Identification of Rare Cells from Single-cell RNA-Seq Data using Machine Learning

Soumi Ghosh

Identification of Rare Cells from Single-cell RNA-Seq Data using Machine Learning

Thesis submitted to the
XIM University
for award of the partial fulfillment of the degree

of

Bachelor of Technology

by

Soumi Ghosh

Under the guidance of

Dr. Monalisa Mandal

School of Computer Science & Engineering
XIM University
Bhubaneswar - 752 050, India
April 2022

CERTIFICATE

Date:___/___/2022

This is to certify that the thesis entitled **Identification of Rare Cells from Single-cell RNA-Seq Data using Machine Learning**, submitted by **Soumi Ghosh** to XIM University, is a record of bona fide research work under my supervision and I consider it worthy of consideration for the award of the partial fulfillment of the degree of **Bachelor of Technology** of the Institute.

Dr. Monalisa Mandal
Supervisor
Assistant Professor
School of Computer Science & Engineering
XIM University
Bhubaneswar - 752 050, India

DECLARATION

I certify that

a. The work contained in the thesis is original and has been done by myself under the general supervision of my supervisor.

b. The work has not been submitted to any other Institute for any degree or diploma.

c. I have followed the guidelines provided by the Institute in writing the thesis.

d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

e. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

f. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Soumi Ghosh

Dedicated to my family

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my supervisor Dr. Monalisa Mandal for her constant guidance and support through all the stages of my research work right from providing me different datasets, suggesting the methodology and helping me in implementing the same. She was patient enough to clear my doubts and encourage me to complete my thesis despite her busy schedule. It was a valuable learning experience for me and will be beneficial for my future works too.

I am specifically thankful to our dean Dr. Rudra Mohan Tripathy and all the faculty members of the School of Computer Science & Engineering for providing me with their feedbacks and giving me the opportunity for my research work. I also extend my thankfulness to our thesis coordinator, Dr. Chandan Misra for specifying and sharing the format to structure my thesis write-up.

Last but not the least, I would also like to thank my parents for their immense support and patience rendered till the completion of my thesis.

Soumi Ghosh

Abstract

Single-cell RNA sequencing (scRNA-seq) has emerged as an essential tool to dissect the cellular heterogeneity and decompose tissues into cell types or cell states. It offers enormous potential for *de novo* discovery. Cells are the basic units of organisms and the building blocks of various complex tissues. They are controlled by many factors that affect their cell status and features (e.g., cell type specific expression, senescence). ScRNA-seq allows measurement of the expression of genes at single-cell resolution in complex disease or tissue. Quantification of the mRNA transcripts in genome-wide basis is useful to characterize the molecular constituents as well as cellular states. Cell to cell differences in RNA transcripts and protein expression can greatly benefit research in areas of cancer, neurobiology, stem cell biology, immunology, developmental biology etc. Rare cells are the cell types that occur in very small numbers within tissues while other cell types surrounding them are highly abundant. Some examples of rare cell types include circulating tumour cells, cancer stem cells, endothelial progenitor cells, antigen-specific T cells, invariant natural killer T cells etc. Rare cells play an important role in the diagnosis and prognosis of many cancers, prenatal diagnosis, diagnosis of viral infections, mediating immune responses etc. Conventional cell-based assays mainly analyse the average responses from a population of cells, without regarding individual cell phenotypes. To better understand the variations from cell to cell, scientists need to use single cell analyses to provide more detailed information for therapeutic decision making in precision medicine. ScRNA-seq needs the isolation as well as lysis of the single cells, the transformation of their corresponding RNA to cDNA, and the amplification of the cDNA to produce the high-throughput sequencing libraries. Although many methods have been developed to detect cell clusters from the scRNA-seq data, this task still remains a major challenge for the researchers. Larger sample size would make the evaluation better but a larger number of cell type-related transcripts are not identified in current scRNA-seq because of the failure during the stage of amplification and the relative limitation of short read coverage. Consequently, a limited number of cell type-related genes might fail to affect the downstream analysis regime in sufficient way. Sparsity in scRNA-seq data can also hinder downstream analyses and is still challenging to model or handle appropriately, calling for further method development. Our proposed method for the detection of rare cells from single cell gene expression profile data is based on Optimized Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to segregate the outliers effectively. The use of these computational techniques will further produce accurate and faster results than the traditional wet-lab experimentation.

Keywords: ScRNA-seq, DBSCAN, Rare Cell, Sparsity

Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

List of Symbols and Abbreviations

$>$	Greater than
$<$	Less than
$=$	Equal To
$\neq$	Not Equal To
$\in$	Belongs to
%	Percentage
$tr()$	Trace
max	Maximum value
$\sum$	Summation
&	Ampersand
W	Weight
c_1	Cognitive constant
c_2	Social constant
P_b	Local best
g_b	Global best
U_1,U_2	Random numbers

List of Abbreviations

RNA	Ribonucleic Acid
DNA	Deoxyribonucleic Acid
ScRNA-seq	Single-cell RNA Sequencing
DBSCAN	Density-based Spatial Clustering of Applications with Noise
PSO	Particle Swarm Optimization
ANN	Artificial Neural Network
PCA	Principal Component Analysis
t-SNE	t-Distributed Stochastic Neighbor Embedding
DEG	Differentially Expressed Genes

Publications out of this Dissertation

Identification of rare cell types from single-cell gene expression profiles using optimized DBSCAN (to be communicated)

Chapter 1

Introduction

Research in the field of genomics has been revolutionized with the development of new techniques and technologies leading to a better understanding of the biology of an organism. Cells are basic units of organisms and the building blocks of various complex tissues; they are controlled by many factors that affect their cell status and features (e.g., cell type specific expression, senescence) [15].Organisms store information as DNA, transmit information as RNA, and transform information into the proteins that perform most of the functions of cells vital for our survival.

Deoxyribonucleic acid or DNA, is the hereditary material in humans and almost all other organisms. Nearly every cell in a person's body contains the same DNA. Most of the DNA is present in the cell nucleus (Nuclear DNA) and a small amount of DNA can also be found in the mitochondria (mitochondrial DNA or mtDNA).The information in DNA is encoded in four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Based on their order or sequence, these bases determine the information available for building and maintaining an organism, like the letters of the alphabet form words and sentences according to a certain order. DNA bases pair up with each other, A with T and C with G, to form base pairs. Each base is bound to a sugar molecule and a phosphate molecule. A base, sugar, and phosphate together constitute a nucleotide.

Nucleotides are arranged in two long strands that form a spiral called a double helix that constitutes the DNA. The genetic code in DNA is the basis for the production of various molecules, including RNA and protein.

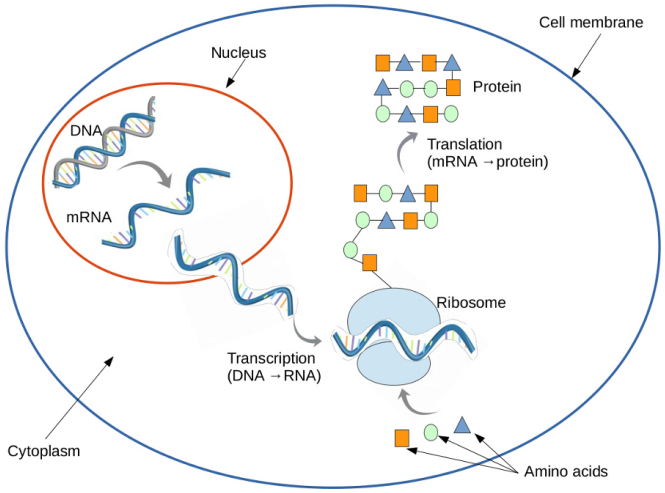

Figure 1.1: The Central Dogma of Molecular Biology [1]

In order to extract the information from DNA and transfer it to the location of cellular machinery that can carry out its instructions (usually the blueprints for a protein), the DNA code is "transcribed" into a corresponding sequence in a carrier molecule called ribonucleic acid or RNA. The portions of DNA that are transcribed into RNA are called "genes". The RNA polymerase enzyme moves stepwise along the DNA, unwinding the DNA helix at its active site. As it progresses, the polymerase adds nucleotides one by one to the RNA chain at the polymerization site using an exposed DNA strand as a template. The RNA transcript is thus a single-stranded complementary copy of one of the two DNA strands. The polymerase has a rudder that displaces the newly formed RNA, allowing the two strands of DNA behind the polymerase to rewind [16]. The sugar in RNA is ribose instead of dexoyribose which makes RNA more flexible and less durable than DNA. The nitrogen base uracil (U) takes the place of thymine (T) in RNA. The pre-messenger RNA is chopped up to remove the introns and create messenger RNA

or mRNA in a process called RNA splicing. The individual exons in mRNA are either spliced or included, producing several different possible mRNA products. each of which codes for a different protein isoform. These protein isoforms differ in their peptide sequence and therefore their biological activity. The mRNA formed in transcription is transported out of the nucleus, into the cytoplasm, to the ribosome (the cell's protein synthesis factory) and directs protein synthesis with the assistance of transfer RNA or tRNA [17].

Sequencing a nucleic acid molecule, such as DNA or RNA, provides information on the sequence of nucleotides or bases(A, C, G, and T) to investigate the functions of genes. In 1977 Frederick Sanger invented Sanger sequencing, a DNA sequencing method that utilizes modified dideoxynucleotides to cause chain-termination for which he was awarded his second Nobel prize in 1980. For the first time, we could read our genetic code. However, this method is expensive and labor-intensive. It took 13 years and about $ 3 billion to complete the "Human Genome Project" by 2003 (the first draft was published in 2001) in which, essentially the entire human genome was sequenced. Sequencing was not economically feasible for an individual research project using microarrays until next generation sequencing (NGS) methods were introduced to the market in 2005. NGS enables high-throughput sequencing by synthesis (SBS), reducing cost and increasing feasibility [18]. Traditional next-generation sequencing (NGS) examines the genome of a cell population, such as a cell culture, a tissue, an organ, or an entire organism. Its output is the "average genome" of the cell population. On the other hand, single cell sequencing measures the genomes of individual cells from a cell population [19]. Technologies like microarray and bulk RNA sequencing can only profile the average gene expression level among a large cell population that often contains significant heterogeneity. As a result, it is likely that many cell types remain unrecognized. ScRNA-Seq allows a transcriptome-wide gene expression measurement at single-cell level, enabling cell type cluster identification, the arrangement of populations of cells according to novel

hierarchies, and the identification of cells transitioning between individual states [20].

The sequencing an entire transcriptome at the level of a single-cell was pioneered by James Eberwine et al and Iscove and colleagues, who expanded the complementary DNAs (cDNAs) of an individual cell using linear amplification by in vitro transcription and exponential amplification by polymerase chain reaction (PCR), respectively. These technologies were initially applied to commercially available, high-density DNA microarray chips and were subsequently adapted for single-cell RNA sequencing (scRNA-seq). The first description of single-cell transcriptome analysis based on a next-generation sequencing platform was published in 2009, and it described the characterization of cells from early developmental stages. Since this study, there has been an explosion of interest in obtaining high-resolution views of single-cell heterogeneity on a global scale. Critically, assessing the differences in gene expression between individual cells has the potential to identify rare populations that cannot be detected from an analysis of pooled cells. [21] Single-cell RNA sequencing can also be used to obtain genomic, transcriptome or other multi-omics information to reveal cell population differences and cellular evolutionary relationships.It offers enormous potential for de novo discovery.

The first and crucial limiting step in scRNA-seq is the isolation of single cells. Cells need to be freed from extracellular matrix and cell-cell adhesion (except for circulating cells).This is achieved by different approaches such as enzymatic treatment, limiting dilution, micromanipulation, flow-activated cell sorting (FACS), laser capture microdissection (LCM) and microfluidics. These single-cell isolation protocols have their own advantages and show distinct performances in terms of capture efficiency and purity of the target cells [22]. Common steps required for the generation of scRNA-seq libraries include cell lysis, reverse transcription into first-strand cDNA, second-strand synthesis and cDNA amplification. The isolated individual cells are lysed to allow capture of as many RNA molecules as possible. In order to specifically analyse polyadenylated mRNA molecules and avoid capturing ribosomal RNAs, poly[T]-primers are commonly

used. Analysis of non-polyadenylated mRNAs is more challenging and requires specialized protocols. Poly[T]-primed mRNA is then converted to complementary DNA (cDNA) by a reverse transcriptase. Depending on the scRNA-seq protocol, the reverse-transcription primers will also have other nucleotide sequences added to them, such as adaptor sequences for detection on NGS platforms, unique molecular identifiers to mark unequivocally a single mRNA molecule, as well as sequences to preserve information on cellular origin. The minute amounts of cDNA are then amplified either by PCR or, in some instances, by in vitro transcription followed by another round of reverse transcription.Some protocols opt for nucleotide barcode-tagging at this stage to preserve information on cellular origin. Then, amplified and tagged cDNA from every cell is pooled and sequenced by NGS, using library preparation techniques, sequencing platforms and genomic-alignment tools [23].Once reads are obtained from scRNA-seq experiments, quality control (QC) is done for inspecting quality distributions across entire reads.

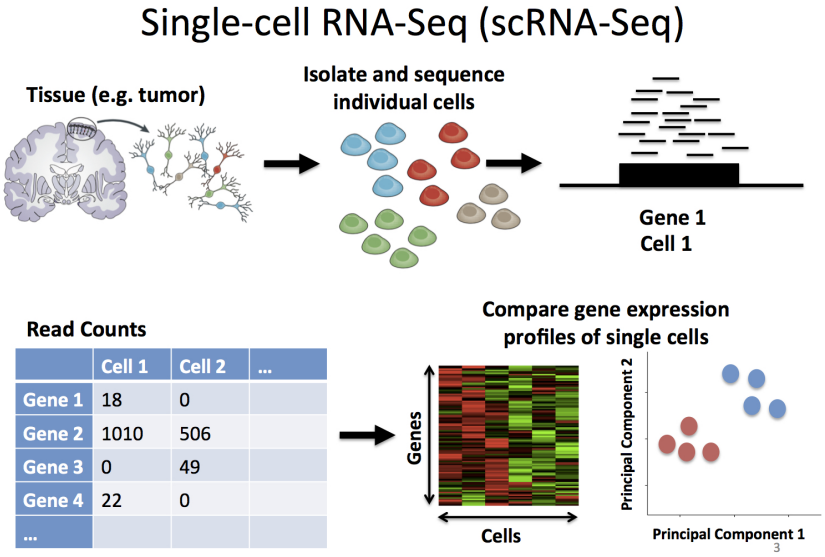

Figure 1.2: Single-Cell RNA-Sequencing [2]

After sequencing, alignment and de-duplication are performed to quantify an initial gene expression profile matrix. Next, normalization is performed with raw expression data using various statistical methods. Additional QC can be performed when using spike-ins by inspecting the mapping ratio to discard low-quality cells. Finally, the normalized matrix is then subjected to main analysis through clustering of cells to identify subtypes. Cell trajectories can be inferred based on these data and by detecting differentially expressed genes between clusters. [21]

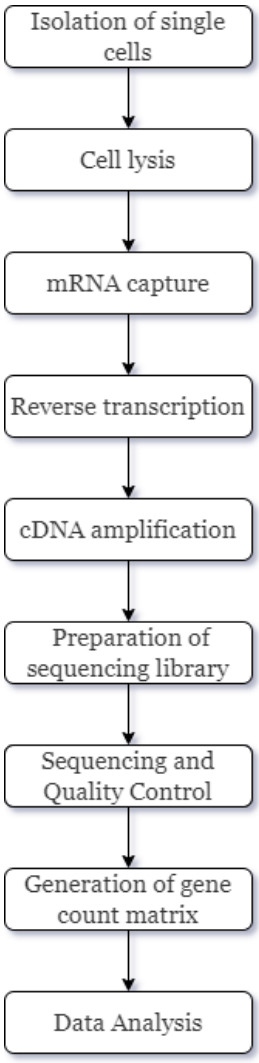

Figure 1.3: General workflow of single-cell RNA-sequencing (scRNA-seq) experiments

With the advent of new technologies and fully commercialized systems, the throughput and availability of single-cell RNA sequencing is increasing rapidly. ScRNA-seq has driven many discoveries and innovations in medicine in recent years. Quantification of the mRNA transcripts in genome-wide basis is useful to characterize the molecular constituents as well as cellular states. It also facilitates the investigation of underlying structures in tissue, organism development and diseases as well as the identification of unique subpopulations in cell populations that were so far perceived as homogeneous. Heterogeneity analysis remains a core reason for embarking on scRNA-seq studies. For instance, the ability to find and characterize outlier cells within a population has potential implications for better understanding of drug resistance and relapse in cancer treatment [24]. Cell to cell differences in RNA transcripts and protein expression can greatly benefit research in areas of cancer, neurobiology, stem cell biology, immunology, developmental biology etc.

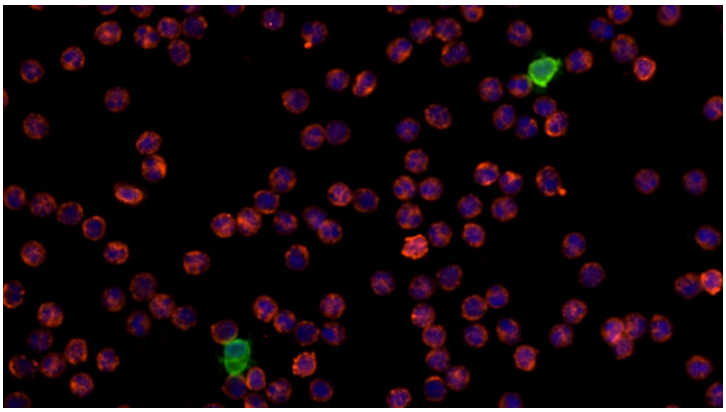

Figure 1.4: Multiple myeloma cells from liquid biopsy sample on an AccuCyte slide. CD45 (red) / CD138 (green) / nuclei (blue) [2]

Rare cells are the cell types that occur in very small numbers within tissues while other cell types surrounding them are highly abundant. Some examples of rare cell types include circulating tumour cells, cancer stem cells, endothelial progenitor cells, antigen-specific T cells, invariant natural killer T cells etc. Rare cells play an important role in the diagnosis and prognosis of many cancers, prenatal diagnosis, diagnosis of viral

infections, mediating immune responses etc. ScRNA-seq has been used to investigate rare antigen-specific T or B cells, measure the composition and structure of human microbiomes, study the origin and development of chemoresistant tumor subpopulations, discover previously unknown genes' functions in plant tissues, study tumor progression mechanisms, base prognostic predictions on intra-tumor cellular heterogeneity etc. [25]

Although many methods have been developed to detect cell clusters from the scRNA-seq data, this task still remains a major challenge for the researchers. The integration of precision, sensitivity, accuracy and the number of cells that have been used in analysis, evaluates the power for identifying the relative differences in the levels of expression. To choose a proper and efficient method among all the state-of-the-art scRNA-seq techniques, it requires to measure those parameters accurately [15]. It is well-known that the characterization of entire cell types in a tissue needs the processing of several thousands of single cells. Larger sample sizes better the odds of capturing minor cell subpopulations in a tissue. However, a larger number of cell type-related transcripts are not identified due to the failure during the amplification stage and the relative limitation of short read coverage. As a result, a small number of cell type-specific genes often fail to influence the downstream analysis regime sufficiently [15]. Sparsity in scRNA-seq data can also hinder downstream analyses and is still challenging to model or handle appropriately, calling for further method development. From an analytical point of view, the high dimensionality and complexity of scRNA-seq data pose significant challenges. Numerous computational methods tailored to scRNA-seq analysis are available, however comprehensive performance comparisons between those are scarce. This is primarily due to the lack of reference datasets with known cellular composition. Prior knowledge or synthetic data are commonly used to circumvent the problem of missing ground truth [26]. As research continues to deepen, the capabilities of single-cell sequencing methods and specialized computational techniques, continue to increase and evolve towards significant lower detection costs.

Chapter 2

Literature Review

Multicellular organisms are composed of diverse cell types with distinct morphologies and functions. Studies have shown that genetic or genomic variation can lead to cells with different genetic and phenotypic characteristics within the tumour tissue, making the it highly heterogeneous.The transcriptome-wide analyses of single cells using single-cell RNA sequencing enables to discover the interesting biomedical insights as well as biological perception. Characterizing the cell-to-cell differences is essential for basic developmental biology research, discovery of many unrecognized cell types in diverse tissues, improved characterization of cancer heterogeneity, clinical diagnosis, drug discovery and treatment of human diseases etc.

In 2018, a team at Massachusetts General Hospital and the Broad Institute used single cell gene expression and in vivo lineage tracing to catalogue cells that line the mouse tracheal epithelium. They found novel cell population is similar to salt-balancing ionocytes in fish gills or frog skin. Although they represent only 1% of airway epithelial cells, pulmonary ionocytes are the primary source of CFTR gene activity in both mouse (Cftr) and human (CFTR) . This discovery redirects efforts to correct CFTR mutations that are responsible for cystic fibrosis [4].

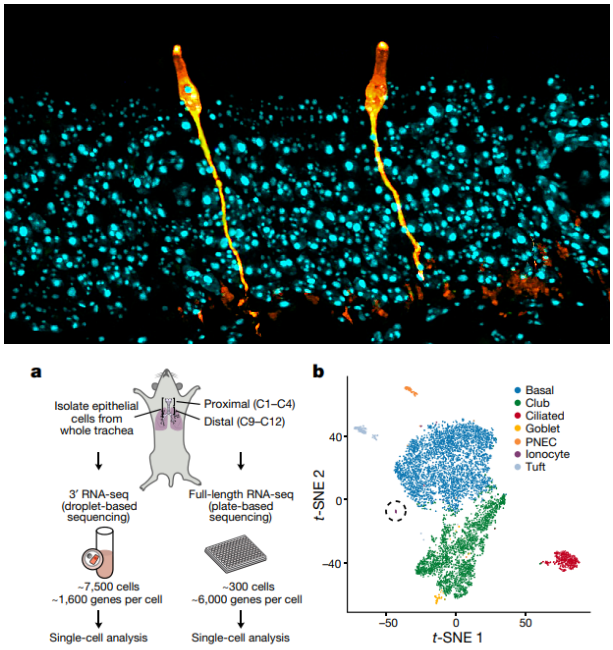

Fig. 1 | A single-cell expression atlas of mouse tracheal epithelial cells. **a**, Overview of the analysis. **b**, *t*-distributed stochastic neighbour embedding (*t*-SNE) of 7,193 3′ scRNA-seq profiles, coloured by cluster assignment and annotated post hoc. The ionocyte cluster is circled. PNEC, Pulmonary neuroendocrine cells.

Figure 2.2: Pulmonary ionocytes (orange) extend through neighboring epithelial cells in the upper respiratory tract of the mouse, to the surface of the epithelial lining. Cell nuclei in cyan. [3] [4]

A novel computational method, called GiniClust is used for detecting rare cell types from single-cell gene expression data with Gini index. It was applied to public single-cell RNA-seq datasets uncovers previously unrecognized rare cell types, including Zscan4-expressing cells within mouse embryonic stem cells and hemoglobin-expressing cells in the mouse cortex and hippocampus. GiniClust also correctly detects a small number of normal cells that are mixed in a cancer cell population.Based on the expression profile of the high Gini genes, cell clusters are identified using the algorithm density-based spatial clustering of applications with noise (DBSCAN). A nonlinear dimensionality reduction method, t-distributed stochastic neighbor embedding (t-SNE) is used to examine whether identified clusters are visually distinct. Differential gene expression analysis is used to identify the gene signature associated with each detected rare cell type [27].

Researchers at the University of Pennsylvania used single cell gene expression to make a comprehensive map of the mouse kidney. From nearly 58,000 cells, they identified 18 previously described kidney cell types, along with three novel cell populations. One new cell type was a group of transitional cells in the collecting duct. This showed that cells focused on sodium/potassium balance can switch, becoming cells focused on acid/base balance (and vice versa). They also mapped kidney disease-associated genes to specific cell types to help infer cellular drivers of diseases, such as proteinuria and metabolic acidosis [3].

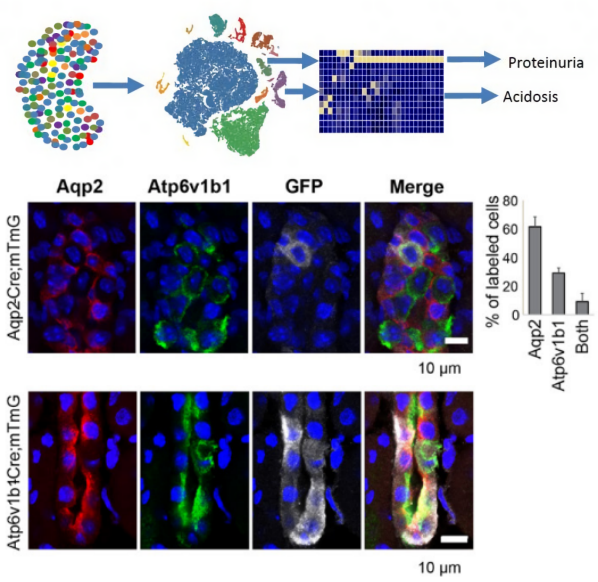

Figure 2.4: Determining kidney cell function by disease gene mapping. [5] [6]

Existing algorithms used to find rare cells scale turns out to be unbearably slowly or terminate, as the sample size grows to the order of tens of thousands. An algorithm, Finder of Rare Entities (FiRE), is proposed that assigns a rareness score to every individual expression profile under study in a matter of seconds. Further it is demonstrated how FiRE scores can help bioinformaticians focus the downstream analyses only on a fraction of expression profiles within ultra-large scRNA-seq data. When applied to a large scRNA-seq dataset of mouse brain cells, FiRE recovered a novel sub-type of the

pars tuberalis lineage. The efficacy of FiRE is also demonstrated in delineating human blood dendritic cell sub-types using 68k single-cell expression profiles of human blood cells. [28]

In 2019, a team at Max Planck Institute published a detailed map of cell populations in the human liver. They performed single cell gene expression profiling of over 10,000 liver cells from nine healthy donors. This high-resolution analysis revealed subtypes of endothelial cells, Kupffer cells, and hepatocytes with discrete gene expression profiles, which are believed to be previously unknown. They also identified a rare subpopulation of bile duct cells capable of forming organoids. This precursor cell type can differentiate into either hepatocytes or bile duct cells, and may play an important role in liver regeneration [29].

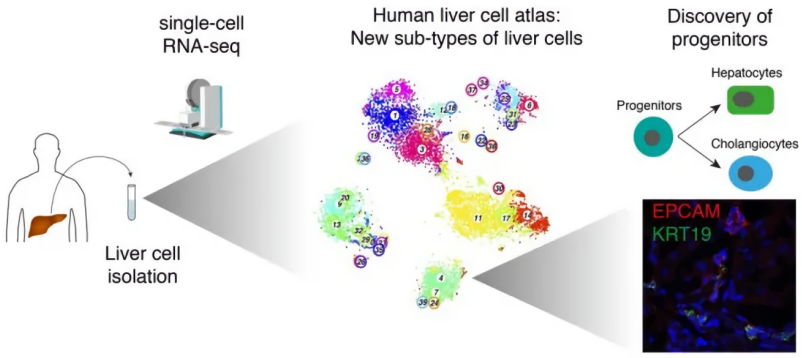

Figure 2.5: Using single cell RNA sequencing to create a comprehensive map of human liver cells [3]

Tirosh et al.identified cancer stem cells and their differentiated progeny through a single cell map in human oligodendroglioma [30]. The findings support the cancer stem cell hypothesis and confirm that cancer stem cells were the main source of growth for oligodendroglioma. These studies suggest that these cells may become new therapeutic targets. Researcher Suvà suggested that immunotherapy can be used to attack specific cell types, thereby terminating tumor growth. It is of great significance for the treatment

of such diseases. Nguyen et al. mapped the single cell pattern of human mammary epithelial cells by single-cell sequencing technology and analyzed the diversity and status of cell types in breast cells. The findings help to understand the early origins of breast cancer and provide the basis for improving the early detection of cancer and preventing cancer progression [19].

Chinese researchers at the single-cell level mapped the immunological map of the liver cancer microenvironment and the T cell immune map of lung cancer [31] [32]. The subgroup classification, tissue distribution characteristics, tumor heterogeneity and drug target gene expression of immune cells in liver cancer and lung cancer were revealed. The above studies are helpful to understand the immune microenvironment of liver cancer and lung cancer, to find effective biomarkers, new tumor immunotherapy, and drug targets. It is of great significance for the diagnosis and treatment of liver cancer and lung cancer [19].

Single-cell analysis is applied to systematically characterize the heterogeneity within leukemic cells using the MLL-AF9 driven mouse model of acute myeloid leukemia. Initially fluorescence-activated cell sorting analysis with seven surface markers, and extend by using a multiplexing quantitative polymerase chain reaction approach to assay the transcriptional profile of a panel of 175 carefully selected genes in leukemic cells at the single-cell level. By employing a set of computational tools like unsupervised hierarchical clustering, t-SNE analysis, SPADE analysis, two-sided Wilcoxon-Mann-Whitney rank sum test and weighted gene co-expression network analysis (WGCNA), striking heterogeneity within leukemic cells is found. Mapping to the normal hematopoietic cellular hierarchy identifies two distinct subtypes of leukemic cells; one similar to granulocyte/monocyte progenitors and the other to macrophage and dendritic cells. Further functional experiments suggest that these subtypes differ in proliferation rates and clonal phenotypes. Finally, co-expression network analysis reveals similarities as well as organizational differences between leukemia and normal granulocyte/monocyte progenitor

networks [7].

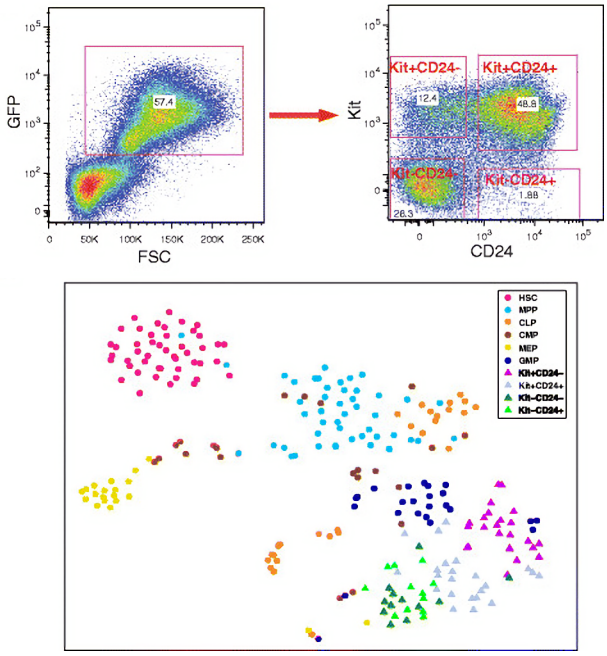

Figure 2.7: t-Distributed stochastic neighbor embedding (t-SNE) analysis of FACS and qPCR single-cell gene expression data [7]

Villani et al. identified multiple subtypes of DC cells and monocytes in human blood by single-cell RNA-seq and revealed a new subset of DC cells that have the properties of plasmacytoid DCs but are effective in activating T cells. In response to this finding, one of the researchers suggested "stimulating such cells or potentially enhancing the body's immune system to fight cancer." Such cells may become a new anti-cancer means to remove tumor cells through their own immune system, avoiding the damage of normal chemotherapy drugs to normal cells. The study redefines the relationship among DC cells, helps analyze the developmental processes and functions of the immune system, and completes immune surveillance under normal and disease states [33].

Xin et al. used single-cell RNA sequencing to study immune cells and cytokines during persistent infection. It was found that the heterogeneity of IL-10 expressing CD4 T cells and the production of IL-10 by a subset of helper cells during different infections

play an important role in promoting humoral immunity [34].

An ensemble method is proposed for simultaneous dimensionality reduction and feature gene extraction (EDGE) of scRNA-seq data. The proposed method implements an ensemble learning scheme that utilizes massive weak learners for an accurate similarity search. The datasets used in the study are Jurkat dataset, PBMC dataset and mouse brain dataset. Based on the similarity matrix constructed by those weak learners, the low-dimensional embedding of the data is estimated and optimized through spectral embedding and stochastic gradient descent. Comprehensive simulation and empirical studies show that EDGE is well suited for searching for meaningful organization of cells, detecting rare cell types, and identifying essential feature genes associated with certain cell types [8].

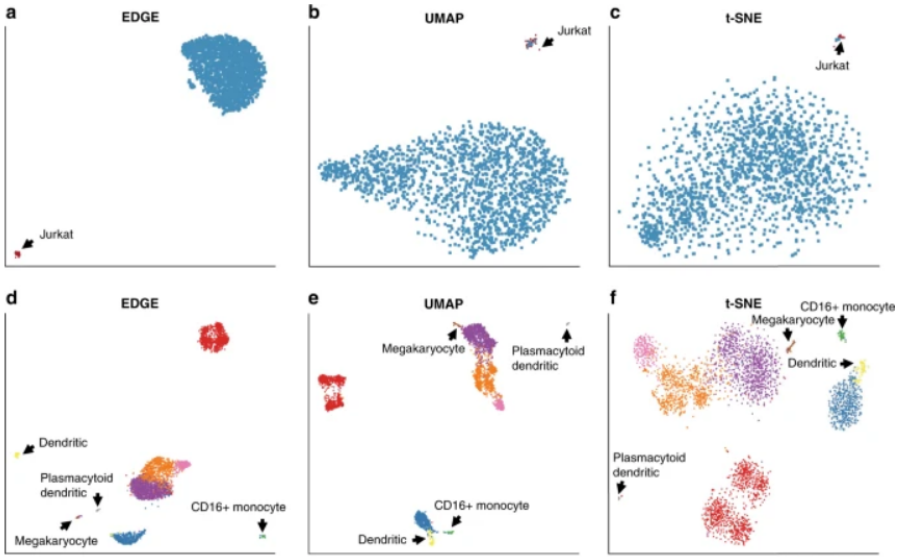

Figure 2.8: The two-dimensional embeddings learnt by EDGE, UMAP, and t-SNE [8]

Fan et al. identified multiple subpopulations of cells in each region in multiple regions of human metaphase embryos by single-cell sequencing techniques, and analyzed gene expression and neuronal maturation in these regions. Zhong et al. mapped a single-cell transcriptome map of human prefrontal embryonic development through single-cell

sequencing. From this map, the diversity of cell types in the prefrontal lobe of the human embryonic brain and the developmental relationship among different cell types were analyzed, and the molecular regulation mechanism of neuron production and loop formation was further revealed [35].

Single cell sequencing technology can detect the advantages of a small number of cells, which can be applied to prenatal diagnosis and assisted reproduction. Detection of female egg cells, polar cells or embryonic cells by single-cell sequencing to select healthy embryo transfer can reduce the birth rate of newborns with congenital genetic diseases and help prevent genetic diseases [36].

Lan et al. performed high-throughput single-cell genome sequencing on bacterial and fungal synthetic communities to analyze the distribution of antibiotic resistance genes, virulence factors and phage sequences in environmental microbial communities. These play an important role in the classification, evolution and drug resistance of microorganisms and finding ways to resist antimicrobial resistance [37].

In this paper, we develop a new algorithm, called GiniClust, for rare cell type detection and show that it outperforms RaceID for both simulated and biological datasets. The most important feature in GiniClust is a novel gene selection method that is particularly suitable for rare cell type identification, borrowing ideas from the social science domain. At first, a bidirectional Gini index is defined to identify genes that are specifically unexpressed in a rare cell type This extension is used only for qPCR data analysis but not for RNA-seq data analysis. Secondly, the Gini index values are normalized to remove a systematic bias toward lowly expressed genes. After normalization, the genes with highest Gini index values are selected for further analysis and referred to as high Gini genes. Based on the expression profile of the high Gini genes, cell clusters are identified by using the algorithm density-based spatial clustering of applications with noise DBSCAN. Two additional steps are added to interpret the clustering results. T-distributed stochastic neighbor embedding (t-SNE), a nonlinear dimensionality reduction

method, is used to examine whether identified clusters are visually distinct. Secondly, we differential gene expression analysis is done to identify the gene signature associated with each detected rare cell type [27].

Chapter 3

Methodology

Our proposed method for the detection of rare cells from single cell gene expression profile data is based on Optimized Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to segregate the outliers effectively.

3.1 Dataset Description

In this experiment, two different single-cell RNA-sequencing (scRNA-seq) datasets and a synthetic dataset were used to evaluate the performance of the algorithm.

- **Synthetic Data:** This data has been artificially generated and contains 505 samples. It is used to test the algorithm and model multiple outcomes.

- **ScRNA-seq Data:** This dataset contains expression profiles of 1872 cells. There are 13 different cell types annotated as I, I5d, II, II5d, III, III5d, IV, IV5d, LI, Lgr5SC, Lgr5SC2, V5d and X . As a result, original labels could also be used directly as ground truth annotation.

- **Usoskin Data:** RNA-Seq was performed on 799 dissociated single cells dissected from the mouse lumbar dorsal root ganglion distributed over a total of nine 96-well

plates. The data is available at the Gene Expression Omnibus under the accession number GSE59739.

3.2 Dataset Preprocessing

The raw scRNA-seq expression profiles are preprocessed as follows:

- **Data Filtering:** Cell filtering was not required in the data sets as there were few genes whose transcripts were not detected. The genes (if any) present with missing or Nan values were removed from further analysis owing to their very less number.
- **Median Normalization:** RobustScaler was applied on the datasets. It removes the median and scales the data according to the quantile range.
- **Log Transformation:** 1 is added as pseudocount to the median normalized count matrix followed by $\log_2$ transformation.
- **Dimensionality Reduction:** Dimensionality reductions methods such as PCA and t-SNE were applied on the datasets for feature extraction.

Single-cell RNA sequencing data are with a high dimensionality, which may involve thousands of genes and a large number of cells. Dimensionality reduction and feature selection are two main strategies for dealing with high dimensional ScRNA-seq data. Dimensionality reduction methods generally project the data into a lower dimensional space by optimally preserving some key properties of the original data.Feature selection removes the uninformative genes and identifies the most relevant features to reduce the number of dimensions used in downstream analysis. Feature selection can largely speed up the calculations of large-scale scRNA-seq data. Differential expression is a widely used method for feature selection in bulk RNA-seq experiments, but it is hard to apply to scRNA-seq data since the information of predetermined and/or homogeneous subpopulations needed for differential expression calling of scRNA-seq data is often unavailable.

Principal Component Analysis (PCA) is a linear dimensionality reduction technique. It transforms a set of correlated variables (p) into a smaller k (k < p) number of uncorrelated variables called principal components while retaining as much of the variation in the original dataset as possible.It is essential to perform feature scaling before running PCA if there is a significant difference in the scale between the features of the dataset.

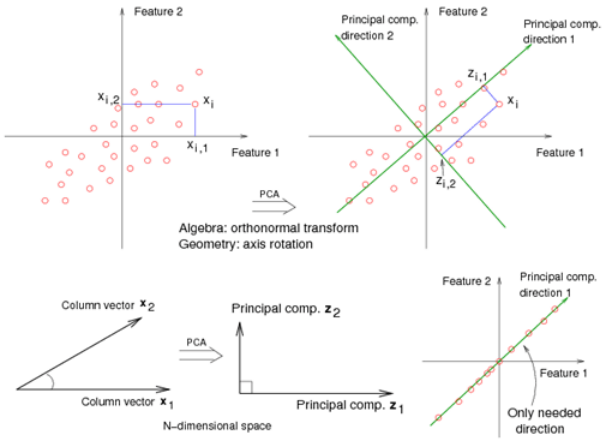

Figure 3.1: PCA applied on given data [9]

T-distributed Stochastic Neighbor Embedding (t-SNE) is a tool to visualize high-dimensional data. It converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. t-SNE has a cost function that is not convex, i.e. with different initializations we can get different results [38].

3.3 Clustering

The objective of of scRNA-seq data analysis is to identify cell subpopulations or distinct cell types within a certain tissue or condition in order to investigate cell heterogeneity. Unsupervised clustering methods such as K-means, DBSCAN etc can be used for de novo identification of cell populations with scRNA-seq data. KMeans is vulnerable to outliers. As the algorithm iterates through centroids, outliers have a significant impact

on the way the centroids moves before reaching stability and convergence. It cannot be applied on nonlinearly separable input data. On the other hand, DBSCAN does not require us to specify the number of clusters, detects outliers, and works quite well with arbitrarily shaped and sized clusters. It does not have centroids, the clusters are formed by a process of linking neighbor points together.

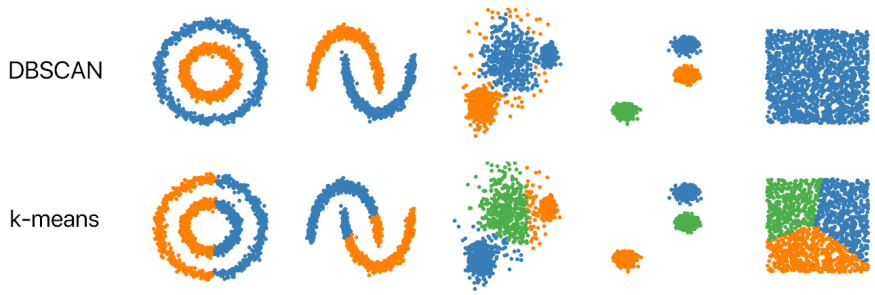

Figure 3.2: Data Clustering using DBSCAN and K-means [10]

The two required parameters for applying the DBSCAN algorithm are Epsilon and Minimum Points.Data points will be valid neighbors if their mutual distance is less than or equal to the specified epsilon. Minimum Points is the minimum number of data points within the radius of a neighborhood (ie. epsilon) for the neighborhood to be considered a cluster.

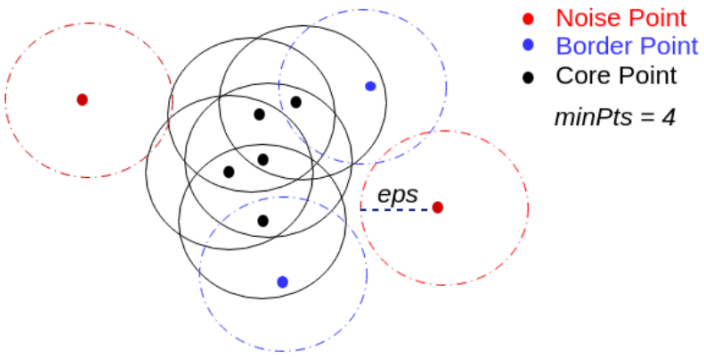

Figure 3.3: A single DBSCAN cluster with Core, Border, and Noise Points [11]

3.4 Optimization

Particle Swarm Optimization (PSO) is a stochastic optimization technique based on the movement and intelligence of swarms. It uses a number of particles that constitute a swarm moving around in the search space, looking for the best solution.

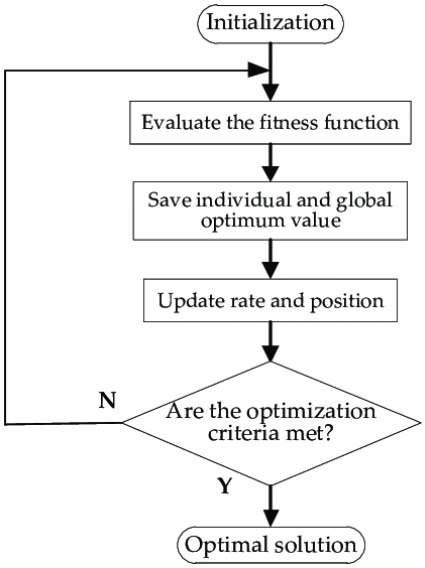

Figure 3.4: Particle Swarm Optimisation Algorithm [12]

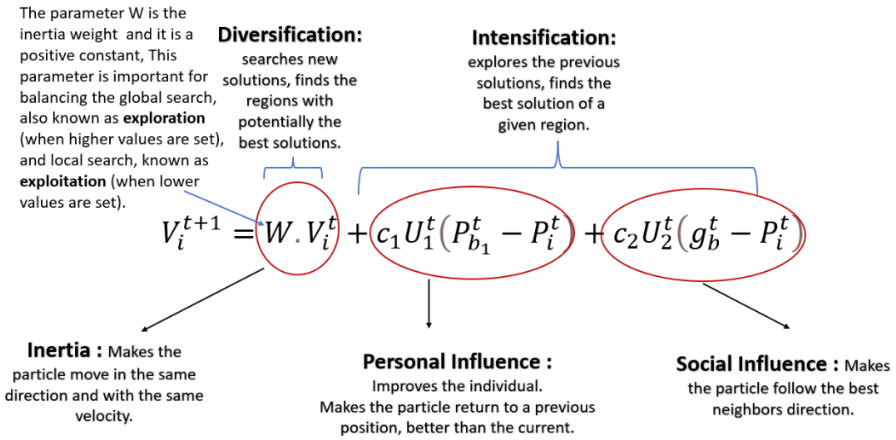

Figure 3.5: Particle Swarm Optimisation Algorithm [13]

Working of PSO:

1. To create a 'population' of agents (particles) which is uniformly distributed over X.

2. To evaluate each particle's position considering the objective function.

3. If a particle's present position is better than its previous best position, update it.

4. To find the best particle (according to the particle's last best places).

5. Update particles' velocities.

$$V_i^{t+1} = W.V_i^t + c_1 U_1^t (P_{b1}^t - P_i^t) + c_2 U_2^t (g_b^t - P_i^t)$$

6. To move particles to their new positions.

$$P_i^{t+1} = P_i^t + V_i^{t+1}$$

7. To go to step 2 until the stopping criteria are satisfied.

The optimal values for the DBSCAN parameters (epsilon and minimum points) to detect the rare cell types from the single cell gene expression profile are obtained using Particle Swarm Optimization executed for ten iterations.

3.5 Validation

Cluster validity indices are used for estimating the quality of partitions produced by clustering algorithms and for determining the number of clusters in data. Cluster validation is difficult task, because for the same data set more partitions exists regarding the level of details that fit natural groupings of a given data set. Even though several

cluster validity indices exist, they are inefficient when clusters widely differ in density or size [39].

- **Silhouette score:** If the ground truth labels are not known, evaluation must be performed using the model itself. A higher Silhouette Coefficient score relates to a model with better defined clusters. The Silhouette Coefficient is defined for each sample and is composed of two scores: a: The mean distance between a sample and all other points in the same class. b: The mean distance between a sample and all other points in the next nearest cluster.

  The Silhouette Coefficient s for a single sample is then given as:

$$s = \frac{b - a}{max(a, b)}$$

- **Davies-Bouldin index:** If the ground truth labels are not known, the Davies-Bouldin index can be used to evaluate the model, where a lower Davies-Bouldin index relates to a model with better separation between the clusters. This index signifies the average 'similarity' between clusters, where the similarity is a measure that compares the distance between clusters with the size of the clusters themselves. Zero is the lowest possible score. Values closer to zero indicate a better partition.

  $\mathbf{s_i}$, the average distance between each point of cluster and the centroid of that cluster, also know as cluster diameter.
  $\mathbf{d_{ij}}$, the distance between cluster centroids

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

  Then the Davies-Bouldin index is defined as:

$$DB = \frac{1}{k} \sum_{i=1}^{k} max_{i \neq j} R_{ij}$$

- **Calinski-Harabasz index:** If the ground truth labels are not known, the Calinski-Harabasz index also known as the Variance Ratio Criterion can be used to evaluate the model. A higher Calinski-Harabasz score relates to a model with better defined clusters. The index is the ratio of the sum of between-clusters dispersion and of within-cluster dispersion for all clusters where dispersion is defined as the sum of distances squared.

  For a set of data E of size n_E which has been clustered into k clusters, the Calinski-Harabasz score s is defined as the ratio of the between-clusters dispersion mean and the within-cluster dispersion:

$$s = \frac{tr(B_k)}{tr(W_k)} X \frac{n_E - k}{k - 1}$$

  where $tr(B_k)$ is trace of the between group dispersion matrix and $tr(W_k)$ is the trace of the within-cluster dispersion matrix defined by:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T$$

  with C_q the set of points in cluster q, c_q the center of cluster q, c_E the center of E , and n_q the number of points in cluster q.

- **Artificial Neural Network**: Artificial neural networks (ANNs) are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

They rely on training data to learn and improve their accuracy over time.

With the optimal values for the parameters of DBSCAN, the class labels for the samples are obtained. Then an ANN model is created, trained with 70% of the data and tested with the remaining. The validation accuracy is obtained.

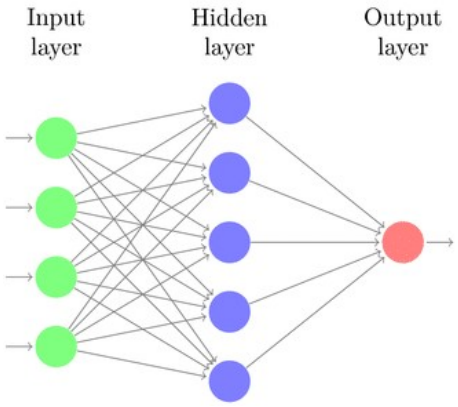

Figure 3.6: Artificial Neural Network [14]

Presence of good training data set results in better and accurate models. When the actual class labels of the samples are not unknown, DBSCAN clustering method is used for generating labels for the samples. The parameters are hyper-tuned using PSO to generate a good training data set with DBSCAN clustering labels. The presence of missing and outlier values in the training data often reduces the accuracy of a model or leads to a biased model. It leads to inaccurate predictions. This is because we don't analyse the behavior and relationship with other variables correctly. So, the outliers are then removed from the training data set. It is then used to train the model to improve the model's performance and get better results.

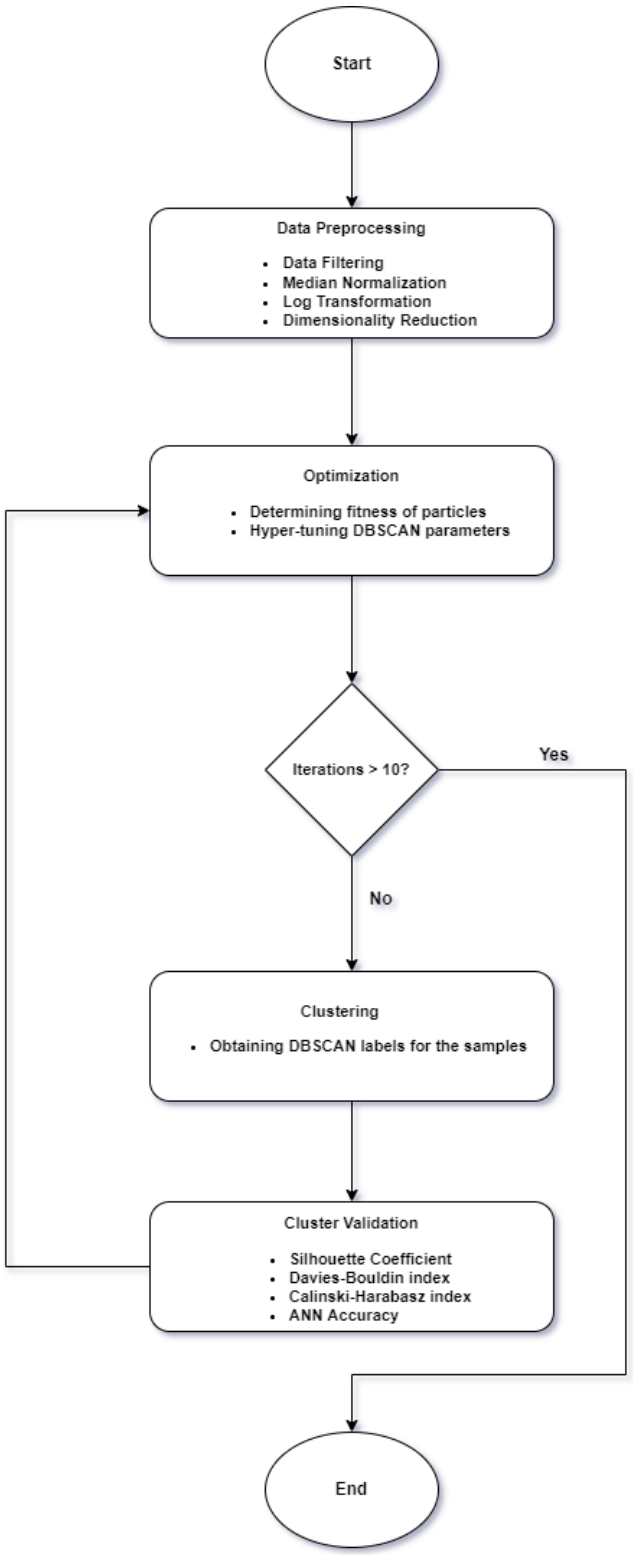

Figure 3.7: Overview of Methodology

Chapter 4

Results

4.1 Synthetic Data

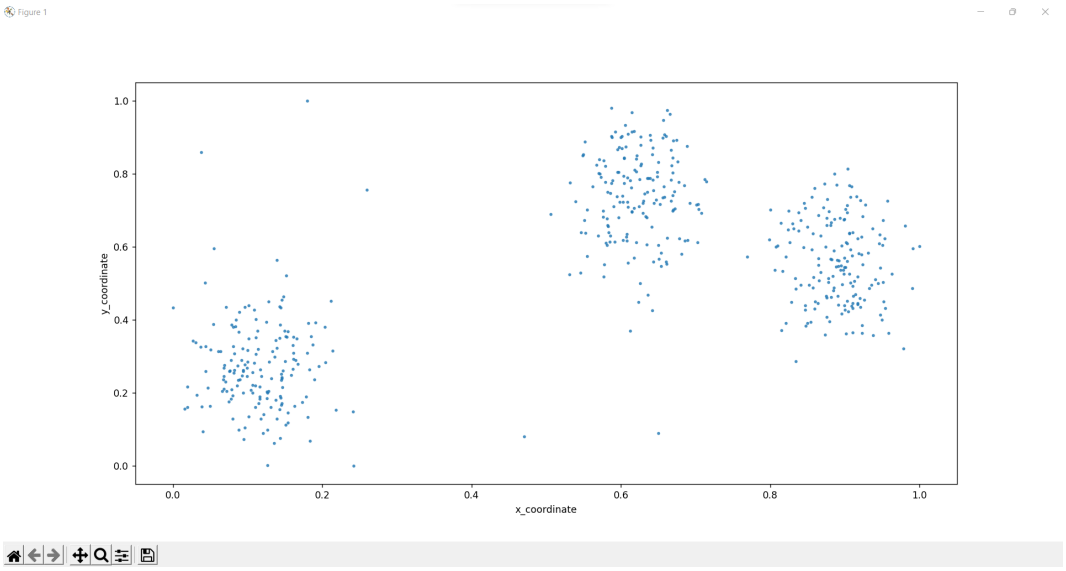

Figure 4.1: Synthetic Data PCA plot

Here the artificially generated data consists of 505 samples with three data clusters and five noisy data points. The cells or samples are arranged in rows and the genes are present in columns in the data frame.

4.1.1 Single-objective optimization (ANN Accuracy) with PCA

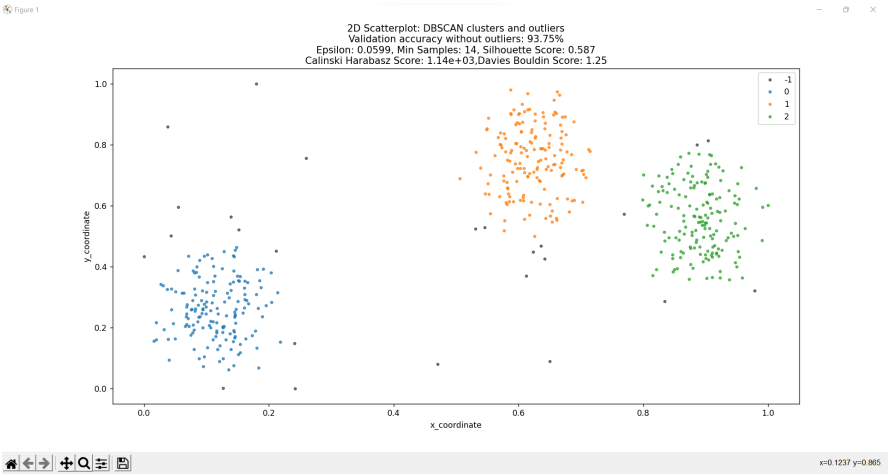

Figure 4.2: Synthetic Data with outliers using single-objective optimization (ANN Accuracy) with PCA

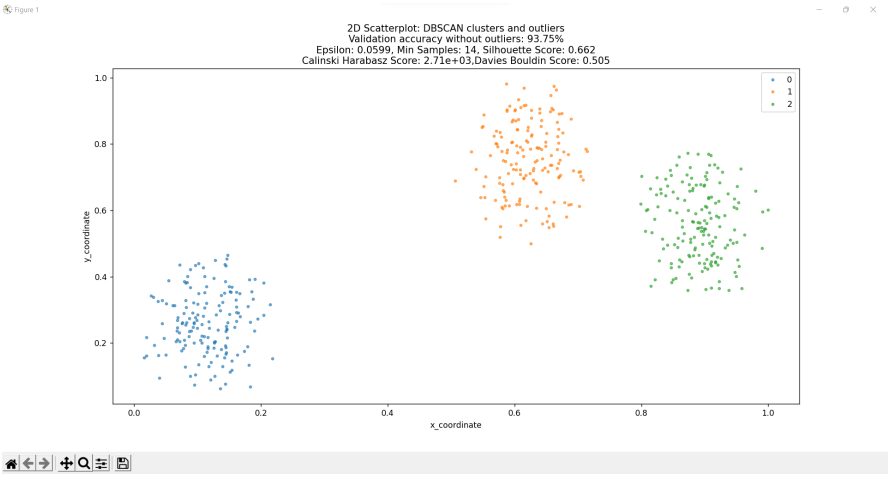

Figure 4.3: Synthetic Data without outliers using single-objective optimization (ANN Accuracy) with PCA

Here the ANN accuracy is considered as the fitness for different combinations of the DBSCAN parameters and is maximised using PSO executed for 10 iterations. The outliers or detected rare cells are labelled as -1 and represented as black dots. The other cells are groped together in different clusters with assigned numeric labels (starting from 0) and distinct colours (other than black).

4.1.2 Challenges with single-objective optimization (ANN Accuracy) with PCA

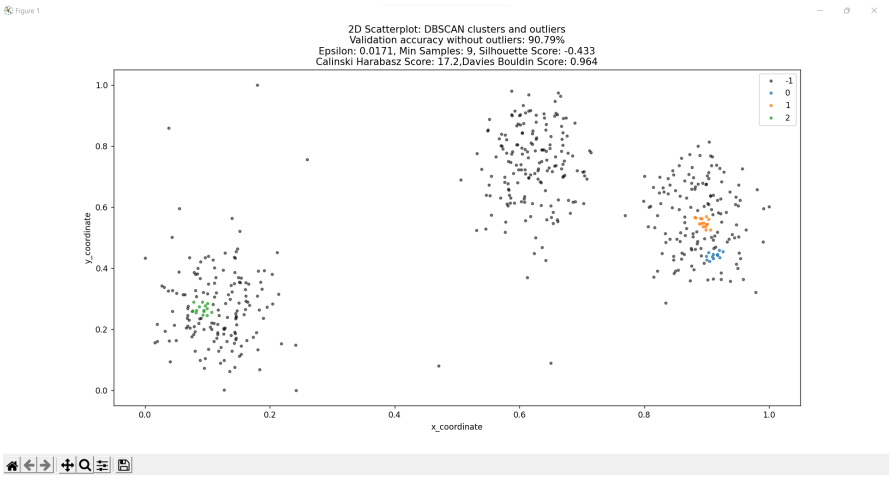

Figure 4.4: Synthetic Data with outliers using single-objective optimization (ANN Accuracy) with PCA

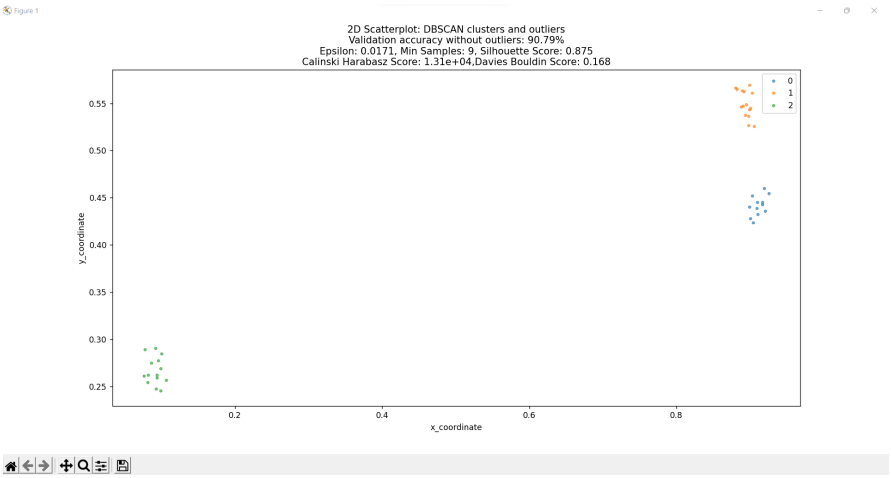

Figure 4.5: Synthetic Data without outliers using single-objective optimization (ANN Accuracy) with PCA

It is observed that considering only the ANN accuracy to determine the optimal DBSACN parameters also gives biased results. Although the ANN accuracy is over 90%, majority of the data points here are segregated as outliers. To overcome such poor outcomes, multi-objective optimisation along with cluster validation is to be performed.

4.1.3 Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA

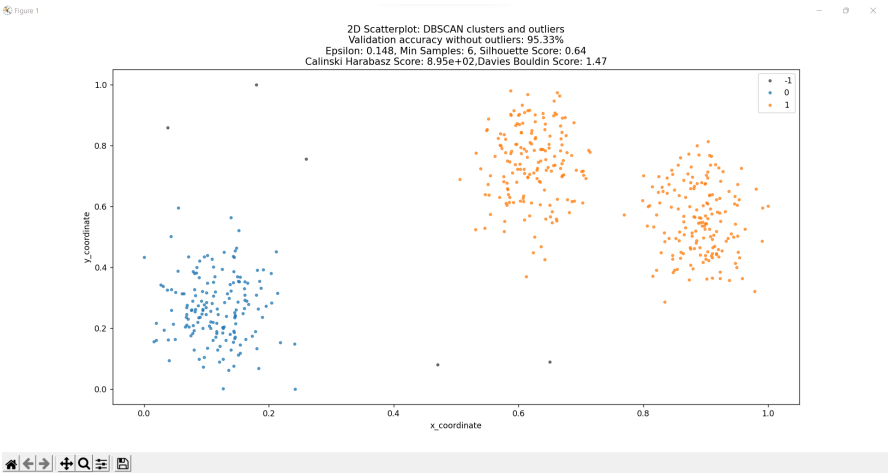

Figure 4.6: Synthetic Data with outliers using single-objective optimization (Silhouette Score & ANN Accuracy) with PCA

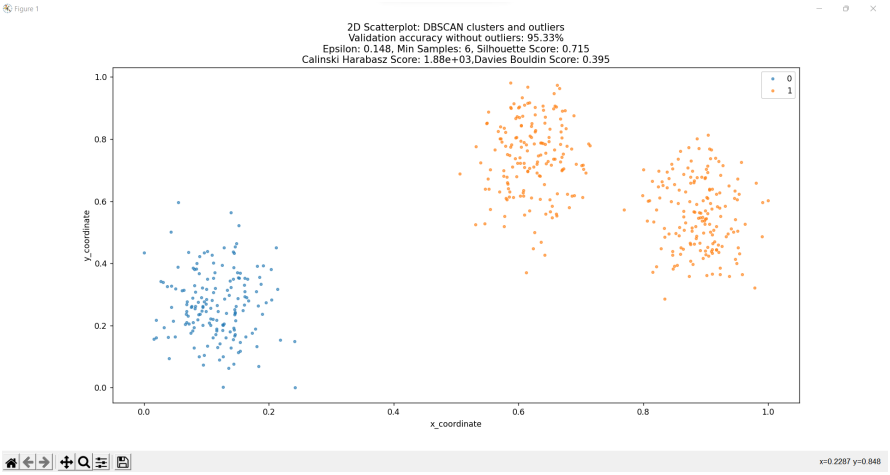

Figure 4.7: Synthetic Data without outliers using single-objective optimization (Silhouette Score & ANN Accuracy) with PCA

Here the average of ANN accuracy and Silhouette index is considered as the fitness for different combinations of the DBSCAN parameters and is maximised using PSO executed for ten iterations.

4.1.4 Single-objective optimization (Silhouette Score) with t-SNE

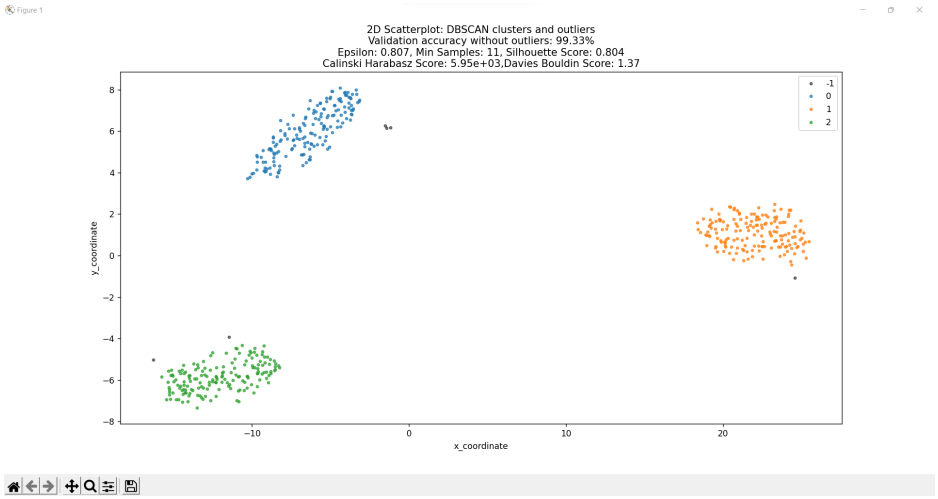

Figure 4.8: Synthetic Data with outliers single-objective optimization (Silhouette Score) with t-SNE

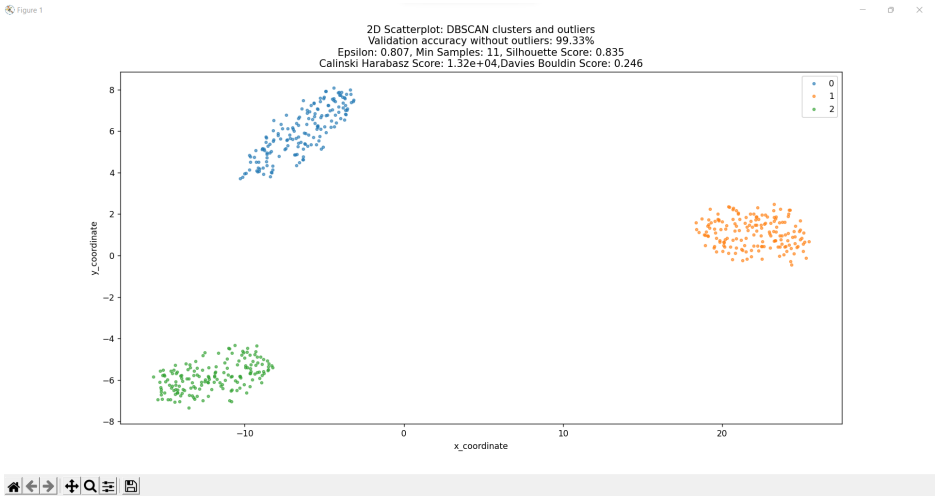

Figure 4.9: Synthetic Data without outliers single-objective optimization (Silhouette Score) with t-SNE

Here the x-coordinate and the y-coordinate represents t-SNE1 and t-SNE2 components respectively. The perplexity and number of iteration parameters are set to 100 and 5000.

4.2 ScRNA-Seq Data

	0	1	...	998	999
Unnamed: 0	Defa24__chr8	AY761184__chr8	...	Exoc3l4__chr12	Ndufb3__chr1
LI_1	3.017717	1.001958	...	0.0	4.031579
LI_2	0.0	0.0	...	0.0	0.0
LI_3	4.031579	0.0	...	2.007853	4.031579
LI_4	4.031579	1.001958	...	0.0	4.031579
...	...	...	...	...	...
I_92.3	439.433228	50.705898	...	0.0	0.0
I_93.3	256.481962	0.0	...	0.0	1.001958
I_94.3	1.001958	0.0	...	0.0	0.0
I_95.3	398.38551	115.565374	...	0.0	0.0
I_96.3	0.0	0.0	...	0.0	0.0

[1873 rows x 1000 columns]

Figure 4.10: ScRNA-Seq Data

Here the cells or samples are arranged in rows and the genes are present in columns. There are 1873 cells and 1000 genes as obtained from the transpose of the gene count matrix.

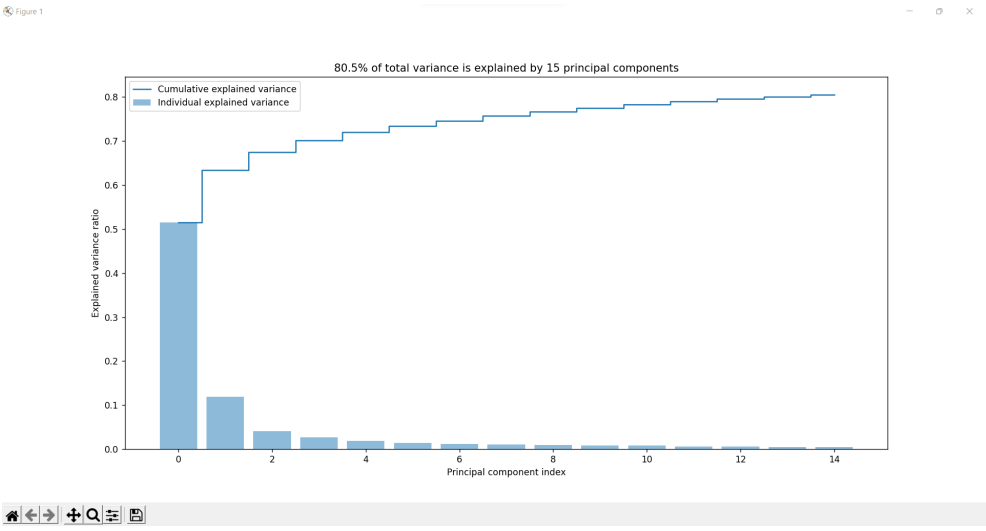

Figure 4.11: ScRNA-Seq Data variance plot

From the given variance plot, it is observed that 15 PCA components retains about 80% of the original data and these components have been used for further analysis.

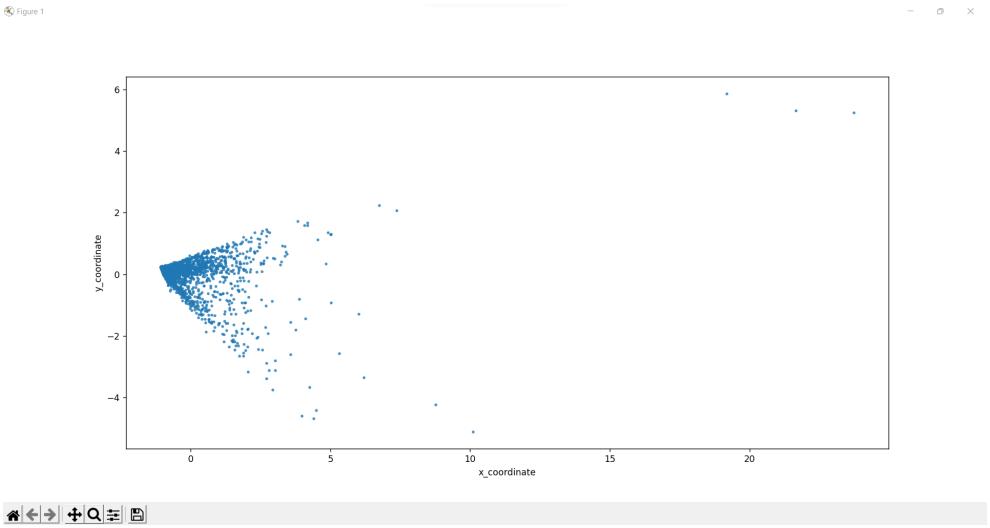

Figure 4.12: ScRNA-Seq Data PCA plot

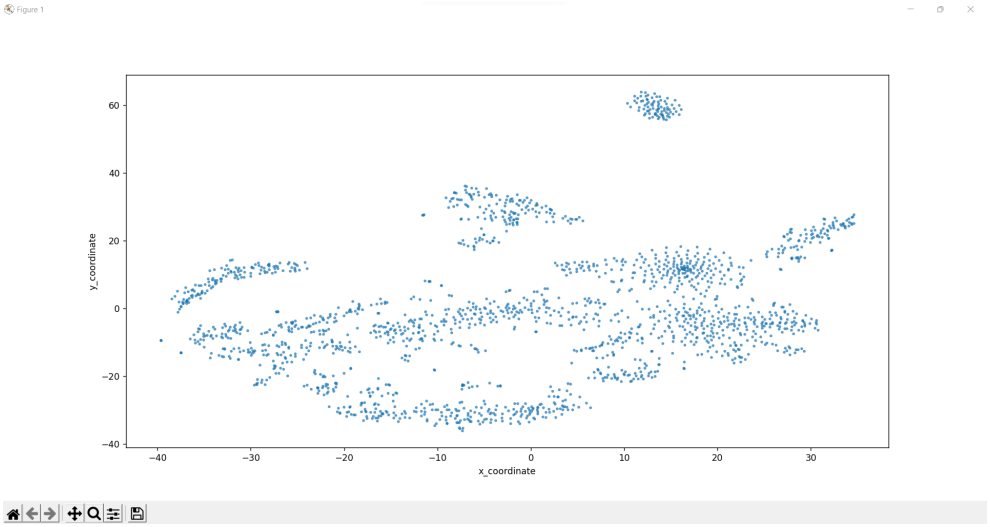

Figure 4.13: ScRNA-Seq Data t-SNE plot

The x-coordinate and the y-coordinate represents PCA1 or t-SNE1 and PCA2 or t-SNE2 components respectively.

4.2.1 Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA

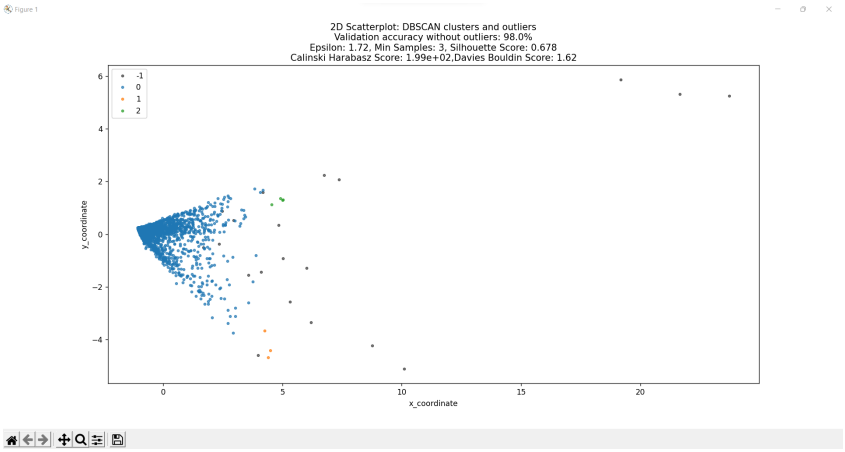

Figure 4.14: ScRNA-Seq Data with outliers using single-objective optimization (Silhouette Score & ANN Accuracy) with PCA

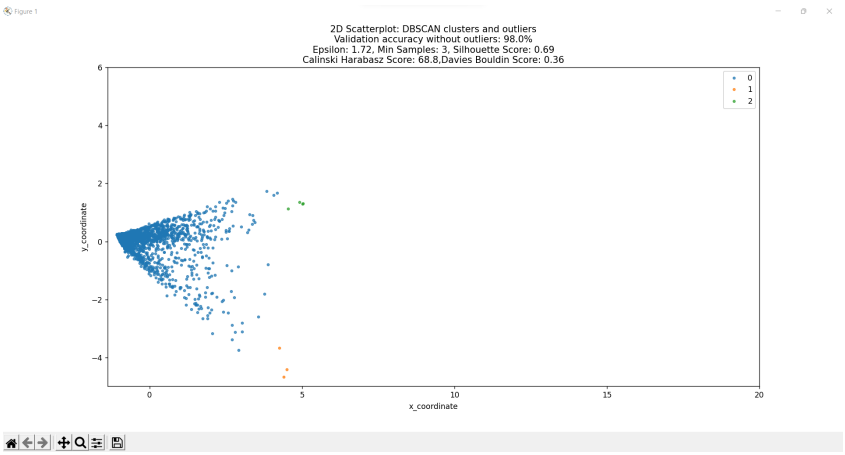

Figure 4.15: ScRNA-Seq Data without outliers using single-objective optimization (Silhouette Score & ANN Accuracy) with PCA

Here the the average of ANN accuracy and Silhouette index is considered as the fitness for different combinations of the DBSCAN parameters and is maximised using PSO executed for 10 iterations. The outliers or detected rare cells are labelled as -1 and represented as black dots. The other cells are groped together in different clusters with assigned numeric labels (starting from 0) and distinct colours (other than black).

4.2.2 Challenges with Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA

It is observed that PCA cannot be applied to find the structure to non-linear and sparse data distributions. While clustering is performed with the PCA components using DBSCAN to avoid the curse of dimensionality, most of the samples are still placed in one big cluster. This does not generate a good training data set and might lead to a biased result similar to that of single-objective optimization as the ANN is trained with the all the features and the obtained DBSCAN labels. Furthermore, it is observed that it takes a long time to obtain the final result as the ANN model is trained and tested for n*p times where n is the number of iterations and p is the population size for the particle swarm optimization.

Sl.No.	Optimization	Number of samples (cells)	Number of features (genes)	Execution Time (in s)
1	Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA	1873	1000	1165.44
2	Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA	1873	1000	1050.23
3	Single-objective optimization (Silhouette Score) with t-SNE	1873	1000	226.76
4	Single-objective optimization (Silhouette Score) with t-SNE	1873	1000	199.03

Table 4.1: Execution Time Analysis for ScRNA-Seq data set

Here Table 4.1 shows time analysis on the ScRNA-seq data. It is much smaller compared to other data sets and still the execution time turns out to be about 20 minutes. Therefore, it would take hours to determine the results with larger data sets.

4.2.3 Single-objective optimization (Silhouette Score) with t-SNE

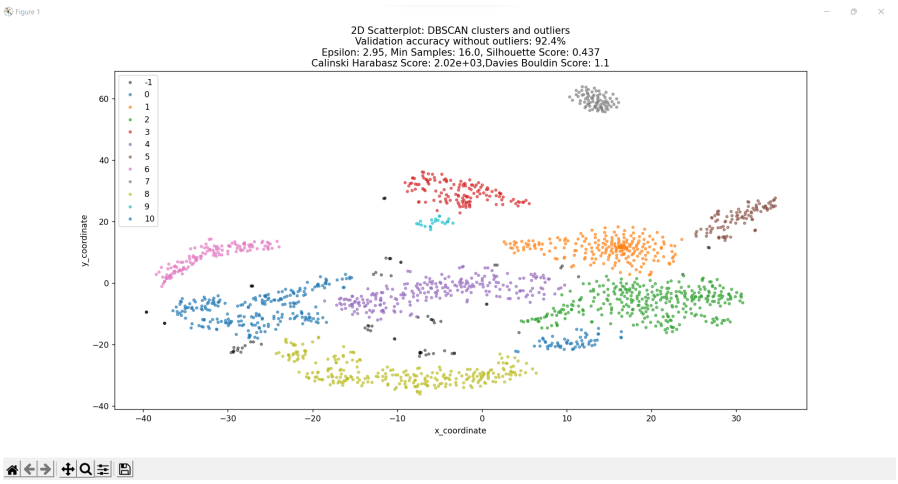

Figure 4.16: ScRNA-Seq Data with outliers with single-objective optimization (Silhouette Score) with t-SNE

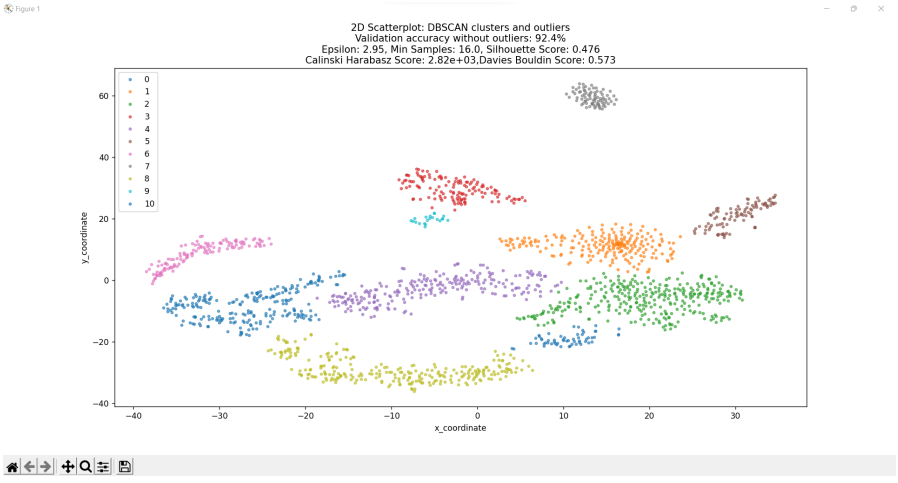

Figure 4.17: ScRNA-Seq Data without outliers single-objective optimization (Silhouette Score) with t-SNE

Here the Silhouette index is considered as the fitness for different combinations of the DBSCAN parameters and is maximised using PSO executed for ten iterations. The outliers or detected rare cells are labelled as -1 and represented as black dots. The other cells are groped together in different clusters with assigned numeric labels (starting from 0) and distinct colours (other than black).

4.3 Usoskin Data:

	0	1	2	...	25330	25331	25332
Unnamed: 0	Cd3d	Beta-s	Ubl5	...	r_tRNA-Arg-CGA_	r_U14	r_(CGTAG)n
L128_A01	0.0	0.0	0.0	...	0.0	0.0	0.0
L128_B01	0.0	0.0	723.81	...	0.0	0.0	0.0
L128_C01	0.0	0.0	1063.7	...	0.0	0.0	0.0
L128_D01	0.0	0.0	626.27	...	0.0	0.0	0.0
...	...	...	...	...	...	...	...
L282_D06	322740.0	0.0	765.7	...	0.0	0.0	0.0
L282_E06	2913.9	184490.0	10.091	...	0.0	0.0	0.0
L282_F06	557740	273	0	...	0	0	0
L282_G06	1306.5	0.0	0.0	...	0.0	0.0	0.0
L282_H06	762.95	0.0	0.0	...	0.0	0.0	0.0

[865 rows x 25333 columns]

Figure 4.18: Usoskin Data

Here the cells or samples are arranged in rows and the genes are present in columns. There are 864 cells and 25333 genes as obtained from the transpose of the gene count matrix.

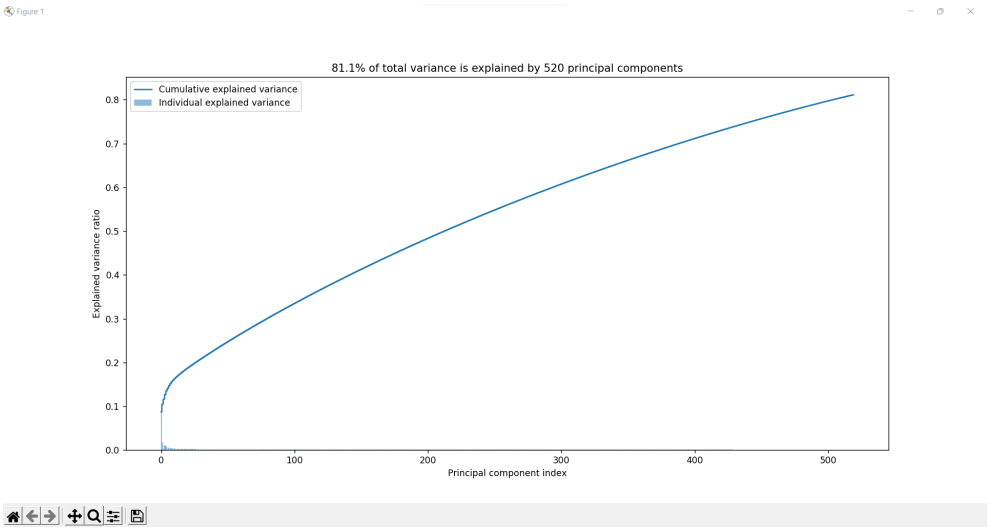

Figure 4.19: Usoskin Data variance plot

From the given variance plot, it is observed that 520 PCA components retains about 81.1% of the original data and these components have been used for further analysis.

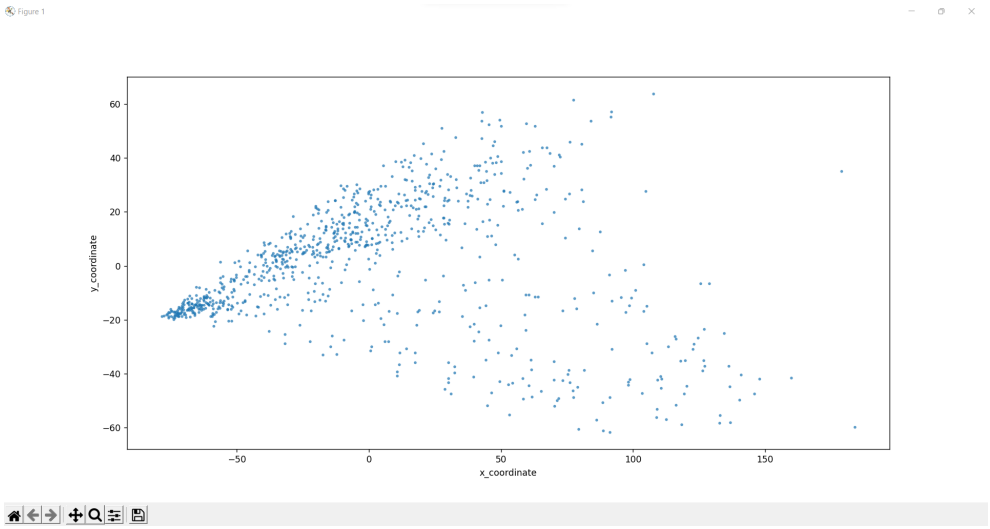

Figure 4.20: Usoskin Data PCA plot

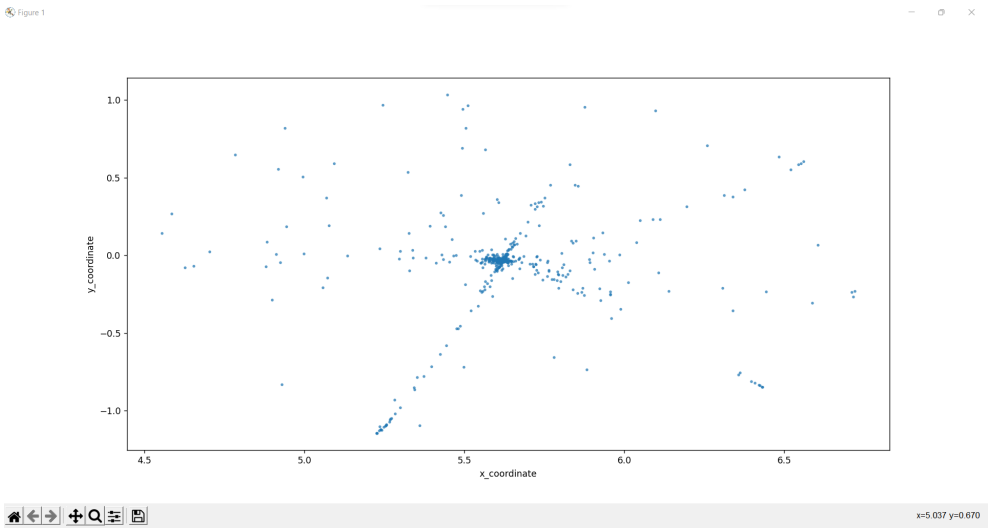

Figure 4.21: Usoskin Data t-SNE plot

The x-coordinate and the y-coordinate represents PCA1 or t-SNE1 and PCA 2 or t-SNE2 components respectively. The perplexity and number of iterations parameters are set to 800 and 5000.

4.3.1 Challenges with single-objective optimization (Silhouette Score & ANN Accuracy) with PCA

Multidimensional data analysis typically requires some components that are needed to be used to produce a visual encoding. Typical components are clustering approaches or projections from higher dimensional spaces into 2D or 3D visual spaces. Clustering models don't work well for data with large dimensions. They require the choice of an appropriate size of locality for density estimation. Using a too large size leads do not properly resolve the clusters as the clusters may not be separated. Hence, a small size is required. Due to the curse of dimensionality, clusters fall apart when using a too small size and one ends up with individual data points rather than clusters [40].

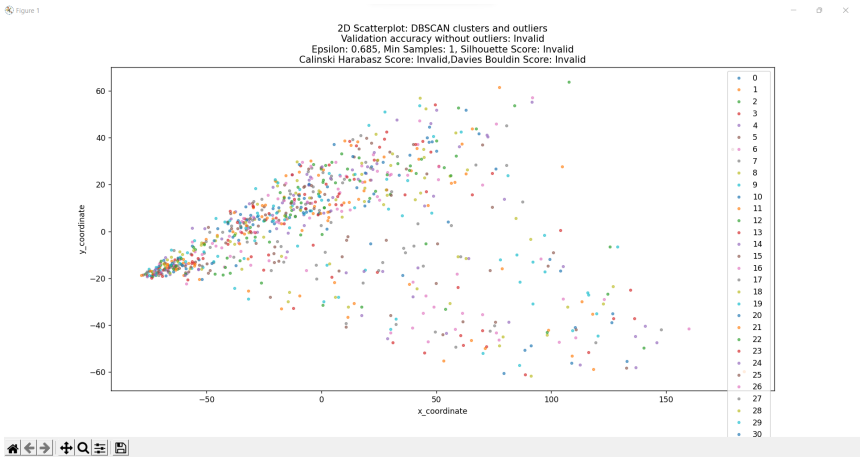

Figure 4.22: Usoskin Data with outliers with single-objective optimization (Silhouette Score & ANN Accuracy) with PCA

The curse of dimensionality is observed in the PCA plot of Usoskin Data. It is observed that PCA cannot be applied to find the structure to the non-linear and sparse data distributions. So t-SNE can be used instead as it groups the data points based on their similarity with features. It then tries to minimize the difference between these conditional probabilities between data points in higher-dimensional and lower-dimensional space for a perfect representation of data points. Also, the differencially expressed genes out of the features can be identified and used for further analysis.

4.3.2 Single-objective optimization (Silhouette Score) with t-SNE

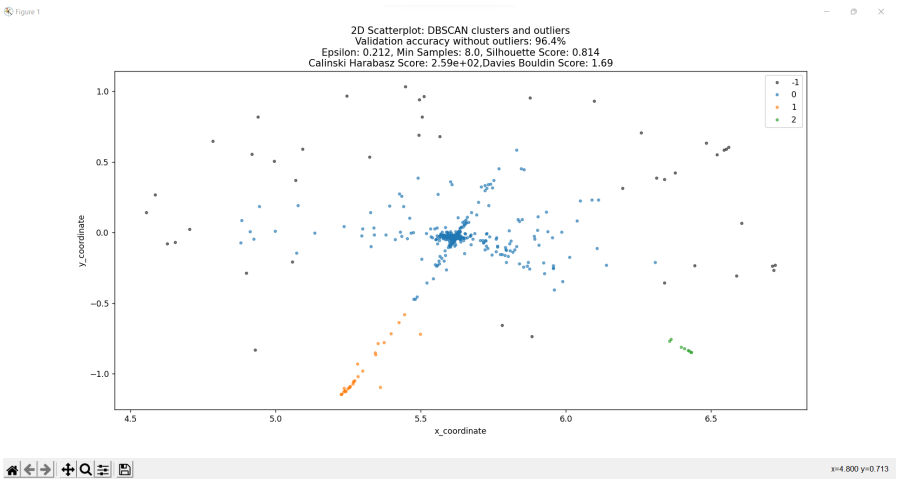

Figure 4.23: Usoskin Data with outliers single-objective optimization (Silhouette Score) with t-SNE

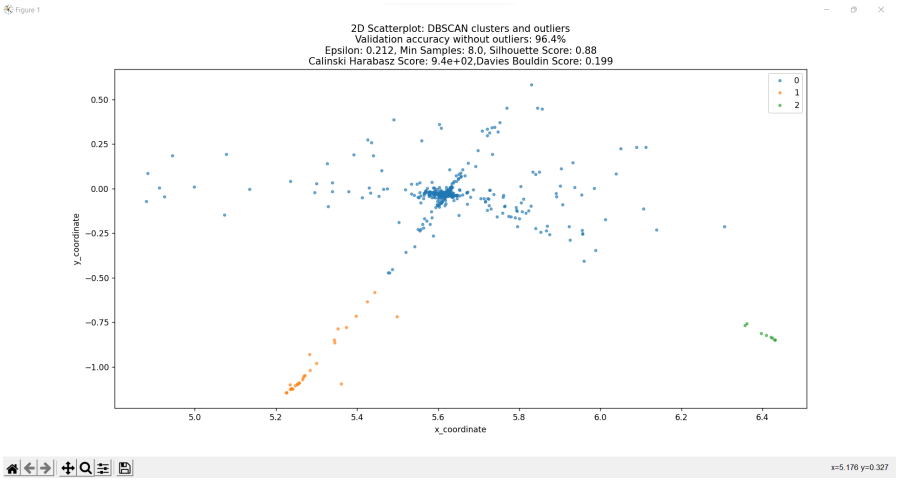

Figure 4.24: Usoskin Data without outliers single-objective optimization (Silhouette Score) with t-SNE

Here the Silhouette index is considered as the fitness for different combinations of the DBSCAN parameters and is maximised using PSO executed for ten iterations. The outliers or detected rare cells are labelled as -1 and represented as black dots. The other cells are groped together in different clusters with assigned numeric labels (starting from 0) and distinct colours (other than black).

4.4 Resultant Cluster Validation

4.4.1 Silhouette Index

Data set	Optimization	Silhouette Score with outliers	Silhouette Score without outliers	Actual Number of clusters	Number of clusters obtained without outliers
Synthetic Data	Single-objective optimization (ANN Accuracy) with PCA	0.587	0.662	3	3
Synthetic Data	Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA	0.64	0.715	3	2
Synthetic Data	Single-objective optimization (Silhouette Score) with t-SNE	0.804	0.835	3	3
ScRNA-seq Data	Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA	0.678	0.69	13	3
ScRNA-seq Data	Single-objective optimization (Silhouette Score) with t-SNE	0.437	0.476	13	11
Usoskin Data	Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA	Invalid	Invalid	13	864
Usoskin Data	Single-objective optimization (Silhouette Score) with t-SNE	0.814	0.88	4	3

Table 4.2: Silhouette Indices of the data sets used

4.4.2 Davies-Bouldin Index

Data set	Optimization	Davies-Bouldin index with outliers	Davies-Bouldin index without outliers	Actual Number of clusters	Number of clusters obtained without outliers
Synthetic Data	Single-objective optimization (ANN Accuracy) with PCA	1.25	0.305	3	3
Synthetic Data	Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA	1.47	0.395	3	2
Synthetic Data	Single-objective optimization (Silhouette Score) with t-SNE	1.37	0.246	3	3
ScRNA-seq Data	Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA	1.62	0.36	13	3
ScRNA-seq Data	Single-objective optimization (Silhouette Score) with t-SNE	1.1	0.573	13	11
Usoskin Data	Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA	Invalid	Invalid	13	864
Usoskin Data	Single-objective optimization (Silhouette Score) with t-SNE	1.69	0.199	4	3

Table 4.3: Davies-Bouldin Indices of the data sets used

4.4.3 Calinski-Harabasz Index

Data set	Optimization	Calinski-Harabasz index with outliers	Calinski-Harabasz insex without outliers	Actual Number of clusters	Number of clusters obtained without outliers
Synthetic Data	Single-objective optimization (ANN Accuracy) with PCA	1.14e+03	2.71e+03	3	3
Synthetic Data	Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA	8.95e+02	1.88e+03	3	2
Synthetic Data	Single-objective optimization (Silhouette Score) with t-SNE	5.95e+03	1.32e+04	3	3
ScRNA-seq Data	Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA	1.99e+02	68.8	13	3
ScRNA-seq Data	Single-objective optimization (Silhouette Score) with t-SNE	2.02e+03	2.82e+03	13	11
Usoskin Data	Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA	Invalid	Invalid	13	864
Usoskin Data	Single-objective optimization (Silhouette Score) with t-SNE	2.59e+02	9.4e+02	4	3

Table 4.4: Calinski-Harabasz Indices of the data sets used

4.4.4 ANN Accuracy

Data set	Optimization	ANN Accuracy without outliers	Actual Number of clusters	Number of clusters obtained without outliers
Synthetic Data	Single-objective optimization (ANN Accuracy) with PCA	93.75%	3	3
Synthetic Data	Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA	95.33%	3	2
Synthetic Data	Single-objective optimization (Silhouette Score) with t-SNE	99.30%	3	3
ScRNA-seq Data	Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA	98%	13	3
ScRNA-seq Data	Single-objective optimization (Silhouette Score) with t-SNE	92.4%	13	11
Usoskin Data	Single-objective optimization (Silhouette Score & ANN Accuracy) with PCA	Invalid	13	864
Usoskin Data	Single-objective optimization (Silhouette Score) with t-SNE	96.4%	4	3

Table 4.5: ANN Accuracy of the data sets used

Here Table 4.2 shows the silhouette scores with and without outliers, the actual number of clusters and the number of clusters obtained on removing the detected outliers with reference to the different types of optimizations done on the three different data sets used for the experiment. The best value is 1 and the worst value is -1. Negative values generally indicate that a sample has been assigned to the wrong cluster. For all the above cases, it is observed that the silhouette scores are nearly above 0.5 and further increases on removing the detected outliers.

Here Table 4.3 shows the Davies-Bouldin indices with and without outliers, the actual number of clusters and the number of clusters obtained on removing the detected outliers with reference to the different types of optimizations done on the three different data sets used for the experiment. The minimum score is zero and lower values indicating better clustering. For all the above cases, it is observed that the Davies-Bouldin indices are nearly about 0.5 and further decreases on removing the detected outliers.

Here Table 4.4 shows the Calinski-Harabasz indices with and without outliers, the actual number of clusters and the number of clusters obtained on removing the detected outliers with reference to the different types of optimizations done on the three different data sets used for the experiment. The higher the score , the better the performances. For all the above cases except multi-objective optimization with PCA on ScRNA-seq data, it is observed that the Calinski-Harabasz indices increases on removing the detected outliers.

Here Table 4.5 shows ANN accuracies with and without outliers, the actual number of clusters and the number of clusters obtained on removing the detected outliers with reference to the different types of optimizations done on the three different data sets used for the experiment. The accuracies are determined after removing the detected outliers as these tend to be the noisy samples not belonging to any of the clusters. For all the above cases, it is observed that the ANN accuracy is over 92% and is not affected with the number of clusters obtained.

Chapter 5

Conclusion

From, the experimental results, it is observed that the rare cells can be more accurately segregated with increased Silhouette score. Further it can be inferred that when both the testing accuracy and Silhouette index are considered for evaluating the fitness and hyper-tuning the DBSCAN parameters, the results improve. The testing accuracy obtained for the datasets using ANN is above 92% for all the data sets used. The Silhouette index and the Davies-Bouldin index are nearly 0.5 in all the cases after discarding the detected outliers, which is a decent score for each of the parameters.

However, since that data sets are quite large, it takes a long time when ANN accuracy is also considered for evaluation of clusters. It is observed that PCA cannot be applied to find the structure to the non-linear and sparse data distributions so t-SNE is used instead. Single-objective optimisation for silhouette index is done to get faster and unbiased results. On removing the outliers detected through DBSCAN from the data sets not only the Silhouette index and Calinski-Harabasz index increases but also the Davies-Bouldin index decreases.

The novelty of the research work done lies in the determination of optimal DBSCAN parameters for clustering using Particle Swarm Optimization (PSO) as such parameters have been experimentally determined so far [27, 28]. The future work includes the in-

dentification and analysis of differentially expressed genes from the gene-count matrix as these genes represent potential molecular biomarkers. The DEGs would be used for the investigation of the rare cells from the large and sparse ScRNA-Seq data sets. This would aid in reducing the dimensionality to catch useful indicators and obtain a more accurate result. Also, the fitness function can be further modified to perform multi-objective optimization that would improve the functioning of the proposed algorithm and make it work on multiple varied data sets.

The continuous innovation of scRNA-seq technologies and advances in bioinformatics approaches will largely facilitate transcriptomic studies, leading to insightful findings in gene expression variability, cell dynamics, disease diagnosis, precision medicine and other biological and clinical researches etc. It has enabled robust analysis of millions of single cells in an efficient manner with high throughput. The use of above mentioned computational techniques will further produce accurate, cost-effective and faster results than the traditional wet-lab experimentation.

References

[1] "The path to sequencing nucleic acids," https://www.whatisbiotechnology.org/index.php/exhibitions/sanger/path.

[2] "Welcome to the world of single-cell rna-sequencing," https://speakerdeck.com/stephaniehicks/welcome-to-the-world-of-single-cell-rna-sequencing?slide=3.

[3] "Molecular Biology of the Cell," https://www.10xgenomics.com/blog/discovering-rare-cell-types-with-single-cell-rna-seq.

[4] D. T. Montoro, A. L. Haber, M. Biton, V. Vinarsky, B. Lin, S. E. Birket, F. Yuan, S. Chen, H. M. Leung, J. Villoria *et al.*, "A revised airway epithelial hierarchy includes cftr-expressing ionocytes," *Nature*, vol. 560, no. 7718, pp. 319–324, 2018.

[5] J. Park, R. Shrestha, C. Qiu, A. Kondo, S. Huang, M. Werth, M. Li, J. Barasch, and K. Suszták, "Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease," *Science*, vol. 360, no. 6390, pp. 758–763, 2018.

[6] "Massive single-cell survey of kidney cell types reveals new paths to disease," https://www.eurekalert.org/news-releases/726691.

[7] A. Saadatpour, G. Guo, S. H. Orkin, and G.-C. Yuan, "Characterizing heterogeneity in leukemic cells using single-cell gene expression analysis," *Genome biology*, vol. 15, no. 12, pp. 1–13, 2014.

[8] X. Sun, Y. Liu, and L. An, "Ensemble dimensionality reduction and feature gene extraction for single-cell rna-seq data," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[9] "Applied Data Mining and Statistical Learning," https://onlinecourses.science.psu.edu/stat857, 2018.

[10] Andrewngai, "Understanding DBSCAN Algorithm and Implementation from Scratch," https://towardsdatascience.com/understanding-dbscan-algorithm-and-implementation-from-scratch-c256289479c5, 2020.

[11] A. Srivastava, "DBSCAN Clustering Algorithm," https://blog.knoldus.com/dbscan-clustering-algorithm/, 2021.

[12] Y. Xiao, Y. Wang, and Y. Sun, "Reactive power optimal control of a wind farm for minimizing collector system losses," *Energies*, vol. 11, p. 3177, 11 2018.

[13] S. Sanyal, "An Introduction to Particle Swarm Optimization (PSO) Algorithm," https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-particle-swarm-optimization-algorithm/, 2021.

[14] G. Holmgren, P. Andersson, A. Jakobsson, and A. Frigyesi, "Artificial neural networks improve and simplify intensive care mortality prognostication: a national cohort study of 217,289 first-time intensive care unit admissions," *Journal of intensive care*, vol. 7, no. 1, pp. 1–8, 2019.

[15] S. Mallik and Z. Zhao, "Multi-objective optimized fuzzy clustering for detecting cell clusters from single-cell expression profiles," *Genes*, vol. 10, no. 8, p. 611, 2019.

[16] "Molecular Biology of the Cell," https://www.ncbi.nlm.nih.gov/books/NBK26887/.

[17] "Dna, rna and protein synthesis."

[18] E. Hedlund and Q. Deng, "Single-cell rna sequencing: technical advancements and biological applications," *Molecular aspects of medicine*, vol. 59, pp. 36–46, 2018.

[19] X. Tang, Y. Huang, J. Lei, H. Luo, and X. Zhu, "The single-cell sequencing: new developments and medical applications," *Cell & Bioscience*, vol. 9, no. 1, pp. 1–9, 2019.

[20] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz *et al.*, "Eleven grand challenges in single-cell data science," *Genome biology*, vol. 21, no. 1, pp. 1–35, 2020.

[21] B. Hwang, J. H. Lee, and D. Bang, "Single-cell rna sequencing technologies and bioinformatics pipelines," *Experimental & molecular medicine*, vol. 50, no. 8, pp. 1–14, 2018.

[22] G. Chen, B. Ning, and T. Shi, "Single-cell rna-seq technologies and related computational data analysis," *Frontiers in genetics*, vol. 10, p. 317, 2019.

[23] A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg, "A practical guide to single-cell rna-sequencing for biomedical research and clinical applications," *Genome medicine*, vol. 9, no. 1, pp. 1–12, 2017.

[24] S. M. Shaffer, M. C. Dunagin, S. R. Torborg, E. A. Torre, B. Emert, C. Krepler, M. Beqiri, K. Sproesser, P. A. Brafford, M. Xiao *et al.*, "Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance," *Nature*, vol. 546, no. 7658, pp. 431–435, 2017.

[25] S. Vaga, "Understanding Single Cell Sequencing, How It Works and Its Applications," https://www.technologynetworks.com/genomics/articles/understanding-single-cell-sequencing-how-it-works-and-its-applications-357578, 2020.

[26] R. Wegmann, M. Neri, S. Schuierer, B. Bilican, H. Hartkopf, F. Nigsch, F. Mapa, A. Waldt, R. Cuttat, M. R. Salick *et al.*, "Cellsius provides sensitive and specific detection of rare cell populations from complex single-cell rna-seq data," *Genome biology*, vol. 20, no. 1, pp. 1–21, 2019.

[27] L. Jiang, H. Chen, L. Pinello, and G.-C. Yuan, "Giniclust: detecting rare cell types from single-cell gene expression data with gini index," *Genome biology*, vol. 17, no. 1, pp. 1–13, 2016.

[28] A. Jindal, P. Gupta, D. Sengupta *et al.*, "Discovery of rare cells from voluminous single cell expression data," *Nature communications*, vol. 9, no. 1, pp. 1–9, 2018.

[29] Q. H. Nguyen, N. Pervolarakis, K. Blake, D. Ma, R. T. Davis, N. James, A. T. Phung, E. Willey, R. Kumar, E. Jabart *et al.*, "Profiling human breast epithelial cells using single cell rna sequencing identifies cell diversity," *Nature communications*, vol. 9, no. 1, pp. 1–12, 2018.

[30] I. Tirosh, A. S. Venteicher, C. Hebert, L. E. Escalante, A. P. Patel, K. Yizhak, J. M. Fisher, C. Rodman, C. Mount, M. G. Filbin *et al.*, "Single-cell rna-seq supports a developmental hierarchy in human oligodendroglioma," *Nature*, vol. 539, no. 7628, pp. 309–313, 2016.

[31] X. Guo, Y. Zhang, L. Zheng, C. Zheng, J. Song, Q. Zhang, B. Kang, Z. Liu, L. Jin, R. Xing *et al.*, "Global characterization of t cells in non-small-cell lung cancer by single-cell sequencing," *Nature medicine*, vol. 24, no. 7, pp. 978–985, 2018.

[32] C. Zheng, L. Zheng, J.-K. Yoo, H. Guo, Y. Zhang, X. Guo, B. Kang, R. Hu, J. Y. Huang, Q. Zhang *et al.*, "Landscape of infiltrating t cells in liver cancer revealed by single-cell sequencing," *Cell*, vol. 169, no. 7, pp. 1342–1356, 2017.

[33] A.-C. Villani, R. Satija, G. Reynolds, S. Sarkizova, K. Shekhar, J. Fletcher, M. Griesbeck, A. Butler, S. Zheng, S. Lazo *et al.*, "Single-cell rna-seq reveals new types of human blood dendritic cells, monocytes, and progenitors," *Science*, vol. 356, no. 6335, 2017.

[34] G. Xin, R. Zander, D. M. Schauder, Y. Chen, J. S. Weinstein, W. R. Drobyski, V. Tarakanova, J. Craft, and W. Cui, "Single-cell rna sequencing unveils an il-10-producing helper subset that sustains humoral immunity during persistent infection," *Nature communications*, vol. 9, no. 1, pp. 1–14, 2018.

[35] X. Fan, J. Dong, S. Zhong, Y. Wei, Q. Wu, L. Yan, J. Yong, L. Sun, X. Wang, Y. Zhao *et al.*, "Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell rna-seq analysis," *Cell research*, vol. 28, no. 7, pp. 730–745, 2018.

[36] Y. Hou, W. Fan, L. Yan, R. Li, Y. Lian, J. Huang, J. Li, L. Xu, F. Tang, X. S. Xie *et al.*, "Genome analyses of single human oocytes," *Cell*, vol. 155, no. 7, pp. 1492–1506, 2013.

[37] F. Lan, B. Demaree, N. Ahmed, and A. R. Abate, "Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding," *Nature biotechnology*, vol. 35, no. 7, pp. 640–646, 2017.

[38] "T-distributed Stochastic Neighbor Embedding." https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html.

[39] K. R. Žalik, "Cluster validity index for estimation of fuzzy clusters of different sizes and densities," *Pattern Recognition*, vol. 43, no. 10, pp. 3374–3390, 2010.

[40] V. Molchanov and L. Linsen, "Overcoming the curse of dimensionality when clustering multivariate volume data." in *VISIGRAPP (3: IVAPP)*, 2018, pp. 29–39.