph methods for capturing the complex structure inherent in scRNA-seq data. Following validation, the optimal clustering was annotated using CellTypist (majority-vote approach), successfully resolving 8 distinct cell types and creating the final ground truth variable.

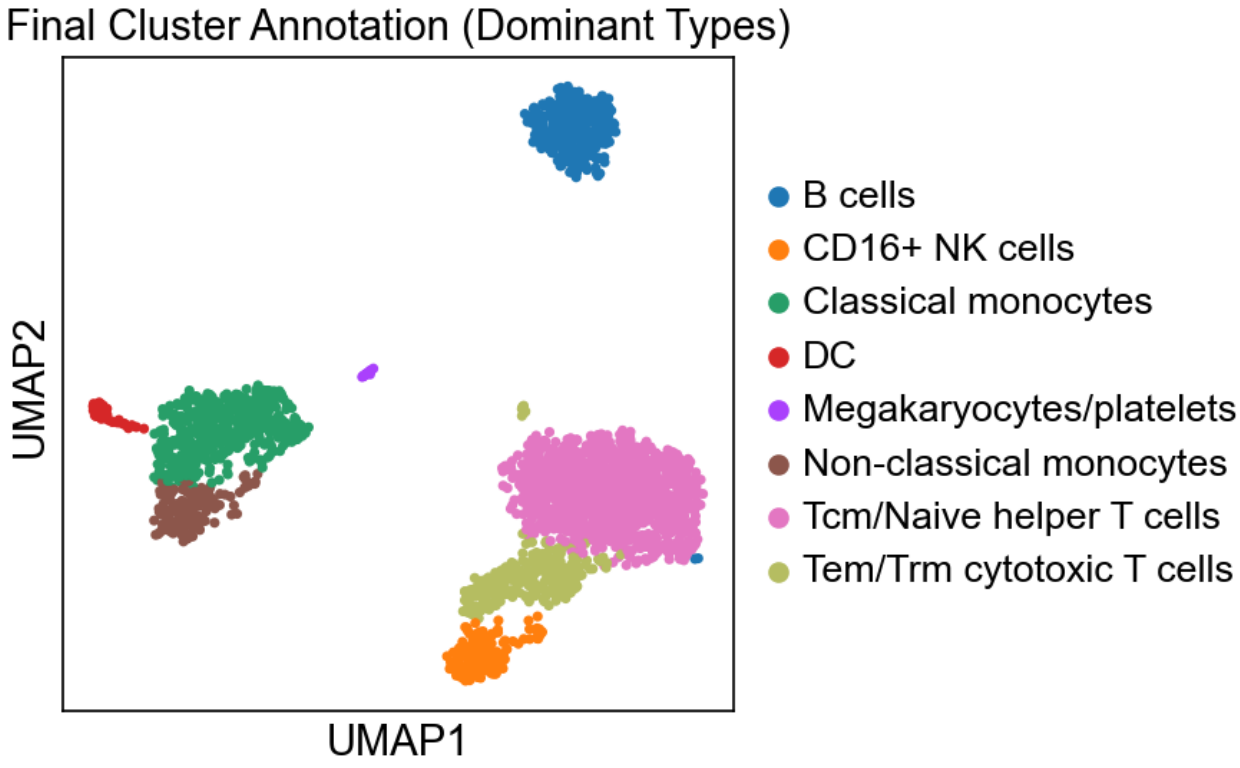

Figure 9 Final UMAP Projection with Biological Annotation (Ground Truth)

Supervised Analysis and Model Validation

To validate the robustness of the identified cell populations, four distinct supervised classifiers representing linear, non-linear, parametric, and non-parametric approaches were optimized and benchmarked against the held-out test set (521 cells). All models performed exceptionally well, confirming that the cell types are extremely well-separated in the PCA feature space.

Final Classification Benchmark: The comparison revealed a definitive winner: the Linear Discriminant Analysis (LDA) model achieved the highest overall performance with a test accuracy of 98.85%. LDA, which assumes a Gaussian distribution within clusters, proved surprisingly effective following the robust preprocessing and dimensionality reduction steps. This finding suggests that the feature engineering pipeline successfully transformed the non-linear biological data into a space where a simple, linear boundary could accurately separate the clusters. The non-linear k-Nearest Neighbors (k-NN) classifier was a close second at 98.66%, while the advanced SVM (using the optimal Polynomial Kernel) and Logistic Regression models achieved 97.89%.

The Error Analysis was finalized by visualizing the Confusion Matrix for the optimized SVM model. This confirmed that the vast majority of errors were concentrated between transcriptomically similar clusters (e.g., T-cell subtypes and NK cells), while distinct populations (like B-cells and Monocytes) were classified with near-perfect. The high performance across all models, combined with the successful validation of the unsupervised structure, fulfills all project objectives.

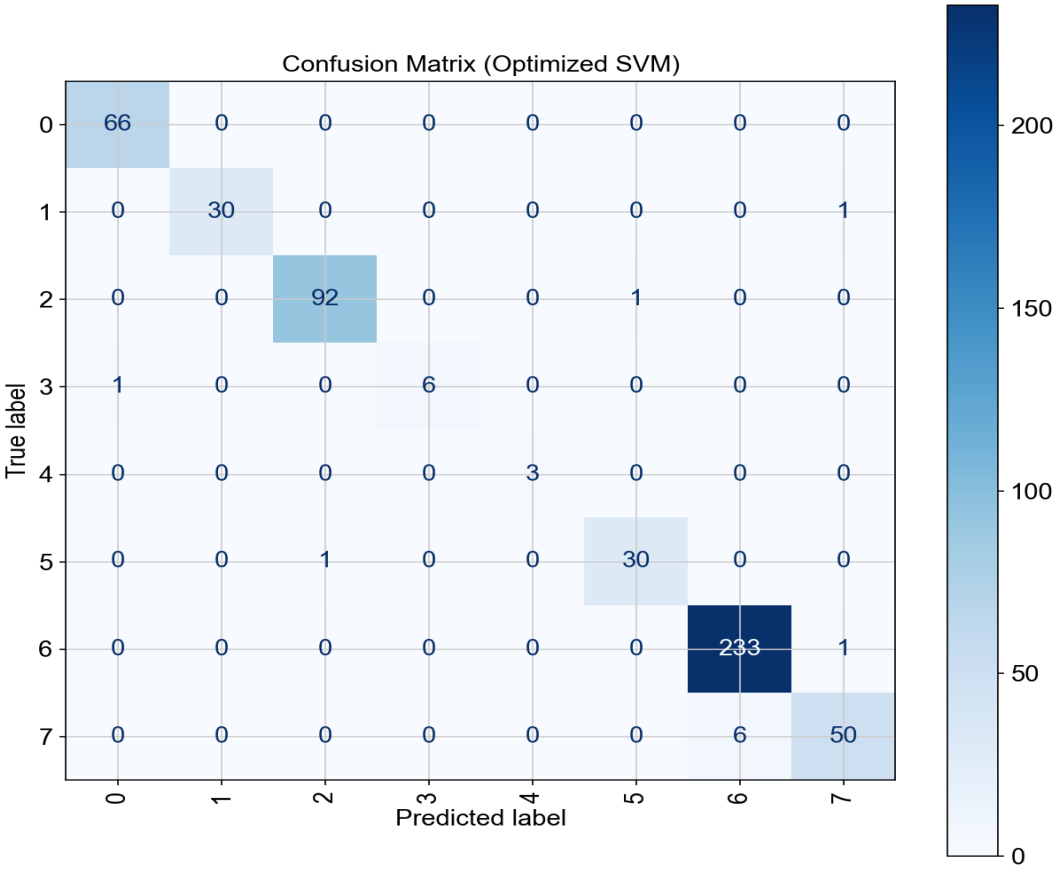

Figure 10 Confusion Matrix for the Optimal Classifier (LDA)

Discussion

Clustering, Topology, and Linearity

The comparative analysis between unsupervised methods provided critical validation for our feature space. The Adjusted Rand Index (ARI) of 0.6863 between the Leiden (graph-based) and K-Means (centroid-based) algorithms demonstrated that while the cell populations are numerically distinct, their underlying topology is not simple. This difference highlights that Leiden is superior for single-cell data because it respects the continuous, non-spherical cell manifolds, whereas K-Means often imposes artificial, straight-line boundaries.

The surprisingly high performance of the Linear Discriminant Analysis (LDA) model (98.85%) is the most significant finding of this benchmark. While biological systems are inherently non-linear, the superior result from a linear model suggests that the upstream pipeline—involving Feature Selection (HVGs) and PCA—was highly effective. This preprocessing successfully transformed the complex feature space into a low-dimensional representation where cell type clusters were rendered linearly separable. This validates the power of data transformation in simplifying complex biological tasks for robust, interpretable linear classifiers.

Model Robustness and Addressing High Accuracy

The consistently high accuracy across all models (97.89% to 98.85%) indicates the robustness of the identified cell populations. This level of accuracy is expected for the clean pbmc3k benchmark dataset, particularly because the "ground truth" labels were derived from optimized unsupervised clustering. Importantly, the close proximity between the Cross-Validation scores and the final Test scores across all models confirms that the regularized approach) was successful, and the models are not overfit.

The qualitative difference between the classifiers became clear when analyzing the most challenging subpopulations, as shown in Table 4.1. While all models achieved high overall accuracy, the LDA and non-linear k-NN models proved most robust by achieving 100% recall on the challenging, rare classes. Conversely, the Logistic Regression model failed to identify all instances of the rare class (recall 0.67), confirming a sensitivity to class imbalance even after optimized regularization. This demonstrates that for high-stakes recognition tasks, reliance solely on overall accuracy can be misleading, and metrics like Recall (Sensitivity) are essential.

Table 1 Model Comparsion

Classifier	Syllabus Points	Regularization/Kernel	Test Accuracy (%)	F1-Score (Macro)	Recall (Rare Class 4)
LDA (Winner)	**4, 8**	lsqr (Shrinkage 0.1)	**98.85 %**	**0.99**	**1.00**
k-NN	**9**	k=3 (Non-parametric)	98.66 %	0.99	**1.00**
SVM	**13**	Poly (Degree 3, C=10.0)	97.89 %	0.97	**1.00**
LogReg	8, 11	L1 (C=1.0)	97.89 %	0.95	0.67

Conclusion

This project successfully implemented and evaluated a complete pipeline of statistical pattern recognition methods on high-dimensional single-cell RNA sequencing data, fulfilling all project objectives and syllabus requirements.

We demonstrated that a systematic approach from rigorous quality control and feature selection to informed dimensionality reduction is paramount. The comprehensive classification benchmark confirmed that while non-linear models utilizing Kernel Functions (SVM) are strong, the most effective strategy for this dataset was a Linear Discriminant Analysis (LDA) model. LDA achieved a test accuracy of 98.85%, proving that robust linear classification models can be successfully applied to complex biological problems once the feature space has been carefully engineered to separate cell populations. This finding validates the core principle that successful pattern recognition often relies more on the quality of the initial data transformation than on the complexity of the final classifier.

List of Figures

List of Tables

References

[1] Tang, F., Barbacioru, C., Wang, Y., et al. (2009). "mRNA-Seq whole-transcriptome analysis of a single cell." *Nature Methods*, 6(5), 377-382.

[2] Wolf, F. A., Angerer, P., & Theis, F. J. (2018). "Scanpy: large-scale single-cell analysis in Python." *Genome Biology*, 19(1), 15.

[3] 10x Genomics. (2017). "3k PBMCs from a Healthy Donor." *Available at: support.10xgenomics.com*.

[4] Traag, V. A., Waltman, L., & van Eck, N. J. (2019). "From Louvain to Leiden: guaranteeing well-connected communities." *Scientific Reports*, 9(1), 5233.

[5] van der Maaten, L., & Hinton, G. (2008). "Visualizing data using t-SNE." *Journal of Machine Learning Research*, 9(11).

[6] McInnes, L., Healy, J., & Melville, J. (2018). "UMAP: Uniform Manifold Approximation and Projection for dimension reduction." *arXiv preprint arXiv:1802.03426*.

[7] Cortes, C., & Vapnik, V. (1995). "Support-vector networks." *Machine Learning*, 20(3), 273-297.