An Analysis of Colorectal Cancer Using Machine Learning

Jai Mehta[1]
jmehta2@vols.utk.edu

Angelina Ju[1]
aju@vols.utk.edu

Eli Fisk[1]
efisk@vols.utk.edu

Andrew Berard[1]
aberard@vols.utk.edu

Sidarth Santhosh Kumar[1]
ssanthos@vols.utk.edu

University of Tennessee, Knoxville[1]

Abstract

Colorectal cancer (CRC) is currently the second leading cause of cancer-related deaths worldwide, accounting for an estimated 52,900 deaths in 2025 (8.6% of all cancer deaths) [1]. Deepening our understanding of CRC is crucial to determining its underlying mechanisms. The use of unsupervised machine learning (ML) algorithms on complex, high-dimensional single-cell RNA sequencing (scRNA-seq) data could provide insight into cancer through mapping patterns within genomic sequences. To effectively analyze scRNA-seq data, however, it is crucial to use dimensionality reduction techniques first. Testing several reduction techniques found that Uniform Manifold Approximation and Projection (UMAP) was the most effective for reducing the non-linear scRNA-seq data's dimensionality. When UMAP was combined with the Leiden clustering algorithm, it produced the strongest cluster recovery, as indicated by the highest Adjusted Random Index (ARI) score. This study demonstrates that readily available unsupervised ML methods can be semi-reliably used to identify true patterns in genomic data. These results also encourage further research into leading dimensionality reduction and clustering algorithms to develop faster and more effective tools for identifying CRC biomarkers, as well as biomarkers for other cancers, for prevention and treatment.

1. Introduction

CRC is currently the second-most deadly cancer in the world. To understand the underlying mechanisms of disease, researchers have engaged in Next-Generation (next-gen) sequencing, using highly parallel and specific techniques to generate information. Of the next-gen sequencing technologies, the most prominent type is single-cell RNA-sequencing (scRNA-seq) data. scRNA-seq data has been shown to provide extensive insight into human health by capturing information at the single-cell level, allowing for specific cell types to be captured [2]. This has allowed for cancers such as CRC to be mapped according to their single-cell genetic expression, where researchers can discover specific biomarkers and gene pathways related to the cancer.

However, a strong challenge to biomarker discovery is the extremely high dimensionality of the datasets. With cell quantity ranging from several thousand to several hundred thousand, along with tens of thousands of genes for each cell, datasets quickly grow. Thus, finding relationships within non-linear, high-dimensional data is quite challenging. To combat this issue, researchers have harnessed dimensionality-reduction techniques to reduce redundancy and increase inference potential. Additionally, various clustering techniques have been used to find relationships between cells for cell type analysis.

The first challenge of dimensionality can be addressed through the use of Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP). PCA is primarily used for linearly variate data, while t-SNE and UMAP serve to preserve non-linear information in datasets, and the difference in use cases will allow for greater difference in visualization and applied clustering to pinpoint the best dimensionality reduction technique(s).

The second challenge of identifying an optimal clustering technique will be addressed through testing algorithms such as K-means, Leiden, Louvain, DBSCAN, and Hierarchical DBSCAN on the dimensionally reduced data. The use of an Adjusted Rand Index (ARI) is the critical component to evaluating the accuracy of techniques when compared to ground truth labels associated with our dataset.

2. Methods

This study uses two scRNA-seq datasets: a SMART-seq dataset, and a 10x Genomics dataset [3], where the 10x dataset size (43817 cells by 13538 genes) is much larger than the SMART-seq (10468 cells by 15179 genes) dataset. This larger set provides samples if needed, but it can be difficult to run with constrained RAM, requiring the use of the SMART-seq dataset for most trials. The data needed pre-processing to account for missing values, which were filled with zeros because of the difficulty imputing gene expression since each cell is different. For normalization, the scale of counts in each cell needed to be normalized so that the total counts per cell equal a target value, which was set to 10,000. After this was done, a natural logarithm transformation was applied to the dataset to reduce the asymmetric values in the data and to help stabilize the variance.

2.1 Feature engineering

To avoid potential doublets or high-library cells, cells with an unusually high count (>35,000) were removed, and cells with a high mitochondrial content (>20%) were removed since it often indicates stressed or dying cells. Genes expressed in fewer than 10 cells were removed to reduce noise from less expressed genes. These methods were applied to clean and normalize the data, resulting in a much more reliable and interpretable dataset that will be suitable for later calculations. Using the Seurat method to help capture the most biological variability across cells, only the top 2,000 highly variable genes were selected. The final size of the SMART-seq dataset is 10288 cells by 15178 genes, while the final size of the 10x dataset is 43817 cells by 13526 genes.

2.2 Dimensionality Reduction

Dimensionality reduction techniques were utilized to distill the high-dimensional data into latent embeddings. PCA, t-SNE, and UMAP were used due to their high efficacy for transforming RNA-seq data into a lower dimensional space [4].

PCA utilizes linear eigendecomposition of a covariance matrix to generate principal components (PCs) that capture maximum uncorrelated variance. These principal components are projected onto a lower dimensional subspace. PCA is often used as a baseline dimensionality reduction technique due to its computational efficiency and linear representation.

t-SNE uses perplexity to control neighborhood sizes of local neighborhoods in a high-dimensional space. The algorithm minimizes Kullback–Leibler divergences between high-dimensional and low-dimensional probability distributions of their respective pairwise distances. This non-linear dimensionality reduction technique allows for capture of complex relationships between genes, preserving the local structure but distorts the global structure.

UMAP assumes that the data lies on a Riemann manifold and constructs a low-dimensional space that preserves local and some global structure of the high-dimensional pairwise distance space.

These methods were applied to the SMART-seq's normalized gene expression data to generate a 2-dimensional representation of the data. To observe the similarity of the reduced data in 2D space, Leiden clustering was applied to the reduced data. The same hyperparameters were applied to the provided "Global" UMAP plots provided in the dataset.

The clusters were compared using the ARI score. ARI is utilized to compare unsupervised clusters to each other using the similarity of clusters. Values range from –1 to 1, where 1 indicates perfect cluster similarity, and –1 indicates no similarity at all. Cluster differences were visualized using confusion matrices.

2.3 Clustering Comparison

Various clustering methods were used to observe their similarity to the "true label" of cell type. K-means clustering was used as a baseline to generate unsupervised clusters of the normalized gene expression data. Louvain, Leiden, Hierarchical, Density-based (DBSCAN), and hierarchical density-based (HDBSCAN) clustering approaches were used to generate clusters from the data.

K-means clustering uses a hyperparameter value, *k*, to determine the number of clusters to create. K centroids are randomly initialized, and all cells within a radius of the centroid are placed within the cluster. The centroid is then updated to be the center of the clusters; this process is repeated until the assignments remain unchanged.

Louvain clustering, instead of k-means, determines the number of clusters inherently through a greedy approach. Louvain optimizes the modularity between clusters in an agglomerative approach, resulting in a number of clusters with the highest modularity score. Leiden clustering builds upon this algorithm by ensuring that each cluster is one component instead of separate clusters, which is more biologically relevant. Leiden also prevents local minima convergence by performing a refinement of the communities found in each cluster after linkage.

Hierarchical clustering, similar to Leiden and Louvain, is an algorithm that connects clusters based on the distances between clusters. For this project, agglomerative clustering was used, and distances were determined using complete linkages. Based on the dendrogram produced by the model, a certain number of optimal clusters were determined and used to fit the model.

Density-based clustering was used to understand if the data exists in particularly dense versus sparse regions. DBSCAN was used to separate dense regions from potential noise. HDBSCAN builds upon this method by identifying different levels of density and separates classes by varying densities through a hierarchical approach.

ARI was used to compare the generated clusters to the true labels, and cluster differences were visualized with confusion heatmaps.

2.4 Fine-tuning

Optuna[5] was used to optimize the hyperparameters of both the dimensionality reduction and clustering tools. Grid Search was utilized to find optimal hyperparameters that generated the highest ARI.

2.5 Initial Dataset

The provided t-SNE and UMAP coordinates along with the labels were plotted using Scanpy [6].

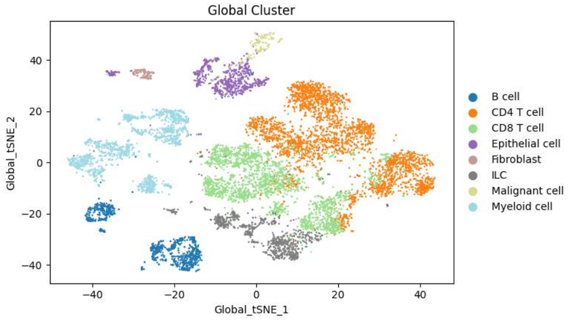

Figure 1. Provided t-SNE plot of normalized data, colored by provided cell type, the “true label.”

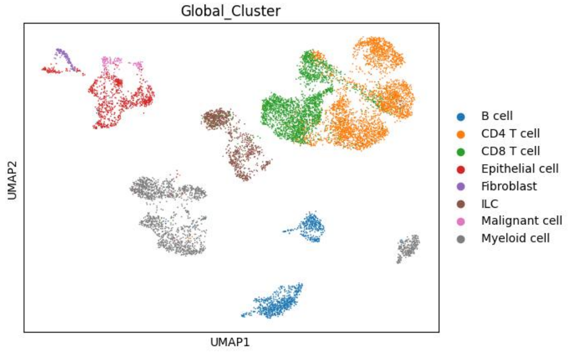

Figure 2. Provided UMAP plot of normalized data, colored by provided cell type.

The above figures show how each cell type is neatly separated between clusters in dimensional space.

In this work, we compare various dimensionality reduction and classification techniques to find the optimal techniques to generate clusters representing cell types in scRNA-seq data.

3. Results

3.1 Dimensionality-Reduction Evaluation

Due to the high dimensionality of the scRNA-seq dataset, dimensionality reduction was required to allow for efficient and meaningful clustering. The central challenge in this process involved selecting a technique capable of keeping a biologically relevant structure, particularly the separation between distinct cell populations, so that clustering algorithms could accurately separate differing cell types.

3.1.1 t-SNE Dimensionality-Reduction Evaluation

The performance of the t-SNE dimensionality-reduction technique was evaluated on the processed SMART-seq dataset to determine its ability to preserve a biologically meaningful structure. The resulting two-dimensional t-SNE embeddings demonstrated a substantial mixing of cell populations, with limited boundary formations between known cell types. Visual inspection revealed that many clusters displayed by the Global t-SNE reference embedding were not preserved in the t-SNE projections. This indicated that the t-SNE dimensionality reduction technique did not generate a manifold that separated the biological identities present in the ground-truth annotations.

To assess whether hyperparameter selection contributed to the weak separation observed in the initial t-SNE embeddings, hyperparameter optimization was performed using Optuna with 20 iterations. The optimization process explored perplexity values in the range of 20-70 and learning rates spanning 1e-3 to 1e2 on a logarithmic scale. Each trial generated a new t-SNE embedding, which was evaluated using the trustworthiness metric to quantify the preservation of the high-dimensional neighborhood structure in the reduced space. The optimization consistently converged toward parameter configurations that maximized trustworthiness; however, even under the best-performing settings, the resulting embeddings continued to display significant mixing of biologically distinct cell types. The SMART-seq dataset t-SNE projections were plotted with cell-type coloring to help visualize overlapping clusters.

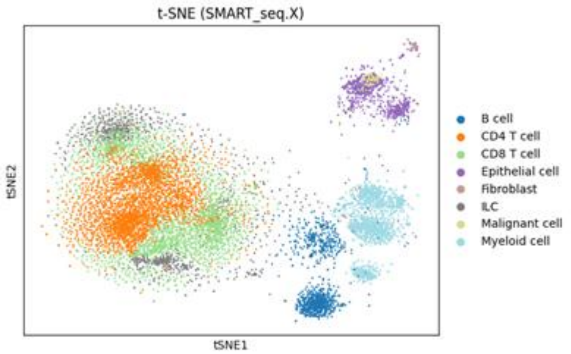

Figure 3. SMART-seq dataset after performing t-SNE dimensionality reduction, colored by provided cell type.

Fig. 3 shows that applying t-SNE directly to the raw data does not produce cluster separations that divide the cells correctly based on the cell types. Thus, t-SNE would not be a good choice for dimensionality reduction.

3.1.2 PCA Dimensionality-Reduction Evaluation

Similarly, PCA was applied to the SMART-seq dataset to evaluate its ability to capture biological expression while reducing the feature set. This results in the highest variable genes being incorporated into the PCA embedding. This resulted in a structure not true to the dataset.

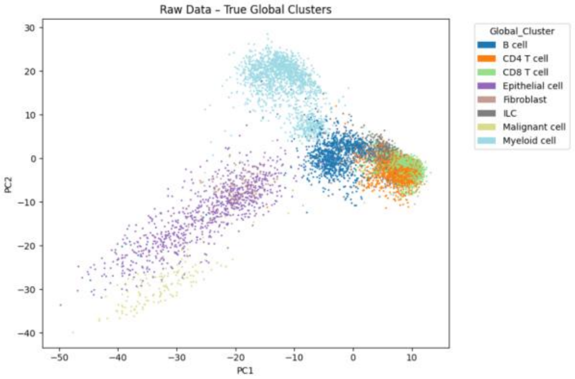

Figure 4. SMART-seq dataset after performing PCA dimensionality reduction, colored by provided cell type.

This was expected as PCA relies on linear algebra, and the dataset was non-linear. PCA flattens this non-linear data and loses the important geometry. Visually, this can be seen as most of the clusters overlap.

3.1.3 UMAP Dimensionality-Reduction Evaluation

The SMART-seq dataset's dimensions were reduced using the UMAP technique.

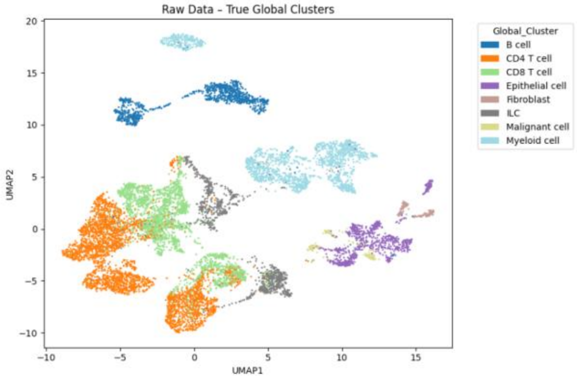

Figure 5. SMART-seq dataset after performing UMAP dimensionality reduction, colored by provided cell type.

Compared to PCA, UMAP was more effective at preserving the dataset structure due to its non-linear transformations. The resulting embedding improved visual separability of cell clusters and created distinct regions.

The density of the clusters indicates strong correlation with the dataset's true labels. This suggests that UMAP maintains meaningful biological relationships.

3.2 scRNA-seq Clustering Analysis

Due to the high-dimensional nature of scRNA-seq data, clustering analysis is challenging yet crucial in grouping cells with similar gene expression profiles to identify distinct cell types, cell states, and subpopulations. This process involves unsupervised clustering algorithms to generate cell type-specific clusters and use marker genes to label the clusters.

3.2.1 K-means

K-means clustering was applied to the SMART-seq dataset following dimensionality reduction to evaluate its effectiveness in recovering biologically meaningful cell-type structure. The algorithm was tested across a range of cluster counts, and the optimal value of *k* was determined using Optuna-based hyperparameter optimization, with the ARI score as the objective metric. The optimization procedure consistently selected *k* values that aligned with the approximate number of biological cell types in the dataset.

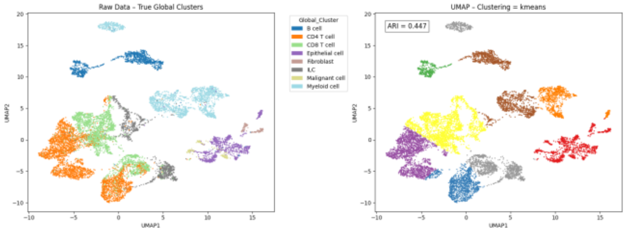

Figure 6. Example of K-means on a UMAP reduced dataset with n-clusters parameter set to 10.

As seen in Fig. 6, the resulting cluster assignments demonstrated limited alignment with the ground-truth labels. As a result, the final ARI score for clustering with K-means was 0.447, which will be used as a baseline for the other clustering methods.

3.2.2 Louvain and Leiden Clustering Comparison

The Louvain and Leiden clustering algorithms were applied to classify the normalized gene expression data, which offers several areas of improvement in comparison to K-means. These graph-based community methods better capture the structure of the data, and evaluating the results will help identify whether each algorithm is over-splitting or under-splitting biologically distinct cell groups.

Both the Louvain and Leiden clustering algorithms take a resolution hyperparameter, which controls the scale of the partitions and the number of clusters that will appear. By increasing this value, the algorithms will prioritize smaller and denser communities, creating smaller and therefore a larger amount of clusters. A lower resolution value results in fewer, larger clusters.

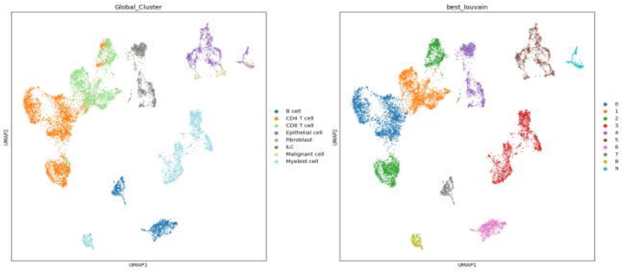

Figure 7. Louvain algorithm clustering with a resolution hyperparameter of 0.1472, determined by Optuna study.

As Louvain and Leiden heavily rely on this resolution parameter to provide meaningful results, the Optuna framework can be utilized to systematically search the parameter space, evaluate a certain number of trials, and select a resolution that best matches the biological labels. Fifty trials of a Louvain study were conducted with a resolution range of 0.05 to 1.75, allowing Optuna to explore any extremes without bias. Each trial was evaluated by computing the ARI between the known "true label" and the algorithmically computed clusters, and the maximized ARI result was selected as the optimal resolution parameter.

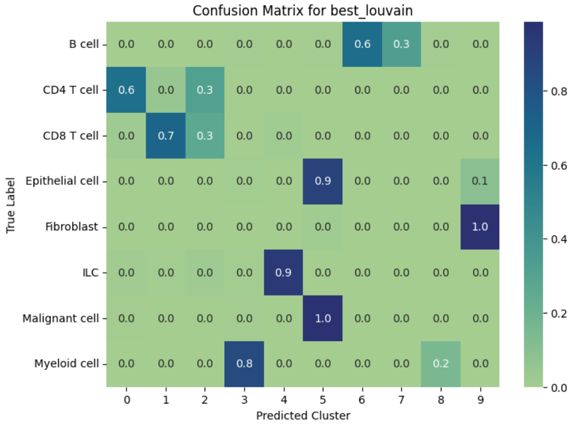

Figure 8. Louvain algorithm confusion matrix plotting "true label" and the predicted cell clusters.

The optimal Louvain model returned an ARI of 0.6161 and a confusion matrix that shows over-splitting along the rows, where the CD4 T-cells and CD8 T-cells have been incorrectly split into multiple predicted clusters. These are both improved upon in the Leiden algorithm, which builds upon the Louvain shortcoming of its tendency to produce inconsistent partitions, resulting in potentially enhanced results.

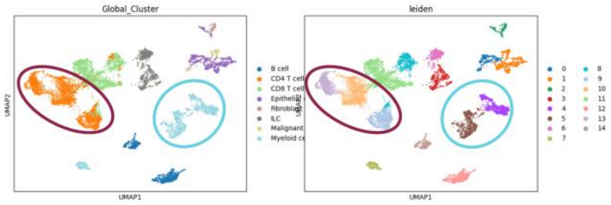

Figure 9. Example of over-clustering with the Leiden algorithm with a hyperparameter resolution of 0.5.

Prior to hyperparameter finetuning, a standard resolution of 0.5 is used to display over-clustering, as shown in Fig. 9. Both areas circled on the graph show examples of over-clustering, where the CD4 T-cells have been grouped into three separate clusters and the myeloid cells are grouped into two separate clusters. As discussed previously, utilizing Optuna to fine-tune the resolution of the Leiden algorithm will result in significantly fewer clusters.

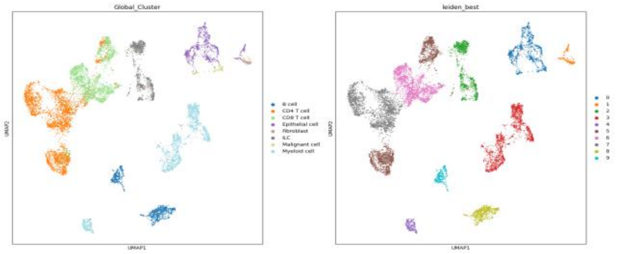

Figure 10. Leiden algorithm clustering with a resolution hyperparameter of 0.14, determined by Optuna study.

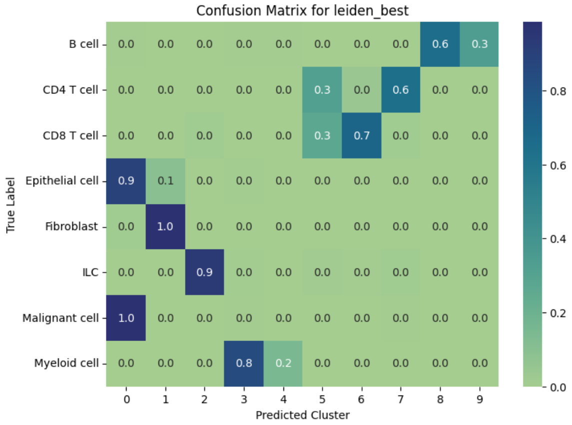

Figure 11. Leiden algorithm confusion matrix.

Fig. 10 and Fig. 11 show Leiden's improvements, where the number of clusters has been reduced and more appropriately fit. While the confusion matrix still displays over-splitting, it is less severe, resulting in an ARI of 0.6220.

3.2.3 DBSCAN, HDBSCAN, and Hierarchical

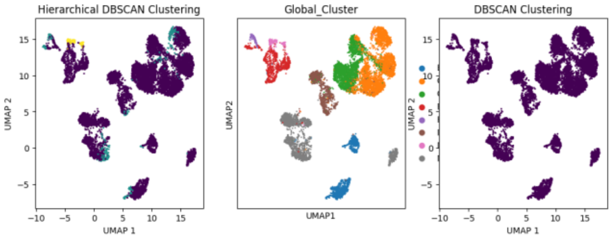

Figure 12. Provided UMAP plot colored by HDBSCAN, cell type, and DBSCAN cluster.

Both DBSCAN and HDBSCAN appear to form up to 3 clusters, which do not match the true labels. Hierarchical clustering produced a greater number of clusters but were similarly poor in quality. This is likely attributed to nuances within biological data that may not directly represent cell types, but potential subtypes of malignancies, immune, and cell types.

4. Conclusion

Dimensionality-reduction techniques and clustering methods were analyzed to determine which most effectively displays the true biological structure of CRC scRNA-seq data. Among the dimensionality-reduction techniques, UMAP produced the clearest separation of cell types, outperforming t-SNE and PCA. For the clustering algorithms, K-means, DBSCAN, and HDBSCAN failed to successfully reproduce the true label, either producing too few clusters or having poor alignment with the true labels. However, the graph-based Louvain and Leiden methods achieved significantly stronger cluster generation with Optuna-optimized hyperparameters, improving alignment to true labels, and reducing over-splitting. Leiden achieved the highest ARI, indicating an accurate match to the biological cell types and therefore providing the most reliable unsupervised identification of cell types within the SMART-seq dataset.

Future work will focus on expanding the range of dimensionality-reduction and clustering strategies used on the dataset. One direction involves exploring additional combinations of the three primary reduction techniques already used, evaluating whether sequential or hybrid approaches (e.g., PCA followed by t-SNE or UMAP) create embeddings with better structure preservation. More hyperparameter tuning for each method may also provide more stable, low-dimensional representations. Future dimensionality-reduction experiments could also include non-linear neural approaches. In particular, a variational autoencoder-based latent space could be trained to learn compact representations that preserve biologically relevant variation more effectively than the methods used so far.

Further refinement of the clustering stage is another possible area for development. Additional hyperparameter tuning for the clustering methods used may lead to better partitions of the underlying cell population. Techniques such as consensus clustering or cluster-stability analysis could also be incorporated to better analyze the reliability of different cluster assignments.

5. Workload and Distribution

Andrew Berard - Tested effectiveness of PCA and UMAP dimensionality reduction and compared all the clustering techniques to find the best performing model
Eli Fisk - Tested the effectiveness of t-SNE dimensionality reduction and compared all the clustering techniques with the t-SNE reduced data.
Angelina Ju - Visualized Louvain and Leiden clustering algorithms with UMAP, conducted Optuna study trials to fine-tune hyperparameters, and generated confusion matrices with ARI to demonstrate importance of results.
Sidarth Kumar – Tested RI vs ARI as accuracy metrics on K-means clustered data compared to ground truth labels. Interpreted project overview in abstract and introduction with the help of Jai.
Jai Mehta – Preprocessed the code. Developed the confusion matrix. Performed Hierarchical, DBSCAN, and HDBSCAN clustering. Interpreted the outputs and related to a biological background. Led the project.

References

[1] American Cancer Society. *Cancer Facts & Figures 2025.* Atlanta: American Cancer Society; 2025.

[2] Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: A brief overview. *Clin Transl Med*. (2022)

[3] Zhang L, Li Z, Skrzypczynska KM, Fang Q et al. Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. *Cell* (2020)

[4] Nadjafikhah, M., Nasiri, M. A comparative study of manifold learning methods for scRNA-seq with a trajectory-aware metric. *Sci Rep* **15**, 28923 (2025).

[5] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In KDD.

[6] Wolf, F., Angerer, P. & Theis, F. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 19, 15 (2018).