BIOMEDICAL NETWORK SCIENCE PROJECT Exploring the Impact of Hyper-parameter Variation in Single Cell Analysis

1st Ismam Hussain Khan
Department of Artificial Intelligence
Friedrich Alexandar University of Erlangen Nuremberg
Erlangen, Germany
ishmam.hussain.khan07@gmail.com

***Abstract*—Single-cell RNA sequencing (scRNA-seq) has emerged as a powerful tool for dissecting cellular heterogeneity, but the analysis is heavily influenced by hyperparameter selection. This study presents a comprehensive investigation into the effects of varying two critical hyperparameters: the number of neighbors and the number of principal components, on the performance of single-cell classification models across three datasets - diabetesII-h1, mpn-H108, and pbmc3k . An automated pipeline was developed for preprocessing, clustering using the Leiden algorithm, cell type annotation via SCSA, evaluation metric calculation and summary table generations for all the dataset. The pbmc3k dataset consistently exhibited the highest performance metrics, suggesting more distinct cell populations in non-diseased states. Conversely, the disease datasets (diabetesII-h1 and mpn-H108) displayed lower metrics, potentially due to increased cellular heterogeneity. Notably, an intermediate number of neighbors tended to balance capturing local structures and generalization for improved performance. The mpn-H108 dataset exhibited higher precision compared to diabetesII-h1 when varying principal components, suggesting better identification of true positive cell types associated with MPNs. The high performance on pbmc3k indicates robust classification of healthy cell populations, while lower metrics for disease datasets highlight challenges in increased complexity and heterogeneity. This study underscores the importance of hyperparameter tuning in scRNA-seq analysis and provides insights into optimizing parameters across disease contexts to enhance accuracy, reproducibility through the pipeline and interpretability.**
Github Link of repository: https://github.com/ishmam367/Single-Cell-Analysis

***Index Terms*—Single Cell Analysis, Cell type annotation, Hyper-parameter selection**

I. Introduction

The central dogma of molecular biology, first proposed by Francis Crick in 1958, describes the fundamental flow of genetic information within biological systems. It states that the information stored in DNA is transcribed into messenger RNA (mRNA), which is then translated into proteins, the functional molecules that drive cellular processes. This unidirectional transfer of information, from DNA to RNA to proteins, underpins the intricate mechanisms governing gene expression and regulation. [2].

Single cell segmentation is a crucial technique in biomedical image analysis that aims to delineate individual cells within microscopic images. It plays a vital role in various applications, such as cell tracking, cell counting, and quantitative analysis of cellular morphology and behavior. The accurate segmentation of individual cells is a challenging task due to factors like cell clustering, variations in cell shape and size, and imaging artifacts. Effective single cell segmentation algorithms leverage techniques from image processing, machine learning, and pattern recognition to address these challenges. The development of robust and accurate segmentation methods is crucial for advancing our understanding of cellular processes and enabling quantitative analysis in fields like developmental biology, cancer research, and regenerative medicine. [5].

The standard Scanpy workflow only works with one dataset and preselected hyperparameters which are finely tuned to work on that dataset. However, it does not give a clear idea about the impact of varying hyperparameters. Also, since there is no annotation provided for the cell types, it is hard to find any meaningful outcome or to use any evaluation metric on them for comparison and analysis. Thus, the problem addressed in this work is the exploration of the impact of varying hyperparameters, specifically the number of principal components and the number of neighbors, on the clustering and cell type annotation of scRNA-seq data. These hyperparameters play a pivotal role in dimensionality reduction and neighborhood graph construction, respectively, which can significantly influence the downstream analysis and interpretation of the data. A key task achieved here is the development of an automated pipeline for preprocessing, clustering, and cell type annotation of scRNA-seq data across multiple datasets, enabling efficient and reproducible analysis and comparison among these datasets. By addressing these objectives, this work aims to contribute to the advancement of scRNA-seq data analysis methodologies, enabling researchers to make informed decisions regarding hyperparameter selection and enhancing the interpretability and reliability of downstream analyses.

II. Dataset

In total there are 3 diffferent datasets from 2 different resources. Two data sources of two diseases, such as type

II Diabetes Mellitus and Myeloproliferative Neoplasm (MPN) have been obtained from the Gene Expression Omnibus repository.Another dataset consist of 3k Peripheral Blood Mononuclear Cells (PBMCs) from a Healthy Donor and are freely available from 10x Genomics.Here is the download Link . From the link download the 'GENE/cell matrix filtered'. Details of the dataset are given below,

A. *Peripheral Blood Mononuclear Cells (PBMCs)*

The dataset comprises 3,000 peripheral blood mononuclear cells (PBMCs) from a healthy donor, sequenced using Cell Ranger 1.1.0. Each cell contains approximately 1pg RNA. A total of 2,700 cells were successfully detected, sequenced on Illumina NextSeq 500 with an average of 69,000 reads per cell. The sequencing setup includes a 98bp transcript read, 8bp I5 sample barcode, 14bp I7 GemCode barcode, and 10bp UMI read. The analysis was conducted with a focus on 3,000 cells. Here is the source-url

B. *Diabetes Mellitus Type II*

This dataset examines the systemic immunological changes induced by type 2 diabetes mellitus (DM) in individuals diagnosed with periodontitis (PD). Utilizing single-cell RNA sequencing (scRNA-seq) analysis of peripheral blood mononuclear cells (PBMCs), the study aims to contrast the immune response in patients with PD alone versus those with both PD and DM (PDDM). By comparing these groups, researchers aim to deepen the understanding of the intricate immunological interplay between PD and DM. The sample distribution consists of 11 healthy control subjects, 10 PD patients without DM, and 6 patients diagnosed with PDDM. Here is the source-url

C. *Myeloproliferative Neoplasm (MPN)*

This dataset presents an in-depth examination of platelets obtained from patients diagnosed with myeloproliferative neoplasms (MPNs), with a specific focus on essential thrombocythemia (ET). Carried out by researchers, the study unveiled noteworthy metabolic changes influencing abnormal platelet function and inflammation in MPNs, employing single-cell RNA sequencing (scRNA-Seq) analysis of primary PBMC samples. Particularly, transcripts associated with platelet activation, mTOR, and oxidative phosphorylation (OXPHOS) were observed to be heightened in ET platelets. Here is the source-url.

III. Methodology

In this project, an automated pipeline to facilitate the analysis of single-cell data was implemented. The pipeline was implemented using Python programming language and relied on widely-used libraries such as Scanpy and Anndata. Custom scripts were written to orchestrate the preprocessing, clustering, cell-type annotation, measuring evaluation metrics to compare the impact of hyperparameters and finally, generating summary results to compare the outcome of the 3 different datasets side by side. The pipeline architecture was designed to ensure modularity and scalability, enabling seamless integration of additional datasets and analysis techniques.

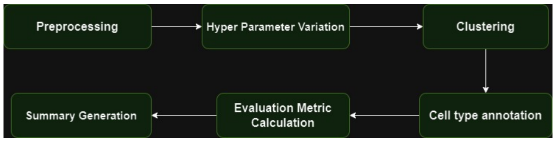

Fig. 1. Pipeline Framework for Single-Cell Data Analysis and Evaluation

A. *Preprocessing*

In preprocessing the raw scRNA-seq data, we implemented a series of quality control measures to filter out low-quality cells based on mitochondrial gene expression, UMIs, and detected genes. Subsequently, we applied logarithmic transformation and Pearson correlation-based normalization to mitigate technical biases and ensure comparability across cells. We then selected highly variable genes to prioritize those contributing most significantly to cellular heterogeneity. Finally, Principal Component Analysis (PCA) was employed for dimensionality reduction, transforming the high-dimensional gene expression space into a lower-dimensional subspace while retaining maximal variance. These preprocessing steps collectively aim to enhance the accuracy and robustness of downstream analyses by addressing technical variability and capturing the underlying biological signal present in the scRNA-seq data.

B. *Hyper-Parameter Variation*

To explore the impact of hyper-parameter variation on single-cell analysis outcomes, we designed a systematic experiment. Specifically, we varied two key hyper-parameters: the number of principal components and the number of neighbors while keeping the other parameters the same. These hyper-parameters were selected based on their significance in single-cell analysis and their potential influence on clustering and cell-type annotation results. The experiment was structured to vary each hyper-parameter within a predefined range, allowing for comprehensive analysis of parameter sensitivity.

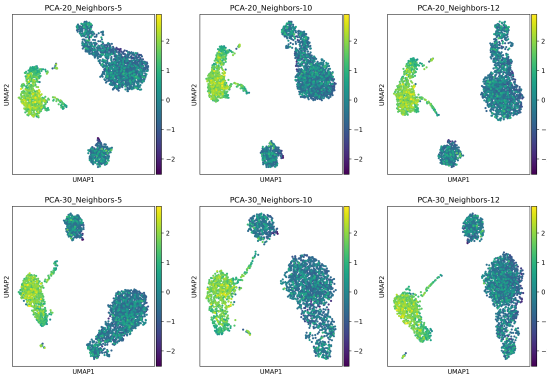

Fig. 2. Impact of Hyper-Parameter variation on spatial clustering

C. Clustering

For clustering the single-cell data, we employed the Leiden algorithm, a state-of-the-art method for identifying clusters in large-scale networks. The Leiden algorithm operates on a neighborhood graph constructed from the high-dimensional single-cell data, where cells are represented as nodes, and edges connect cells based on their similarity in gene expression or protein levels [5].This approach effectively captures the underlying structure and heterogeneity within the dataset. The neighborhood graph was constructed using the UMAP (Uniform Manifold Approximation and Projection) technique, which projects the high-dimensional data into a lower-dimensional space while preserving local and global data structure. [11]. UMAP has been widely adopted in single-cell analysis due to its ability to capture both local and global structures, making it well-suited for visualizing and exploring complex single-cell datasets [6]. Once the neighborhood graph was constructed, the Leiden algorithm was applied to identify clusters of cells with similar expression profiles or phenotypes. This algorithm optimizes for both cluster tightness and cluster separation, ensuring that the identified clusters are well-separated and internally coherent [3]. The resulting clusters were then visualized using UMAP plots, which provide a two-dimensional representation of the high-dimensional data . These UMAP embeddings allow for intuitive exploration of the clustered cell populations, enabling the identification of distinct cell types, rare cell subpopulations, and potential transitional states [11].

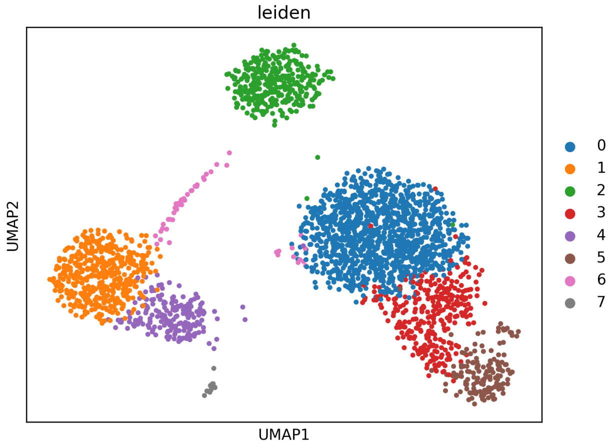

Fig. 3. Clustering with Leiden Algorithm

D. Cell Type Annotation

For annotating cell types in our single-cell dataset, we employed the SCSA (Single-Cell Scoring Assignment) method, an automated approach that leverages established and user-defined cell markers to assign cell type labels [5]. SCSA utilizes a scoring model that integrates differentially expressed genes (DEGs) and confidence levels of cell markers from various sources, providing a robust and accurate annotation strategy. The SCSA scoring model assigns a score to each cell for a given cell type based on the expression levels of known marker genes [5]. It incorporates information from established cell atlases, literature-curated gene sets, and user-provided marker genes, allowing for a comprehensive and flexible annotation process. The model considers both the differential expression of marker genes and the confidence levels associated with each marker, ensuring that the most reliable and informative markers contribute more significantly to the final score [11]. Compared to manual annotation methods, which can be time-consuming, subjective, and prone to inconsistencies, the automated SCSA approach offers several advantages [11]:

1) **Consistency:** SCSA applies a standardized and reproducible scoring algorithm, ensuring consistent annotation across different datasets and experiments.
2) **Accuracy:** By leveraging established cell atlases and literature-curated marker genes, SCSA can accurately annotate cell types, even in complex datasets with rare or novel cell populations.
3) **Scalability:** The automated nature of SCSA allows for efficient annotation of large-scale single-cell datasets, which would be impractical with manual methods.
4) **Flexibility:** SCSA can incorporate user-defined marker genes, enabling customization and adaptation to specific research contexts or experimental conditions.

The use of SCSA for cell type annotation not only streamlines the analysis process but also enhances the reliability and reproducibility of the results, facilitating downstream analyses and biological interpretations [8].

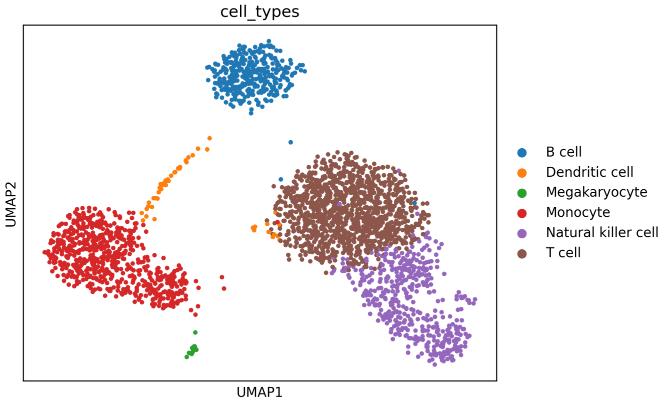

Fig. 4. Clustering after SCSA cell type annotation

E. Evaluation Metric Calculation

After finding the cell type annotation for each subset of PCA and Number of neighbours, we take the cell type annotation of PCA-40 and Neighbour-10 as the golden standard as these were the value used in the standard Scanpy tutorial. Here is a brief description of scanpy

Scanpy is a popular open-source Python library for analyzing single-cell gene expression data. It provides a scalable and efficient workflow for processing, analyzing, and visualizing large-scale single-cell datasets. Scanpy is built on top of powerful libraries like NumPy, SciPy, and Matplotlib, enabling efficient computation and visualization of large datasets . It also integrates with other popular single-cell analysis tools like Seurat (R) and Cellranger, facilitating interoperability and cross-platform analysis [10].
After deciding the golden standard, its cell type annotation is compared with other sets of hyper-parameters cell type annotation to determine evaluation metric such as F1-Score, Precision and Accuracy. At first I keep the value of PCA fixed to 40 and vary the number of neighbours to observe the change of evaluation metrics. Then I keep the number of neighbours fixed to 10 and change the PCA value to observe the change of results. Finally, visualization showing the impact of changing hyper-parameters on evaluations metrics is automatically created for each dataset.

F. Summary Generation

In our study, a comprehensive method was devised for summarizing the evaluation results of each set of hyper-parameters across multiple datasets. This process involved the following key steps:

- **Data Collection:** Evaluation results were collected from various datasets used in our experiments. Each dataset contained evaluation metrics such as accuracy, precision, and F1 score for different combinations of model hyperparameters. They were stored in pickle files for downstream analysis
- **DataFrame Construction:** A pandas DataFrame was constructed to organize these evaluation metrics systematically. Each row in the DataFrame represented a dataset, and each column represented a specific evaluation metric associated with a particular combination of model hyperparameters.
- **Heatmap Visualization:** To provide a visual representation of the evaluation results, a heatmap visualization technique was employed. Heatmap visualization is a powerful technique for representing tabular data in a compact and visually intuitive manner [7]. It employs color gradients to encode numerical values, enabling the identification of patterns and trends across multiple dimensions [7]. Heatmaps offer an effective way to display and compare complex data, making them particularly useful in various domains, including bioinformatics, genomics, and data analysis. In the heatmap, datasets were listed along the y-axis, while evaluation metrics and model hyperparameter combinations were displayed along the x-axis. Each cell in the heatmap contained the corresponding evaluation metric value, allowing for quick interpretation of model performance across different datasets and hyperparameter settings.
- **Exporting Results:** Finally, the generated heatmap was exported as an image file, allowing for seamless integration into research reports, presentations, and publications.

IV. Result and Analysis

The performance of our single-cell classification model was evaluated across three datasets: diabetesII-h1, mpn-H108, and pbmc3k. Two key hyperparameters, the number of neighbors (NB) and the number of principal components (PCA), were varied to assess their impact on the evaluation metrics: Accuracy, Precision, and F1 Score. The results are summarized in the heatmaps shown in figures 5 and 6.

Impact of Varying Number of Neighbors (NB) Figure 5 presents the performance metrics across different datasets and neighbor values (NB6, NB8, NB9, NB11, NB12, NB14, NB16) while keeping the number of principal components fixed at PCA40. Several key trends can be observed: The pbmc3k dataset consistently exhibits the highest performance metrics across all neighbor values, indicating robust classification results for this dataset derived from a healthy donor [4]. The diabetesII-h1 dataset shows moderate performance metrics, reflecting the model's ability to distinguish cell types affected by type II diabetes mellitus with reasonable accuracy. The mpn-H108 dataset generally displays lower performance metrics compared to the other two datasets, highlighting the challenges in classifying cell types in the context of myeloproliferative neoplasm [1]. Across all datasets, varying the number of neighbors affects the performance metrics, underscoring the importance of optimizing this parameter [1]. An intermediate number of neighbors (e.g., NB08, NB09, NB11, NB12) tends to balance capturing local data structures with maintaining generalization, leading to improved performance.

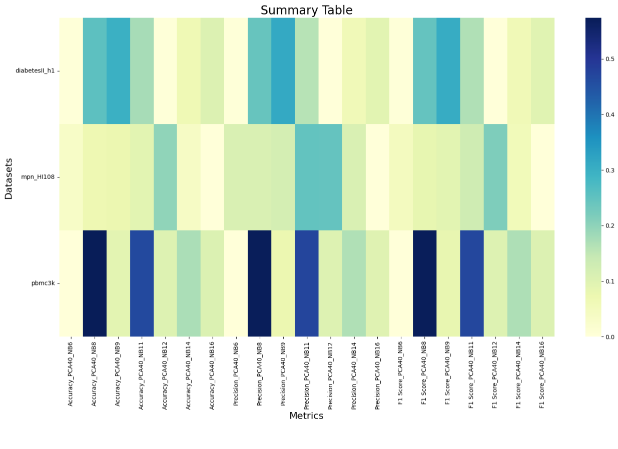

Fig. 5. Impact of Changing Neighbours on Evaluation Metric across all datasets

Impact of Varying Number of PCA: Figure 6 illustrates the performance metrics across different datasets and principal component values (PCA20, PCA25, PCA30, PCA35, PCA38 ,PCA42, PCA45, PCA50, PCA55) while keeping the number of neighbors fixed at NB10. The following observations can be made: The pbmc3k dataset consistently exhibits the highest performance metrics across all PCA values, further reinforcing the model's effectiveness in classifying cell types from healthy donors [4]. The diabetesII-h1 and mpn-H108 datasets show varying performance metrics across different PCA values. However, in this case the mpn-H108 dataset performs better than the diabetesII-h1 specially in terms of precision. [9] An intermediate number of PCA (e.g.,PCA35 PCA38, PCA42) tends to balance capturing local data structures with maintaining generalization, leading to improved performance. The higher precision observed for the mpn-H108 dataset compared to diabetesII-h1 suggests that the model is better able to accurately identify and classify the true positive cell types associated with MPNs, while minimizing false positives [9]. This could be attributed to the distinct molecular and cellular signatures present in MPN-affected cell populations, which may be more easily distinguishable by the model when the appropriate number of principal components is selected. [9] In contrast, the lower precision observed for the diabetesII-h1 dataset indicates a higher rate of false positives, where the model may be misclassifying certain cell types or failing to accurately capture the more subtle or heterogeneous cellular changes associated with type II diabetes mellitus [9], [1]. Furthermore, the higher precision observed for the mpn-H108 dataset could potentially be attributed to the more distinct and pronounced cellular changes associated with hematological malignancies, such as abnormal cell proliferation and differentiation, compared to the more gradual and systemic effects of metabolic disorders like type II diabetes mellitus [9], [1]. In summary, The high performance metrics for the pbmc3k dataset suggest that cell populations in healthy donors are more distinct and easier to classify, likely due to the lower complexity and clearer separation of cell types in a non-disease state [4]. The comparatively lower performance metrics for the mpn-H108 and diabetesII-h1 datasets could be attributed to the increased complexity and heterogeneity of disease-affected cell populations, potentially requiring additional refinement in feature selection or model parameters to improve classification accuracy [1].

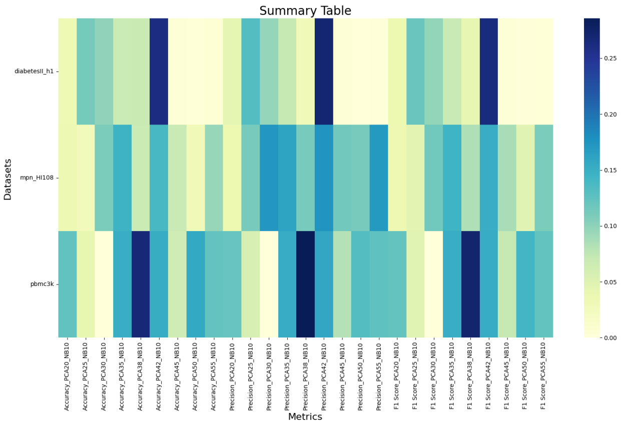

Fig. 6. Impact of Changing PCA on Evaluation Metric across all datasets

V. Conclusion

The comprehensive analysis presented in this study highlights the significant impact of hyperparameter selection on the performance of single-cell classification models. By systematically varying the number of neighbors and principal components across three distinct datasets, we gained valuable insights into the intricate interplay between these hyperparameters and the resulting evaluation metrics. The pbmc3k dataset, derived from a healthy donor, consistently exhibited the highest performance metrics, underscoring the model's ability to accurately classify cell types in a non-diseased state. This observation suggests that cell populations in healthy individuals exhibit more distinct and well-separated characteristics, facilitating robust classification. In contrast, the diabetesII-h1 and mpn-H108 datasets, representing type II diabetes mellitus and myeloproliferative neoplasm respectively, displayed comparatively lower performance metrics. This could be attributed to the increased complexity and heterogeneity of disease-affected cell populations, which may require further refinement in feature selection or model architecture to improve classification accuracy. The analysis also revealed that an intermediate number of neighbors (e.g., NB08, NB09, NB11, NB12) and PCA(e.g.,PCA35 PCA38, PCA42) tends to strike a balance between capturing local data structures and maintaining generalization, leading to improved performance across all datasets. In summary, this study emphasizes the critical role of hyperparameter tuning in single-cell analysis and provides valuable insights into the impact of varying neighbors and principal components on classification performance.This project represents a significant effort toward automating key stages of single-cell RNA sequencing (scRNA-seq) data analysis. By developing robust pipelines for data preprocessing, clustering, cell type annotation, and evaluation metric computation, I aim to streamline the analytical process, enabling researchers to obtain valuable insights more efficiently.The findings highlight the need for careful consideration of dataset characteristics and disease contexts when selecting appropriate hyperparameters. Future research could explore the integration of additional hyperparameters, advanced feature selection techniques, and ensemble models to further enhance the accuracy and robustness of single-cell classification models in biomedical applications.

References

[1] Stephen Cottrell, Yusuke Hozumi, and Guo-Wei Wei. K-nearest-neighbors induced topological pca for single cell rna-sequence data analysis. *ArXiv [Preprint]*, Oct 23 2023. arXiv:2310.14521v1.

[2] Francis Crick. Central dogma of molecular biology. *Nature*, 1970.

[3] Marie Cumberbatch, Geoffrey Ivison, Amy Lam, Aaron T. Mayer, and Milan Bhagat. Abstract 4623: Single-cell spatial proteomic analysis of the tumor microenvironment in treatment-naive nsclc samples with immunotherapy treatment and response data. *Cancer Research*, 2023.

[4] Lars Heumos, Aaron C Schaar, Charles Lance, et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(9):550–572, 2023.

[5] Mena Soliman Asaad Kamel, Amrut Sarangi, Cindy Qin, Het Barot, Pavel Senin, Sergio Villordo, Sunaal Mathew, Albert Pla Planas, and Ziv Bar-Joseph. Deep-learning based cell segmentation and deconvolution in spatial transcriptomics. *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2023.

[6] T. McKee, Mark Zaidi, and Veronica Cojocari. Abstract pr-06: Utilizing biological domain knowledge and machine learning methods to improve cellular segmentation on multiplex fluorescence and imaging mass cytometry datasets improves the quality of single-cell data obtained. *Oral Presentations - Proffered Abstracts*, 2021.

[7] Zulhafizal Othman, Aisyah Mat Jasin, Muhd Eizan Shafiq Aziz, Mohd Khairul Izamil Zolkefley, Ainamardia Nazarudin, Hamizah Mokhtar, and Amminudin Ab. Latif. Visualization of word similarity measurement for messages in sequence diagram using heatmap. *2023 IEEE 13th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pages 360–366, 2023.

[8] MB Pouyan and M Nourani. Clustering single-cell expression data using random forest graphs. *IEEE Journal of Biomedical and Health Informatics*, 21(4):1172–1181, Jul 2017. Epub 2016 May 10.

[9] Forrest W. Townes, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome Biology*, 20:295, 2019.

[10] FA Wolf, P Angerer, and FJ Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, Feb 2018.

[11] Musu Yuan, Liang Chen, and Min Deng. Clustering single cell cite-seq data with a canonical correlation based deep learning method. *bioRxiv*, 2021.