A Comparative Analysis of Clustering Gene Expression Data for Cancer Subtypes

Leila Michal
Ray and Stephanie Lane Computational Biology Department
Carnegie Mellon University
Pittsburgh, PA 15213
Andrew ID: lmichal

Abstract

Accurate identification of cancer subtypes is critical for improving patient outcomes, yet the complex, high-dimensional nature of gene expression data presents substantial analytical challenges. This study investigates how combinations of dimensionality reduction and clustering techniques can reveal biologically meaningful patterns in cancer transcriptomic profiles. Using the RNA-Seq (HiSeq) PANCAN dataset comprising 801 patients across five tumor types, Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) were applied to project the high-dimensional data into lower-dimensional spaces. Subsequently, K-means, hierarchical agglomerative clustering, and Leiden clustering were employed to uncover underlying subtype structures. Clustering performance was evaluated through internal metrics (Silhouette Score) and external validation (Adjusted Rand Index and Jaccard Index) against known tumor labels. The results demonstrate that nonlinear dimensionality reduction methods, particularly UMAP combined with classical clustering algorithms such as K-means and hierarchical clustering, achieve superior performance in recovering biologically coherent groups. Quantitative analysis revealed that UMAP combined with K-means achieved an ARI of 0.9925, a Jaccard Index of 0.9888, and a Silhouette Score of 0.8891, significantly outperforming PCA-based clustering. This work highlights the importance of method selection when analyzing complex genomic data and provides a reproducible computational framework for unsupervised cancer subtype discovery.

1 Introduction

Cancer is an inherently heterogeneous disease, characterized by a complex interplay of genetic, epigenetic, and transcriptomic variations that drive tumor initiation, progression, and response to therapy. Accurately categorizing tumors into biologically meaningful subtypes is essential, as distinct subtypes are often associated with different clinical outcomes and therapeutic strategies. Gene expression profiling, particularly through RNA sequencing (RNA-Seq), has enabled high-throughput measurement of transcriptomic activity across thousands of genes simultaneously, offering an unprecedented opportunity to understand the molecular basis of tumor heterogeneity.

Traditional approaches to cancer classification have relied on expert-curated gene panels or predefined biological pathways. While effective in many contexts, such methods inherently limit discovery, focusing analyses on known markers and potentially overlooking novel or emerging molecular patterns. In contrast, unsupervised clustering techniques, which group samples based solely on similarities in the data without reliance on prior knowledge, offer a powerful alternative for discovering hidden structures within gene expression profiles. These approaches have the potential to reveal previously unrecognized subtypes, providing new insights into cancer biology and identifying patient groups that may benefit from targeted therapies.

Despite their promise, clustering gene expression data presents significant analytical challenges. Beyond the well-known difficulties regarding dimensionality, additional factors such as technical noise from sequencing variability, batch effects arising from differences in sample preparation, and biological variability within and between tumor types can obscure meaningful patterns. Tumor samples often exhibit high intra-class heterogeneity and inter-class similarity, making it difficult to establish clean separations between biological groups. Furthermore, some molecular subtypes may exist along a continuum rather than as discrete clusters, further challenging conventional clustering algorithms. Successfully overcoming these obstacles requires a sophisticated computational approach capable of reducing dimensionality while preserving both local and global biological structures, followed by robust clustering validated through both internal and external metrics.

This project evaluates the performance of several dimensionality reduction and clustering method combinations for the task of identifying cancer subtypes from RNA-Seq data. Dimensionality reduction was performed using Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP), each offering different strengths in preserving variance, local structure, and global relationships. Clustering was subsequently applied using K-means, hierarchical agglomerative clustering, and Leiden community detection methods. The quality of the resulting clusters was assessed through internal validation using the Silhouette Score and external validation against known tumor labels using the Adjusted Rand Index (ARI) and Jaccard Index. By systematically analyzing these combinations, this project aims to identify computational strategies that most effectively recover biologically significant cancer subtypes, providing a reproducible foundation for future studies in precision oncology [10, 4].

2 Methods

2.1 Data Description

The dataset utilized in this study is drawn from the RNA-Seq (HiSeq) PANCAN collection, provided by the UC Irvine Machine Learning Repository. Curated by Fiorini (2016) [8], this dataset captures gene expression profiles for 801 patients diagnosed with one of five cancer types: breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC), colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), and prostate adenocarcinoma (PRAD). Each sample is represented by 20,531 real-valued features corresponding to gene expression levels measured via the Illumina HiSeq platform.

Samples are organized in a row-wise fashion, where each row corresponds to a patient and each column to a gene. Although no missing values were officially reported, preprocessing included a precautionary step using scikit-learn's SimpleImputer to ensure robustness against potential anomalies. To enhance the discriminative power of the dataset and improve clustering performance, genes with low variance (variance < 0.1) across samples were removed, resulting in a refined matrix with 20,264 genes. Tumor labels indicating the known cancer types were preserved separately and only used during evaluation, maintaining the unsupervised nature of the clustering task. The choice of this dataset was deliberate, as it embodies both technical noise and biological heterogeneity typical of real-world clinical genomic data, offering a stringent test for the clustering methodologies applied.

2.2 Computational Approach

The computational workflow for this project was structured into four stages: Data Preprocessing, Dimensionality Reduction, Clustering, and Evaluation.

2.2.1 Data Preprocessing

Gene expression data were first standardized using scikit-learn's StandardScaler, transforming each feature to have a mean of zero and a standard deviation of one. Standardization was chosen over normalization to a [0,1] range to preserve the natural variance structure of the data, a critical property for methods such as Principal Component Analysis (PCA) and K-means clustering, which are sensitive to differences in feature scales. Without this step, highly expressed genes could dominate the analysis, masking important but subtler patterns.

2.2.2 Dimensionality Reduction

Following preprocessing, three-dimensionality reduction techniques were applied independently to the standardized data set. Principal Component Analysis (PCA) was used to linearly project the data onto orthogonal axes that maximize variance. Fifty principal components were retained, chosen to balance the need to capture sufficient variation without reintroducing excessive dimensionality, with the first ten components accounting for approximately 45.9% of the total variance. t-distributed Stochastic Neighbor Embedding (t-SNE) was then employed to create a two-dimensional embedding that preserves local similarities between samples. A perplexity value of 30 was selected to balance the preservation of local neighborhood relationships and global data structure, which is appropriate given the size of the dataset of several hundred instances. Lastly, Uniform Manifold Approximation and Projection (UMAP) was used to generate another two-dimensional embedding, with the number of neighbors set to 15 and the minimum distance between points set to 0.1. These parameters were chosen to prioritize the maintenance of fine-grained local structures while allowing flexibility in cluster separations, both properties critical for uncovering meaningful biological subtypes [1, 5].

2.2.3 Clustering Methods

Clustering was subsequently performed independently on each low-dimensional embedding. K-means clustering was implemented with five clusters, corresponding to the known number of tumor types, and initialized with ten random

seeds to improve robustness against poor centroid initialization. Hierarchical agglomerative clustering was performed using Ward's linkage and Euclidean distance, which favor compact, spherical cluster formation—an appropriate assumption after dimensionality reduction. Additionally, Leiden clustering was applied to graphs constructed from UMAP embeddings, optimizing modularity to detect community structures [9, 11]. While Leiden clustering has the capacity to uncover multi-scale biological relationships, it also introduces a risk of overfragmentation, necessitating careful parameter monitoring.

The combination of centroid-based (K-means), agglomerative (Hierarchical), and graph-based (Leiden) clustering methods allowed for a comprehensive evaluation of how different algorithmic assumptions impact cancer subtype discovery when applied to diverse lower-dimensional representations of the gene expression space.

2.2.4 Evaluation Metrics

To assess the quality of clustering, both internal and external validation metrics were employed. The Silhouette Score [7] was used as the primary internal validation measure, evaluating the cohesion within clusters and separation between clusters without reference to external labels. Silhouette values range from -1 to 1, with higher scores indicating more distinct and well-separated clusters.

External validation was performed using the Adjusted Rand Index (ARI) and the Jaccard Index. The ARI measures the agreement between predicted cluster labels and true tumor type labels, adjusting for random chance. An ARI close to 1 indicates near-perfect agreement, while an ARI of 0 corresponds to random assignment. The Jaccard Index evaluates the proportion of correctly grouped sample pairs relative to the union of predicted and true groupings, offering another perspective on clustering agreement. Both external metrics provided a biologically grounded evaluation of how well the unsupervised clustering algorithms recovered meaningful tumor subtypes.

Using these metrics in combination ensured that clustering performance was comprehensively assessed from both statistical and biological perspectives, critical for evaluating method effectiveness in real-world biomedical applications.

3 Implementation Details

All computational analyses were performed using Python 3.10. The primary libraries utilized included `scikit-learn` for standard machine learning algorithms such as Principal Component Analysis (PCA), K-means clustering, and hierarchical agglomerative clustering; `umap-learn` for Uniform Manifold Approximation and Projection (UMAP); and `leidenalg` for graph-based Leiden clustering, implemented using the `igraph` package. Additional preprocessing and numerical operations were conducted using `pandas` and `NumPy`, while all visualizations were generated using `matplotlib` and `seaborn`.

Hyperparameters for each method were selected through exploratory testing, with values chosen based on empirical performance and visual inspection of clustering results. For dimensionality reduction, fifty components were retained in PCA to balance information retention and computational efficiency. t-SNE was configured with a perplexity of 30 to reflect the expected local neighborhood size for the dataset of 801 samples. UMAP was applied with 15 neighbors and a minimum distance of 0.1 to preserve fine-grained local structure while allowing moderate flexibility in global cluster separation. For clustering, the number of clusters for both K-means and hierarchical clustering was set to $k = 5$, matching the known number of cancer types. K-means was initialized with 10 random seeds to ensure robustness. Leiden clustering was applied to a k-nearest neighbor graph constructed from UMAP embeddings, with community detection driven by modularity optimization.

The full computational pipeline was designed to be modular and reproducible, with each stage—from data preprocessing, dimensionality reduction, and clustering to evaluation—implemented independently and in an easily configurable manner, as summarized below.

1. **Preprocessing:** Gene expression data were log-transformed using `NumPy`'s `log1p` function to reduce skewness. Standardization was performed using `StandardScaler` from `scikit-learn`, centering each gene to mean zero and scaling to unit variance. Low-variance genes (variance < 0.1) were filtered using `NumPy`.
2. **Dimensionality Reduction:** Applied independently using:
   - **PCA:** Implemented via `sklearn.decomposition.PCA`, retaining 50 components.
   - **t-SNE:** Using `sklearn.manifold.TSNE` with 2 components and perplexity = 30.
   - **UMAP:** Applied via `umap.UMAP` with 2 components, $n\_neighbors = 15$, and $min\_dist = 0.1$.
3. **Clustering:** Performed on each embedding using:
   - **K-means:** Via `sklearn.cluster.KMeans` with $k = 5$ and $n\_init = 10$.
   - **Hierarchical Clustering:** Using `AgglomerativeClustering` with Ward linkage.
   - **Leiden Clustering:** Using `leidenalg` on graphs generated from UMAP embeddings.

4. **Evaluation:** Results were assessed using:
   - **Silhouette Score:** From `sklearn.metrics`, for internal cluster validity.
   - **Adjusted Rand Index (ARI):** From `sklearn.metrics`, to compare with ground truth labels.
   - **Jaccard Index:** Implemented using set operations to measure overlap between predicted and true clusters.

By clearly specifying all software tools, hyperparameter selections, and procedural steps, this implementation ensures that results are reproducible, interpretable, and directly extendable to future transcriptomic clustering studies.

4 Results

Dimensionality reduction revealed substantial differences in the structure and separability of the gene expression data depending on the method applied. Principal Component Analysis (PCA) reduced the data to fifty principal components, with the first ten components capturing approximately 45.9% of the total variance. These observations are illustrated in Figure 1. However, the resulting PCA embedding exhibited diffuse sample distributions without clear cluster separations, suggesting that linear projections alone are insufficient to capture the complex biological structures underlying cancer subtypes.

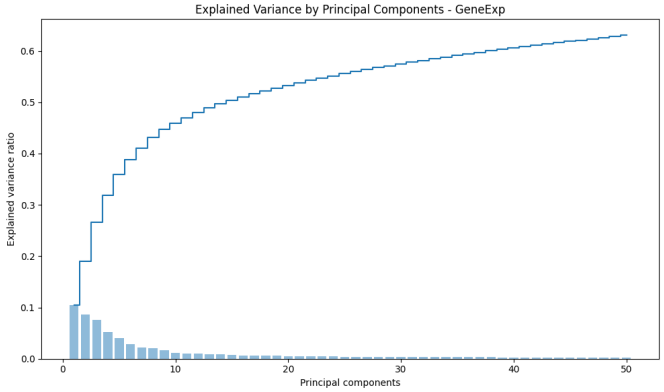

Figure 1: PCA explained variance plot showing the first ten components capture 45.9% of variance.

Visualization of low-dimensional embeddings (Figure 2) further emphasized the differences between dimensionality reduction techniques. PCA produced a continuous elongated cloud of samples with minimal visible clustering, with the exception of some separation between samples from kidney cancer. In contrast, t-SNE and UMAP embeddings displayed well-defined compact clusters. UMAP, in particular, preserved both local neighborhood integrity and global topology, producing separable clusters with minimal overlap, while t-SNE generated dense local groupings but distorted global distances. These observations indicate that non-linear dimensionality reduction methods are better suited to uncover complex, nonlinear biological relationships in gene expression profiles.

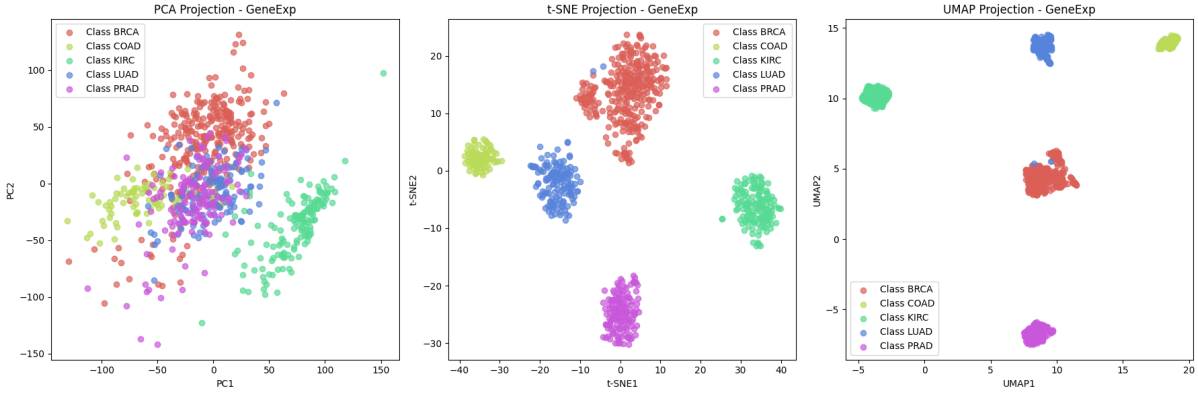

Figure 2: Comparison of PCA, t-SNE, and UMAP embeddings, with UMAP revealing the most distinct clusters.

The application of K-means clustering to the PCA, t-SNE, and UMAP embeddings highlights the dependence of clustering success on the quality of the data projection. Clustering on the PCA-reduced data resulted in overlapping, diffuse clusters with substantial intermixing of samples, achieving an Adjusted Rand Index (ARI) of 0.7979, a Jaccard

Index of 0.7241, and a Silhouette Score of 0.2462. When K-means was applied to t-SNE embeddings, five tight, well-separated clusters emerged, closely matching the known tumor subtype structure, with improved ARI of 0.9925 and Silhouette Score of 0.7341. UMAP embeddings further enhanced cluster separability, achieving an ARI of 0.9925, a Jaccard Index of 0.9888, and a Silhouette Score of 0.8891, indicating near-perfect recovery of biological subtypes. These findings are shown in Figure 3.

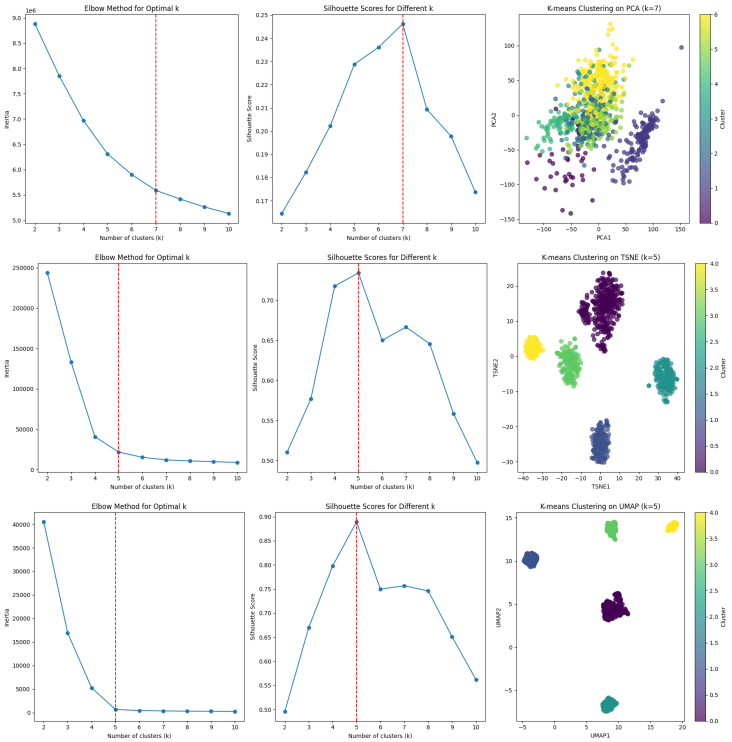

Figure 3: K-means clustering results for PCA, t-SNE, and UMAP embeddings.

Hierarchical agglomerative clustering produced results consistent with those observed for K-means clustering. Clustering based on PCA embeddings yielded less distinct groupings, with a Silhouette Score of 0.2457 and moderate overlap between subtypes. Clustering on t-SNE embeddings showed improved cluster separation, while UMAP embeddings once again yielded the clearest structure, with a Silhouette Score of 0.7341. These trends are visualized in Figure 4.

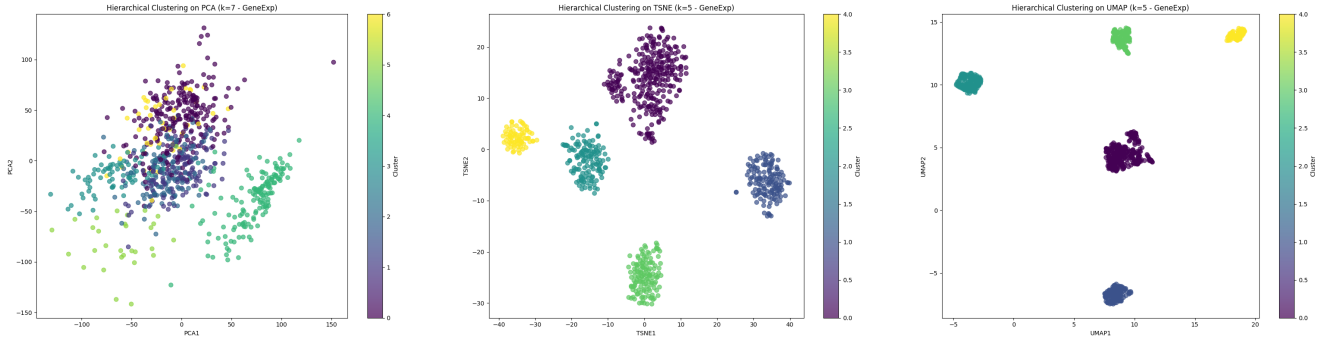

Figure 4: Hierarchical clustering results on PCA, t-SNE, and UMAP embeddings. UMAP clustering achieved the highest Silhouette Score (0.8891).

Following the application of hierarchical clustering, dendrograms were constructed to visualize the nested relationships between samples after dimensionality reduction. The dendrogram generated from PCA embeddings revealed shallow and indistinct branch separations, reflecting the limited ability of linear projections to fully separate biological subtypes. In contrast, dendrograms derived from t-SNE and UMAP embeddings exhibited deeper, well-separated branches, corresponding to distinct tumor subtypes. The UMAP-based dendrogram was particularly informative, showing long, distinct branches that grouped samples into tight, cohesive clusters. This structure indicates that UMAP not only preserved local neighborhood relationships but also enabled hierarchical algorithms to detect meaningful multilevel biological structures. These patterns are visualized in Figure 5.

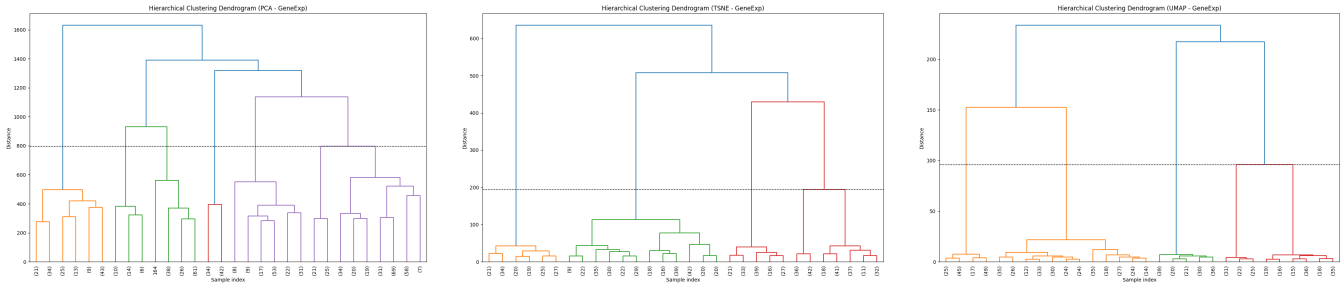

Figure 5: Hierarchical clustering dendrograms constructed from PCA, t-SNE, and UMAP embeddings. UMAP dendrogram exhibits deeper branch separations corresponding to biologically coherent tumor subtypes.

Leiden clustering provided an alternative, graph-based perspective for uncovering community structure in the gene expression data. When applied to PCA embeddings, Leiden clustering at a resolution of 0.5 produced six clusters with a relatively low silhouette score of 0.24, and the resulting cluster boundaries remained diffuse, with considerable overlap among biological classes. On t-SNE embeddings, the method yielded seven clusters with a significantly higher silhouette score of approximately 0.66 at the same resolution, reflecting improved intra-cluster cohesion and inter-cluster separation. However, UMAP embeddings exhibited the highest silhouette score of approximately 0.76 at a lower resolution of 0.2, generating seven compact clusters with excellent visual separability. Despite these improvements, Leiden clustering tended to fragment known tumor types into overly fine subdivisions, particularly at higher resolutions, as evidenced by the steep increase in the number of clusters beyond the optimal point. This overpartitioning, while useful for exploratory subgroup discovery, complicates interpretation when coherent, clinically relevant subtype labels are desired. These observations are summarized in Figure 6.

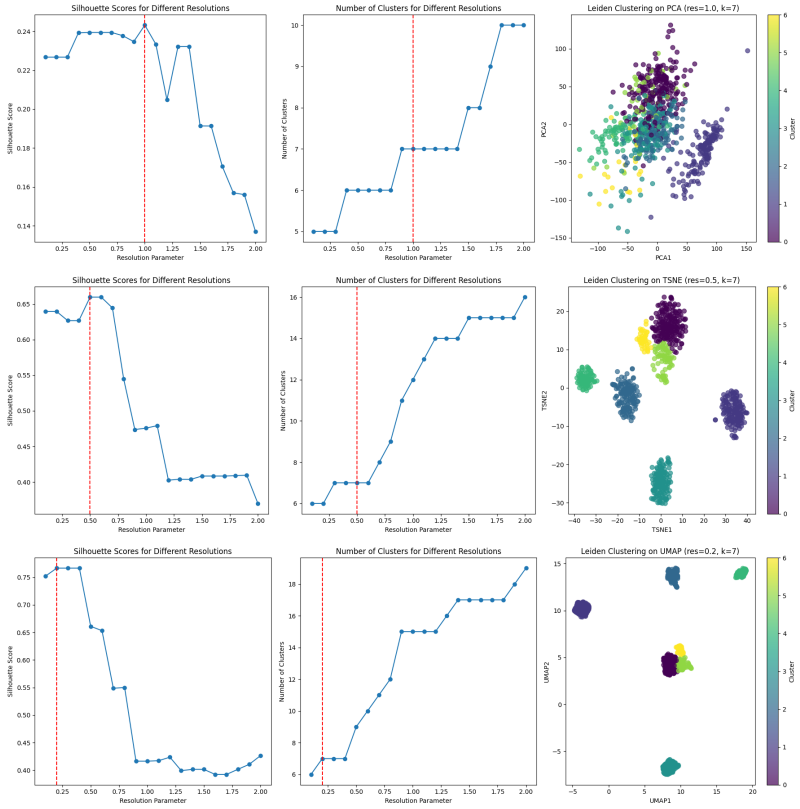

Figure 6: Leiden clustering results showing community detection but overfragmentation on t-SNE and UMAP embeddings.

Quantitative evaluation using internal and external metrics confirmed the superiority of nonlinear dimensionality reduction methods. As summarized in Figure 7, UMAP combined with K-means clustering achieved the highest overall scores, with ARI of 0.9925, Jaccard Index of 0.9888, and Silhouette Score of 0.8891. t-SNE embeddings also achieved strong clustering results, while PCA embeddings consistently underperformed across all metrics.

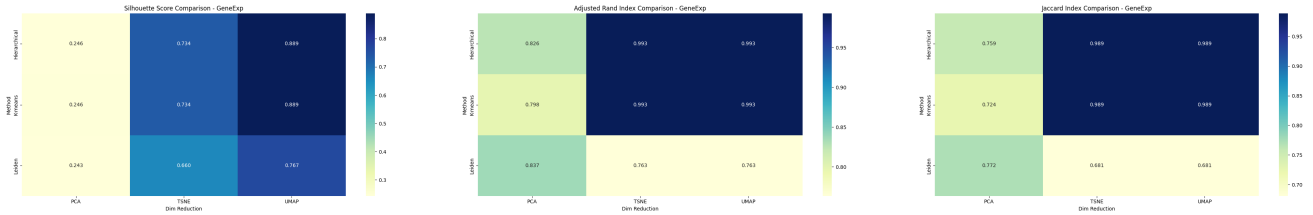

Figure 7: Heatmaps summarizing Silhouette, ARI, and Jaccard scores across all clustering approaches.

Taken together, these results strongly demonstrate that nonlinear dimensionality reduction techniques, particularly UMAP, combined with simple clustering algorithms such as K-means and hierarchical clustering, offer the most robust and biologically coherent recovery of cancer subtypes.

5 Discussion and Conclusions

The findings of this study highlights the crucial role of dimensionality reduction and clustering method selection in the analysis of high-dimensional gene expression data for cancer subtype discovery. The results demonstrated that while Principal Component Analysis (PCA) is effective for capturing large variance components, it falls short in revealing the nonlinear, complex structures underlying transcriptomic profiles. Nonlinear dimensionality reduction methods, particularly Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbor Embedding (t-SNE), preserved both local and global relationships more effectively, enabling downstream clustering algorithms to recover biologically coherent subtypes with higher fidelity.

The choice of clustering method was equally consequential. When applied to embeddings produced by nonlinear dimensionality reduction, traditional distance-based clustering algorithms such as K-means and hierarchical agglomerative clustering consistently outperformed graph-based Leiden clustering. K-means clustering on UMAP embeddings achieved the best overall performance, with an Adjusted Rand Index (ARI) of 0.9925, a Jaccard Index of 0.9888, and a Silhouette Score of 0.8891. These metrics indicate near-perfect recovery of the known tumor subtypes, underscoring the strength of combining nonlinear embedding with simple, interpretable clustering techniques. Hierarchical clustering similarly benefited from UMAP embeddings, producing clear, nested cluster structures as evidenced by dendrogram analysis.

While Leiden clustering exhibited strong modularity optimization scores, particularly on UMAP embeddings, it frequently overpartitioned the data into numerous small communities. This overfragmentation, although useful for exploratory analyses seeking fine-grained structures, complicated the biological interpretation of major tumor subtypes. These results highlight an important tradeoff in clustering cancer genomics data: fine-grained community detection must be balanced against the need for interpretable, biologically meaningful groupings.

An important insight from this study is that dimensionality reduction should not be viewed merely as a visualization tool, but as a fundamental determinant of clustering success in high-dimensional spaces. Without effective dimensionality reduction, even robust clustering algorithms struggle under the curse of dimensionality, leading to poor cluster separation and biological misclassification. Moreover, the use of nonlinear techniques like UMAP enables clustering algorithms to operate on manifolds that more faithfully reflect the underlying biological structures.

Despite the strong performance observed, the project has limitations. Clustering results were sensitive to hyperparameter settings in both dimensionality reduction and clustering algorithms. For instance, changes in UMAP's number of neighbors or t-SNE's perplexity could materially affect cluster tightness and separability. Additionally, external validation relied on known tumor type labels, which, while clinically validated, may obscure the discovery of novel or finer-grained biological subtypes. The reliance on ground truth labels also assumes discrete subtype boundaries, whereas cancer biology often presents continuum-like transitions between molecular states.

Future work should address these limitations through systematic hyperparameter sweeps to evaluate robustness, bootstrapped clustering across random seeds to assess stability, and integration with additional omics data such as copy number variation or methylation profiles to improve biological interpretability. Matrix factorization methods like non-negative matrix factorization (NMF) have also been widely used for cancer subtype discovery [3], and although not explored in this study, they may provide complementary perspectives worth comparing in future analyses. Evaluation frameworks could additionally be expanded to include survival analysis or treatment response association studies, providing deeper insights into clinical relevance.

In conclusion, this project successfully achieved its objective of identifying effective combinations of dimensionality reduction and clustering methods for the unsupervised discovery of cancer subtypes from high-dimensional gene expression data. The results clearly demonstrate that nonlinear dimensionality reduction, particularly UMAP, when combined with classical clustering algorithms like K-means and hierarchical clustering, offers a powerful and

reproducible computational framework for cancer genomics. These findings emphasize that careful method selection, coupled with a deep understanding of the data structure, is essential for advancing precision oncology through unsupervised learning approaches.

References

[1] E. Becht, L. McInnes, J. Healy, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–44, 2019.

[2] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297, 1967.

[3] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.

[4] J. Wang, Z. Zhang, L. Liu, et al. Clustering methods in cancer gene expression data. *Briefings in Functional Genomics and Proteomics*, 7(4):332–338, 2008.

[5] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[6] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

[7] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[8] S. Fiorini. Gene Expression Cancer RNA-Seq. *UCI Machine Learning Repository*, 2016. Available: `https://doi.org/10.24432/C5R88H`.

[9] S. Freytag, T. Tian, S. L. Lonnstedt, M. Bahlo. Comparison of clustering tools in R for medium-sized 10x single-cell RNA-sequencing data. *F1000Research*, 7:1297, 2018.

[10] S. Ramaswamy, P. Tamayo, R. Rifkin, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154, 2001.

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