Clustering and Classification of Single-Cell RNA Sequencing (ScRNA-Seq) Data with Artificial Intelligence (AI)

Faisal Asadia) Nobel Shan Setionob), Roland Febrianc) Bernardus Ignasiod), Jasson Widiartae), and Albert Yulius Ramahalimf)

Department of Computer Science, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

a) Corresponding author: Faisal.asadi@binus.ac.id

b) nobel.setiono@binus.ac.id

c) roland.febrian@binus.ac.id

d) bernardus.ignasio@binus.ac.id

e) jasson.widiarta@binus.ac.id

f) albert.ramahalim@binus.ac.id

**Abstract.** This literature review examines recent advancements in the clustering and classification of single-cell RNA sequencing (scRNA-seq) with Artificial Intelligence (AI). We selected fifteen critical studies and explored six state-of-the-art clustering frameworks, seven classification methods, an evaluation and recommendation system, and a sample application of scRNA-seq clustering. We explored the methods used in each framework and how the models were evaluated. We also reviewed the datasets used in training and benchmarking the models. Finally, in this research paper, we give a comprehensive summary of the frameworks' performance to better illustrate the performance and limitations of these frameworks.

**Keywords:** Single-cell RNA sequencing, clustering frameworks, classification methods, AI in genomics

# INTRODUCTION

In recent years, the existence of single-cell RNA sequencing (scRNA-seq) has helped numerous researchers to gain insight and redefine our understanding of heterogeneity and dynamics of the cells as it refers to the sequencing of single-celled genomics and transcriptomic profiles from a specific tissue [1]. This results in a better understanding of gene expressions in each cell, which helps explore more complex biological systems since it can characterize and differentiate different cell patterns within different tissues and conditions. However, scRNA-seq data contains much noise and has high dimensionality and sparsity, thus leading to increased difficulty in analyzing data. Therefore, researchers have turned to applying artificial intelligence to unravel the intricacies of scRNA-seq datasets. They can develop automatic cell classification using methods such as clustering and classification. This comes with a challenge, as both methods have their problems. Due to the sparsity of the data, clustering can lead to most measurements outputting zeros, resulting in a 'false' zero count observation [2]. Classification algorithms also show a problem as the success of the classification is dependent on the variety of cell types inside the reference dataset, which severely limits the extrapolation of cell types [1]. Through this review, we aim to evaluate existing clustering and classification algorithms and techniques that have already been presented, discussing their challenges, robustness, and complexity. Through this evaluation, we hope to provide a comprehensive perspective on the best practices for utilizing artificial intelligence in analyzing scRNA-seq data.

# Literature Review

We have selected fifteen works that we will analyze in this review. These papers are selected based on specific search procedures and inclusion criteria, which we will outline in the methodology. There have been many existing clustering and classification models for scRNA-seq data, such as scClassify, SINC, hierarchical clustering, K-means clustering, pcaReduce, single-cell analysis in python (SCANPY), and single-cell consensus clustering (SC3) [3-6]. Each of these methods comes with its limitations. Many of these methods implicitly assume that the scRNA-seq data follows a particular distribution, most frequently the Poisson or negative binomial distribution. However, the distribution of a sizeable scRNA-seq dataset may be mixed and complex [7]. Another drawback is the high sensitivity of clustering results to the input parameters. This makes parameter tuning critical, but many parameters are not user-intuitive [8]. To address these limitations, several attempts have been made to propose a new framework for scRNA-seq classification and clustering. An in-depth explanation of these works will be presented in the results.

## MODEL TECHNIQUES

Before performing classification or clustering, dimensionality reduction on scRNA-seq data is expected to be performed first due to its high dimensionality. Frequently used dimensionality reduction techniques include principal component analysis (PCA), uniform manifold approximation and projection (UMAP), and t-distributed stochastic neighbor embedding (tSNE) [9]. Many proposed classification and clustering models are based on machine learning techniques, such as k-nearest neighbors (kNN), k-means, hierarchical clustering, singular value decomposition (SVD), support vector machine (SVM), ensemble learning, and transfer learning [9-14]. Several of the proposed classification models are based on neural networks, such as multi-layer perceptron (MLP), long short-term memory (LSTM), convolutional neural network (CNN), and deep neural network (DNN) [1, 2, 15, 16]. Novel algorithms were also proposed, such as the iterative cluster generation in PanoView [8], which uses a novel density-based algorithm called Ordering Local Maximum by Convex hull (OLMC).

## EVALUATION DATASETS

Assessing the proposed models' performance is commonly done by comparing them to existing models on both simulated and real-world data. Simulated data can be generated using several methods, such as the Splatter Bioconductor package, with adjusted parameters on several clusters, sample sizes, differentially expressed gene proportions and probability of excess zeros [2, 7, 10]. Another generator is Scikit, which is used in [8]. Real-world datasets are obtained from various databases, such as the National Center for Biotechnology Information (NCBI) search database or ArrayExpress [7, 8, 13]. Several works also used primary data obtained directly during the research process. For instance, work [8] used mouse embryonic hypothalamus cells, while the research in [14] gathered scRNA-seq data from human kidneys.

## EVALUATION METRICS

A quantitative measurement of performance is often desired. Two metrics commonly used in evaluating clustering results are the Adjusted Rand Index (ARI) and normalized mutual information (NMI) [17]. For classification, common metrics are accuracy or misclassification rate, AUC or AUROC (area under receiver-operating curve), AUPRC (area under precision-recall curve), recall, precision, and the F1-score [12]. Measures of sensitivity and specificity [13] are often found as well.

# Methods

To conduct the systematic literature review, we conducted a comprehensive search to identify relevant papers on the clustering and classification of single-cell RNA sequencing with AI.

## INCLUSION AND EXCLUSION CRITERIA

We established the following inclusion criteria to ensure the efficiency of our search.

The paper must focus on the clustering or classification of scRNA-seq data and include an application of AI.

The paper must be published in an open-access, peer-reviewed journal or conference.

The paper must be written in English.

The paper must be written during the scope of the past five years (from 2019 to 2024, inclusive).

The exclusion criteria for the target articles are as follows.

1. The paper is a review of other related articles.

The paper does not have full-text availability.

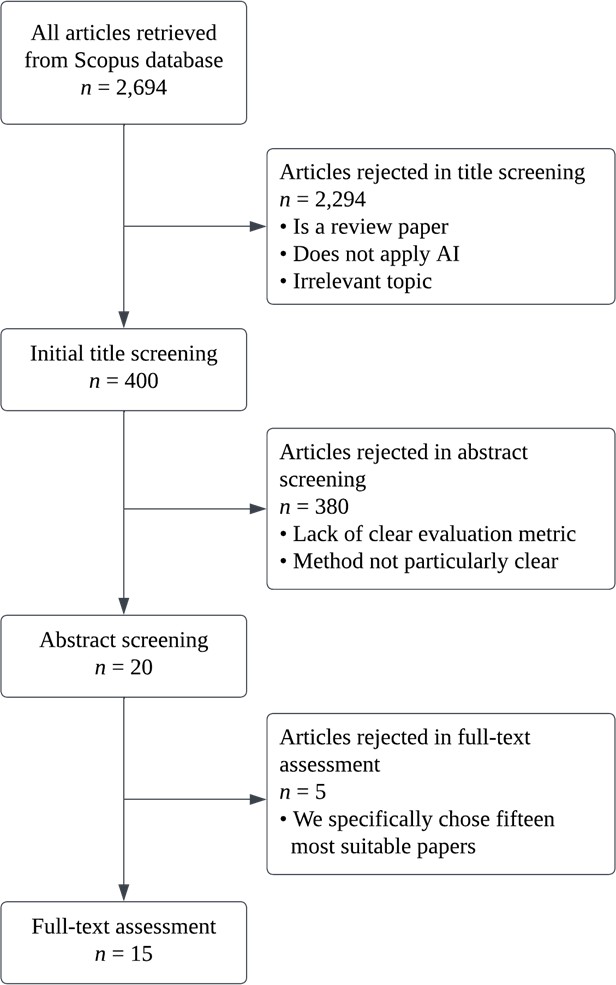
The paper is outside of the considered topic or time.

## PAPER SEARCH

Our workflow for article selection is illustrated in **FIGURE 1**. Furthermore, we initially searched the Scopus database using the following search key: (TITLE-ABS-KEY ("cluster\*") OR TITLE-ABS-KEY ("classify\*")) AND TITLE-ABS-KEY ("single-cell RNA sequence\*") AND PUBYEAR > 2018 AND OA (all) AND LANGUAGE ("English")

Wildcards were used to accommodate the various possible word endings. Paper sources must also come from reputable publishers such as NCBI and Scopus. To ensure the sources' credibility, originality, and quality. The initial search resulted in 2,694 articles. As it is impossible to check the presence of AI techniques in the paper via only a database search, reviewers performed an initial screening of article titles to select 400 papers.

Further reviews via abstract screening reduced the number of candidates to twenty. Finally, fifteen papers were chosen after assessing the full text to ensure suitability. All fifteen papers were analyzed closely to understand each proposed method's function, impact, and efficiency in classifying and clustering single-cell RNA sequencing data.

Moreover, the initial search resulted in 2,694 articles. As it is impossible to check the presence of AI techniques in the paper via only a database search, reviewers performed an initial screening of article titles to select 400 papers. Further reviews via abstract screening were done to reduce the number of candidates to twenty. Finally, fifteen papers were chosen after assessing the full text to ensure suitability. All fifteen papers were analyzed closely to understand each proposed method, function, impact, and efficiency in classifying and clustering single-cell RNA sequencing data.

**Figure 1**. A flowchart outlining our article selection workflow

# RESULTS AND DISCUSSION

The fifteen papers are mainly frameworks for performing clustering or classification. In our review, we have also included an evaluation and recommendation system, as well as a sample application of scRNA-seq clustering with AI.

## CLUSTERING FRAMEWORKS

We reviewed a total of six clustering frameworks as follows. A summary of these frameworks is outlined in **TABLE 1**. Firstly, we found that PanoView is a novel iterative density-based clustering algorithm. Each iteration creates a PCA space and identifies the cluster with the lowest variance. The clusters are created through an algorithm named Ordering Local Maximum by Convex Hull (OLMC). The code for PanoView is publicly available at github.com/mhu10/scpanoview/. PanoView was evaluated with nine other models on simulated and real-world data. On simulated data, it performed better than the other methods together with SCANPY. Eleven real-world datasets were used, where PanoView outperformed other models, especially on larger dataset sizes.

However, PanoView performed slower than other clustering frameworks such as SCANPY, Seurat, and RCA [8]. Secondly, we reviewed Two-Step Clustering (TSC), which performs two main clustering steps: hierarchical clustering core cells and assigning non-core cells to corresponding clusters. Here, core cells are defined as closely connected to their neighbors based on a defined distance metric [11]. The code for TSC is publicly available at github.com/LiRuiyi-raptor/TSC\_Project. TSC was compared to seven other methods on twelve real-world datasets. It was shown that TSC outperformed all other methods on all datasets, achieving an average ARI of 0.79 with the Pearson correlation coefficient as the distance metric [11].

Moreover, we also reviewed the third, Contrastive-sc, a self-supervised clustering algorithm that performs contrastive representation learning to produce an embedding for each cell through an artificial neural network. The embeddings can then be clustered by a general clustering algorithm such as k-means or Leiden [10]. The code for contrastive-sc is publicly available at github.com/ciortanmadalina/contrastive-sc. Contrastive-sc was evaluated on simulated and accurate data by comparing its performance with 11 other models. There was no constant best method among the datasets, but contrastive- sc performed highly together with scanpy-seurat and scziDesk. It was noted that contrastive-sc has a runtime of up to eight times faster than methods with similar architecture [10].

Besides, ItClust is a supervised clustering method utilizing transfer learning, which performs clustering on unlabeled target data after training on labeled source data. It builds a neural network to extract cell-type specific gene expression signatures, then performs iterative fine-tuning to gather suitable parameters for the target data [14]. The code for ItClust is publicly available at github.com/jianhuupenn/ItClust. ItClust was evaluated using 12 real datasets from NCBI and ArrayExpress databases and one self-collected dataset of human kidney cells. ItClust was compared to three other methods (Louvain, DESC, SAVER-X) and outperformed them. Its authors argued that this was because ItClust could learn cell types from the source dataset to perform its clustering [14].

Furthermore, scDeepCluster is a deep learning model using encoder and decoder architecture. Unlike the normal encoder and decoder architecture, it uses ZINB loss, which gives the model zero inflation and overdispersion. Another used trick is the application of KL divergence, which allows it to manipulate the separation of clusters inside the latent space [2]. The code for scDeepCluster is available at the GitHub repository github.com/ttgump/scDeepCluster/. scDeepCluster was compared against seven other methods on simulated data of 1500 cells and 2500 genes. The experiment was repeated 20 times under the same setting to minimize errors. The performance of scDeepCluster results in a robust clustering model by retaining a perfect clustering concordance. It was then evaluated using four real scRNa-seq datasets: PBMC 4k cells, mouse embryonic stem cells, mouse bladder cells, and worm neuron cells. It also results in scDeepCluster outperforming all the other methods in all four datasets [2]. We last found that scMPN uses a combination of MLP and graph neural networks to perform gene imputation and cell clustering. It incorporates a graph attention mechanism and a variation graph auto-encoder (VGAE) [15]. The VGAE was proven to be crucial through ablation study. The authors gave us a link to the code of scMPN, but we could not access it at the time of writing. scMPN was evaluated against five other methods with four real-world datasets. It achieved the best results among all the methods used for all six evaluation metrics [15]. It needs to be noted that the methods scMPN was benchmarked with can perform both gene imputation and clustering.

**TABLE 1.** Summary of clustering frameworks.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Framework Name** | **Ref.** | **Technique** | **Evaluation Results** | | **Availability** |
| **Simulated Data** | **Real-World Data** |
| PanoView | [8] | Iterative density-based clustering in PCA space using Ordering Local Maximum by Convex Hull (OLMC) | * From Scikit * Similar performance with SCANPY, better than other models | * Eleven published datasets * Outperformed other methods, especially on larger datasets | GitHub repository |
| TSC | [11] | Hierarchical clustering of core cells and assignment of non-core cells | N/A | * Twelve published datasets * Outperformed other methods with ARI up to 0.79 | GitHub repository |
| contrastive-sc | [10] | Contrastive learning with ANN, general clustering on embedding | * There is no constant best method * Performed highly alongside scedar, scanpy-seurat, soup | * Fifteen published datasets * Has most agreement with ground truth alongside scziDesk and scDeepCluster | GitHub repository |
| ItClust | [14] | ANN with transfer learning on both labeled and unlabeled data | N/A | * Twelve published datasets * Outperformed other methods due to the learning of cell type | GitHub repository |
| scDeepCluster | [2] | Model-based deep-embedded clustering | * 1500 cells with 2500 genes each * Perfect clustering concordance (NMI ≈ 1) | * 10x PBMC, mouse ES cells, mouse bladder cells, worm neuron cells * Outperformed all other methods in all four datasets | GitHub repository |
| scMPN | [15] | MLP and graph neural network with a graph attention mechanism | N/A | * Four published datasets * Best results among six other methods for imputation and clustering | Unavailable |

## CLASSIFICATION FRAMEWORKS

We reviewed a total of seven classification frameworks as follows. A summary of these frameworks is outlined in **TABLE 2**. Firstly, scDLC is a deep-learning classifier that is based on two-layer LSTMs. Its authors believed that LSTMs could capture short-term and long-term time dependence. Thus, it can classify scRNA-seq data without assuming any underlying distribution of the data [7]. The code for scDLC is publicly available at github.com/scDLC-code/scDLC. scDLC was evaluated and compared with seven other models. The evaluation was done for both simulated and real-world data. Simulated datasets were generated using Splatter. Six datasets from the NCBI database were used for real-world data. scDLC outperformed all other baseline models on simulated data. On real-world data, scDLC and scPred performed similarly and better than other methods [7]. Secondly, scPred is a classifier based on machine learning techniques. It decomposes the gene expression matrix via SVD and trains an SVM on the training data using k-fold cross-validation [12]. The code for scPred is publicly available at github.com/powellgenomicslab/scPred. scPred was evaluated on a real-world dataset of tumor epithelial cells against two baseline models: differentially expressed genes (DEGs) and per-cell mean-of-logarithms. It outperformed them with an F1-score of 0.922.

However, training data shows reduced performance for rare cells [12]. Thirdly, CaSTLe performs classification using a pre-tuned XGBoost classifier after performing several feature engineering steps on scRNA-seq data. It is intended for transfer learning and emphasizes its parallelizability [13]. The code for CaSTLe is publicly available at github.com/yuvallb/CaSTLe. CaSTLe was compared to a majority vote classifier and two linear regression models with different features. The evaluation used nine real-world datasets from NCBI and ArrayExpress, from which twelve pairs were extracted to test transfer learning. On average, CaSTLe outperformed the benchmarks in 9 out of 12 cases. A downside is that it could have performed better on small datasets with many classes [13].

Besides, the fourth kNN-tPCA combines topological PCA with kNN-based filtration. Topological PCA reduces dimensionality through PCA, persistent Laplacian, and 𝐿2,1 norm regularization. The code for kNN-tPCA is publicly available at github.com/seanfcottrell/Topological- PCA. kNN-tPCA was tested only on real-world data, which consisted of 11 datasets from the NCBI database. It outperformed all baseline dimensionality reduction techniques, with a 13% improvement in F1-score over PCA averaged over all datasets [9]. The next is GOWDL, another deep-learning classifier that extends an average WDL model. The normal WDL model comprises a deep feed-forward neural network and a generalized linear model. The GOWDL architecture takes two inputs from both the non-relevant and relevant genes. The non-relevant genes are fed into a GOCNN architecture, while the relevant genes are fed into a generalized linear model [1]. The code for GOWDL is publicly available at github.com/bcb4pm/gowdl. Using the GOCNN rather than the normal deep feed-forward neural network lets the model understand more of the functional annotations of genes. GOWDL has been compared with ten other state-of-the-art tools and two general-purpose tree-based classifiers. This was tested on five independent scRNA-seq datasets from five different human tissues. The results suggest that the proposed architecture yields an equal or even better than the other 12 classifiers, making it viable for integrating more complex bioinformatic tasks [1].

Furthermore, we found ImmClassifier is a classification algorithm designed for immune cells. It is based on knowledge and hierarchical models and made with the target of higher performance in identifying fine-grained cell types. It uses a random forest classifier and a DNN, the former of which feeds class probabilities to the latter to be used for training. The code for ImmClassifier is publicly available at github.com/xliu- uth/ImmClassifier. ImmClassifier was compared with four other classifiers on ten real-world immune cell datasets (7 for training and 3 for testing). At more profound levels of the hierarchy, it performed better than the classifiers by a margin of 15% recall and 14% precision. Despite this, ImmClassifier cannot detect new or intermediate cell types; it will always assign a class to all cells [16], and the last is scAnnotatR is an SVM-based classifier, which consists of a of SVMs arranged in a tree to focus on classifying the closely related cell types. scAnnotatR was reported to be able to distinguish ambiguous classification results while having low execution time. The code for scAnnotatR is publicly available at github.com/grisslab/scAnnotatR. scAnnotatR was compared with ten other models on 19 real-world datasets. There were two types of benchmarks: discrete cell populations and closely related populations. scAnnotatR consistently performed well on both benchmarks and other models, SCINA and scClassify [18].

**TABLE 2.** Summary of classification frameworks.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Framework Name** | **References** | **Technique** | **Evaluation Results** | | **Availability** |
| **Simulated Data** | **Real-World Data** |
| scDLC | [7] | Deep neural network with two-layer LSTMs | * From Splatter * Outperformed other baseline models on misclassification rate | * Six datasets from the NCBI database * Similar performance with scPred, better than other models | GitHub repository |
| scPred | [12] | SVD on gene expression matrix, SVM on results | N/A | * Performed better than baseline models * Reduced performance on rare cells | GitHub repository |
| CaSTLe | [13] | Feature engineering followed by XGBoost, transfer learning | N/A | * Outperformed benchmarks in 9 out of 12 cases * Did not perform well on small datasets with many classes | GitHub repository |
| kNN-tPCA | [9] | Topological PCA results filtered based on the kNN network | N/A | * Eleven published datasets * Outperformed baseline models, 18.3% improvement from PCA | GitHub repository |
| GOWDL | [1] | Wide, deep learning model comprised of GOCNN and generalized linear model | N/A | * NCBI, Human Kidney Atlas, Human Cell Landscape, Immune Cell Atlas, Single Cell Portal * Comparable or better than other methods except when using recall validation | GitHub repository |
| ImmClassifier | [16] | Random forest classifier as input for DNN training | N/A |  +15% recall and +14% precision from other models   * Cannot detect new or intermediate cell types | GitHub repository |
| scAnnotatR | [18] | Tree of SVMs | N/A | * Nineteen published datasets * Consistently performed well together with SCINA and scClassify | GitHub repository |

## EVALUATION AND RECOMMENDATION FRAMEWORK: DDCR

DDCR is a recommendation framework developed to select a suitable downstream clustering method between hierarchical and spectral clustering. DDCR uses the QRS strategy to determine the latent shape of a dataset. This strategy performs dimensionality reduction with UMAP and then constructs the minimum spanning tree to conclude the inter-cluster relationships [17]. DDCR was evaluated on real-world data only, which were gathered from the NCBI database and ArrayExpress. Twenty well-annotated datasets on cells containing specific functional subsets and cells undergoing differentiation were used. Given two clustering methods (SC3 and RAFSIL), DDCR was able to correctly recommend the clustering method with better performance [17].

## SAMPLE APPLICATION OF SCRNA-SEQ CLUSTERING

A study [18] used scRNA-seq clustering to perform an analysis to characterize the heterogeneity of endothelial cells (ECs) in association with myocardial infarction (MI). The study clustered scRNA-seq data using Seurat to gather information on DEGs in the heart cells of mice with MI. Data was collected from two datasets on the NCBI database. Dimensionality reduction via PCA and tSNE was performed, followed by cluster analysis. Notably, scRNA-seq clustering with AI helped perform the study, especially in analyzing the EC subclusters [17]. Finally, we would like to highlight the datasets used in developing and evaluating the frameworks we reviewed. In **TABLE 3**, we present a summary of the eight most frequently used datasets among the clustering and classification frameworks.

**TABLE 3.** Summary of most frequently used dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset Name** | **Origin** | **Usage** | **Cell Resource** | **Number of classes** | **Number of cells** |
| GSE84133 (Baron) | NCBI | [1, 7, 12, 14] | Human pancreas | 14 | 8569 |
| GSE85241 (Muraro) | NCBI | [10, 12-14, 18] | Human pancreas | 9 | 2122 |
| E-MTAB-332 (Goolam) | ArrayExpress | [8, 11, 13] | Mouse preimplantation development | 5 | 124 |
| GSE45719 (Deng) | NCBI | [8, 9, 13] | Mouse embryo | 10 | 268 |
| E-MTAB-2600 (Kolodziejczyk) | ArrayExpress | [11-15] | Mouse embryonic stem cells | 3 | 704 |
| GSE60361 (Zeisel) | NCBI | [8, 17] | Mouse brain cortex and hippocampus | 9 | 3005 |
| GSE65525 (Klein) | NCBI | [11-15] | Human and mouse embryonic stem cells | 4 | 2717 |
| GSE94820 (Villani) | NCBI | [8, 9] | Human peripheral blood mononuclear cells | 10 | 1078 |

# ConclusionS

AI has been revolutionizing the analysis of single-cell RNA sequencing (scRNA-seq) by enabling further analysis through various clustering and classification techniques. Through this systematic literature review, we have highlighted fifteen works on scRNA-seq clustering and classification with AI, significantly state-of-the-art advancements leveraging various machine learning and deep learning techniques to enhance performance, from which we have produced a comprehensive summary together with our review. We have also noted the importance of evaluation and recommendation frameworks and applications of scRNA-seq clustering through our review of sample works. Further research could compare the performance of these clustering and classification frameworks through experiments on various datasets or develop benchmarks to better gauge the performance of the various proposed models.

# References

1. Fiannaca, A., et al., *GOWDL: gene ontology-driven wide and deep learning model for cell typing of scRNA-seq data.* Briefings in Bioinformatics, 2023. **24**(6): p. 1-10.

2. Tian, T., et al., *Clustering Single-Cell RNA-Seq Data with a Model-Based Deep Learning Approach.* Nature Machine Intelligence, 2019. **1**(4): p. 191-198.

3. Lin, Y., et al., *ScClassify: Sample Size Estimation and Multiscale Classification of Cells Using Single and Multiple Reference.* Molecular systems biology, 2020. **16**(6): p. e9389.

4. Žurauskienė, J. and C. Yau, *PcaReduce: Hierarchical Clustering of Single Cell Transcriptional Profiles.* BMC bioinformatics, 2016. **17**: p. 1-11.

5. Wolf, F.A., P. Angerer, and F.J. Theis, *SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis.* Genome biology, 2018. **19**: p. 1-5.

6. Kiselev, V.Y., et al., *SC3: Consensus Clustering of Single-Cell RNA-Seq Data.* Nature methods, 2017. **14**(5): p. 483-486.

7. Zhou, Y., et al., *SCDLC: a Deep Learning Framework to Classify Large Sample Single-Cell RNA-Seq Data.* BMC genomics, 2022. **23**(1): p. 504.

8. Hu, M.-W., et al., *PanoView: An Iterative Clustering Method for Single-Cell RNA Sequencing Data.* PLoS computational biology, 2019. **15**(8): p. 1-17.

9. Cottrell, S., Y. Hozumi, and G.-W. Wei, *K-Nearest-Neighbors Induced Topological PCA for Single Cell RNA-Sequence Data Analysis.* Computers in biology and medicine, 2024. **175**: p. 108497.

10. Ciortan, M. and M. Defrance, *Contrastive Self-Supervised Clustering of SCRNA-Seq Data.* BMC bioinformatics, 2021. **22**(1): p. 280.

11. Li, R., et al., *A New and Effective Two-Step Clustering Approach for Single Cell RNA Sequencing Data.* BMC genomics, 2023. **23**(Suppl 6): p. 864.

12. Alquicira-Hernandez, J., et al., *SCPred: Accurate Supervised Method for Cell-Type Classification from Single-Cell RNA-Seq Data.* Genome biology, 2019. **20**: p. 1-17.

13. Lieberman, Y., L. Rokach, and T. Shay, *CaSTLe–Classification of Single Cells by Transfer Learning: Harnessing the Power of Publicly Available Single Cell RNA Sequencing Experiments to Annotate New Experiments.* PloS one, 2018. **13**(10): p. 1-16.

14. Hu, J., et al., *Iterative Transfer Learning with Neural Network for Clustering and Cell Type Classification in Single-Cell RNA-Seq Analysis.* Nature machine intelligence, 2020. **2**(10): p. 607-618.

15. Feng, X., et al., *Advancing Single-Cell RNA-Seq Data Analysis Through the Fusion of Multi-Layer Perceptron and Graph Neural Network.* Briefings in Bioinformatics, 2024. **25**(1): p. 1-13.

16. Liu, X., et al., *Knowledge-Based Classification of Fine-Grained Immune Cell Types in Single-Cell RNA-Seq Data.* Briefings in bioinformatics, 2021. **22**(5): p. 1-11.

17. Tian, Y., et al., *A Data-Driven Clustering Recommendation Method for Single-Cell RNA-Sequencing Data.* Tsinghua Science and Technology, 2021. **26**(5): p. 772-789.

18. Nguyen, V. and J. Griss, *scAnnotatR: framework to accurately classify cell types in single-cell RNA-sequencing data.* BMC bioinformatics, 2022. **23**(1): p. 44.