HELLENIC REPUBLIC
National and Kapodistrian
University of Athens
EST. 1837

(2023), 1–17

Term Project – Final Report

Machine Learning in Computational Biology - ML pipeline to classify retinal ganglion cells in subclasses based on their gene expression profile.

Marianna Papadopoulou, ID:7115152200022
Department of Informatics and Telecommunications, University of Athens, Greece

Thalassini-Marina Filippidou, ID:7115152200032
Department of Informatics and Telecommunications, University of Athens, Greece

Charalampos Vossos, ID:7115152200037
Department of Informatics and Telecommunications, University of Athens, Greece

Abstract

The present study introduces a classification-based methodology for the analysis of single-cell RNA sequencing (scRNA-seq) data. An attempt was made to reproduce the findings of a prior investigation and ascertain the existence of three distinct subclasses within a cellular system. A pipeline for classification was constructed based on the clustering results of the authors. Feature selection techniques were utilized in this study to optimize the number and selection of features. ANOVA and mRMR were employed for this purpose. Additionally, hyperparameter optimization of the classifiers was performed using stratified 5-fold cross-validation. The utilized classifiers encompassed Random Forests, XGBoost, Support Vector Machines (SVC), Logistic Regression, and Gaussian Naive Bayes (GaussianNB). The performance metric utilized for our optimization procedures was the Matthews Correlation Coefficient (MCC) score. The classifiers demonstrated encouraging outcomes in employing a data-centric methodology for addressing the issue, without relying on marker genes. The comparative analysis unveiled non-marker genes that were found to have significant roles in the classification process, consequently encouraging additional investigation into their biological significance. This study showcases the capacity of gene expression profiling and classification methodologies in comprehending cellular heterogeneity within the context of single-cell RNA sequencing (scRNA-seq) analysis.

Keywords: retinal ganglion cells, classification, scRNA-seq, computational biology, machine learning

1. Introduction

The utilization of single-cell RNA sequencing (scRNA-seq) has become a prominent technique in the field of genomics, facilitating the investigation of cellular diversity with unprecedented resolution. The analysis of single-cell RNA sequencing (scRNA-seq) data allows for the examination of gene expression patterns at the level of individual cells, thereby offering significant insights into the identification and functional implications of distinct cellular subtypes. The objective of this study is to make a valuable contribution to the field of scRNA-seq analysis by introducing an enhanced classification-based methodology for the identification and characterization of unique cell populations.

Prior research has effectively employed scRNA-seq analysis to identify distinct subgroups within particular cellular systems. The identification of these subclasses is commonly based on marker genes that demonstrate distinct patterns of differential expression. Although marker gene-based approaches have demonstrated efficacy, they may fail to consider crucial genes that make substantial contributions to the classification process.

Driven by this constraint, we started upon the replication of a seminal investigation that posited the existence of three subclasses within a particular cellular context. The main aim of our study was investigating alternative approaches that could improve the accuracy of classification and offer further understanding of cellular heterogeneity.

In order to achieve these objectives, the problem was framed as a classification task. A comprehensive pipeline was constructed, incorporating multiple essential components. Initially, feature selection techniques such as ANOVA and minimum Redundancy Maximum Relevance (mRMR) were utilized to ascertain the most informative features from the gene expression profiles. The purpose of this step was to decrease the dimensionality of the data and prioritize the genes that have the greatest impact on the classification process. Subsequently, the pipeline was optimized through the identification of the most suitable number and choice of features, with the aim of maximizing the performance of the classification process. The optimization process encompassed the assessment of multiple classifiers, including Random Forests, XGBoost, Support Vector Machines (SVM), Logistic Regression, and Gaussian Naive Bayes. The process of hyperparameter optimization was conducted by employing stratified 5-fold cross-validation, which was chosen to guarantee a reliable and unbiased selection of the model.

In the evaluation conducted, a variety of performance metrics were utilized, encompassing balanced accuracy, precision, recall, and the F1-score. Our study specifically concentrated on the optimization of the Matthews Correlation Coefficient (MCC), which incorporates the consideration of true positives, true negatives, false positives, and false negatives. The MCC (Matthews Correlation Coefficient) demonstrates a high level of suitability for imbalanced datasets, which are frequently encountered in scRNA-seq analysis.

The outcomes of our classification analysis exhibited significant promise. Interestingly, our classifiers often attributed significance to genes that were not identified as marker genes in the original study. The observed discrepancy served as motivation for us to conduct additional research on these non-marker genes in order to clarify their potential biological significance.

In summary, the primary objective of this study is to enhance the field of scrns analysis through the introduction of an enhanced classification-based methodology. The pipeline employed in our study integrates feature selection methodologies, optimization algorithms, and diverse classifiers in order to attain precise classification of cell populations. By questioning the exclusive reliance on marker genes, we open up new possibilities for exploring the diverse cellular heterogeneity that is captured by single-cell RNA sequencing (scRNA-seq) data. The subsequent sections will present a comprehensive description of our methodology, findings, and the biological knowledge acquired through the classification procedure.

2. Paper Discussion

2.1 Issue investigated in the article

The paper of choice [1] focuses on the Retinal ganglion cells (RGCs). Retrieval of retinal ganglion cells (RGCs) from living donors has posed a significant challenge, impeding the study of these cells’ crucial role in the transmission of visual information from the eye to the brain. Nonetheless, RGCs remain integral to this process. Pluripotent stem cells (PSCs) have been identified as a viable source of retinal ganglion cells (RGCs), and a specific methodology has been established to induce the differentiation of human PSCs into fully functional RGCs. The PSC-derived retinal ganglion cells (RGCs) exhibit similarities to sensory neurons and cells originating from the ganglion cell layer, as evidenced by previous studies. The research paper provides a collection of single-cell RNA sequencing data to enhance the understanding of the transcriptome of Retinal Ganglion Cells (RGCs) that are obtained from human embryonic stem cells (hESCs). The proposed research has the potential to enhance our comprehension of retinal ganglion cells (RGCs) and their implication in pathologies that result in visual impairment, such as glaucoma and optic neuropathies.

2.2 Importance-state of the art of chosen research

The aforementioned research holds significance for multiple reasons. Initially, the statement underscores the prospective utility of pluripotent stem cells (PSCs) in the fields of regenerative medicine and cell replacement therapies. The ability to differentiate pluripotent stem cells (PSCs) into distinct cell lineages, such as retinal ganglion cells (RGCs), holds promise for the treatment of diseases that lead to the degeneration of these cells resulting to permanent vision loss. The study highlights the efficacy of single-cell RNA sequencing (scRNA-seq) technology in delineating the transcriptome of retinal ganglion cells (RGCs) at an individual cellular level. This methodology enables the analysis of infrequent cellular populations and the dissection of cellular composition in apparently homogenous tissues or cell cultures.. Through this approach, an in-depth understanding of the molecular mechanisms underlying human tissues and diseases is achieved, surpassing the limitations of bulk RNA-seq investigations. The present study makes a valuable contribution to the ongoing endeavors of the field to gain a deeper comprehension of retinal ganglion cells (RGCs) in both healthy and pathological human tissue. This has been a challenging task owing to the unavailability of noninvasive techniques for procuring these cells from living donors. The scRNA-seq dataset presented in this study provides a comprehensive characterization of the transcriptome of Retinal Ganglion Cells (RGCs) that are derived from human Embryonic Stem Cells (hESCs).

Since its inception in 2006, induced pluripotent stem cell (PSC) technology has demonstrated reliability in tissue replacement and holds significant potential for innovative therapeutic treatments. Novel methodologies have been established to differentiate retinal ganglion cells (RGCs) from pluripotent stem cells (PSCs), constituting a significant progress in this field. Furthermore, gene-editing methodologies have been implemented to simulate and investigate illnesses. Additional methodologies encompass the utilization of tri-dimensional retinal organoids, which are capable of reproducing the configuration and operation of the retina, and the implementation of optogenetics to scrutinize the interconnectivity of RGCs and their photic stimulus reactions. In general, the current state of research in this field entails an amalgamation of methodologies, encompassing stem cell biology, molecular biology, genetics, and neuroscience, to investigate retinal ganglion cells (RGCs) and associated disorders, with the primary objective of creating new therapies and remedies for these medical conditions.

2.3 Data analysis methods – tools and results

2.3.1 Preprocessing

The initial stage of the study consisted of the development of a cell quality matrix utilizing four distinct data types, namely library size (i.e., total mapped reads), total number of detected genes, percentage of reads mapped to mitochondrial genes, and percentage of reads mapped to ribosomal genes. Cells exhibiting any of the four parameter measurements below 3 times the median absolute deviation (MAD) of all cells were identified as outliers and subsequently excluded from further analysis. Furthermore, cells exhibiting mitochondrial reads exceeding 20% or ribosomal reads surpassing 50% were excluded from the analysis. Subsequently, genes with a frequency of occurrence of less than 1% across all cells were eliminated to eliminate genes that may have been detected due to stochastic noise. Subsequently, the expression data underwent normalization at two distinct levels in order to mitigate any potential systematic bias that may exist between the samples and cells. Prior to data aggregation, the initial stage of normalization involved the utilization of the cellranger aggr depth equalization method. This method involved the subsampling of mapped reads from higher-depth libraries until the number of mapped reads per library were equal. The aforementioned approach mitigates the likelihood of confounding effects that may arise due to variations in sequencing depths across samples. The application of the deconvolution approach by Lun et al. facilitated the execution of the second level of normalization. The methodology employed takes into consideration the sparsity of expression data by aggregating expression counts from cell clusters and mitigates potential bias arising from technical factors such as cDNA synthesis, PCR amplification efficacy, and sequencing depth for individual cells. The size factors that were normalized for each group were further deconvoluted into size factors specific to each cell, which were subsequently utilized to adjust the counts of individual cells. After normalization, abundantly expressed ribosomal protein genes and mitochondrial genes were discarded.

2.3.2 Clustering

The study's authors aimed to detect and eliminate a subset of cells that exhibited suboptimal sequence data quality, which had evaded initial filtration. The authors employed an unsupervised clustering approach and conducted enrichment analysis of genes exhibiting differential expression. Initially, the researchers employed Principal Component Analysis (PCA) to decrease the dimensionality of the transcript count table. This was carried out on the top 1,500 most variable genes. The researchers preserved the initial 20 principal components and employed them to partition cells according to their transcript count profiles. In order to attain clustering with high resolution that is capable of detecting small subpopulations and outliers, the authors implemented a bottom-up agglomerative hierarchical clustering methodology that constructs a dendrogram tree. The Ward's minimum distance method was employed to cluster cells into a dendrogram, thereby facilitating the grouping of cells that exhibit similar characteristics. An unsupervised method was employed to amalgamate branches into subpopulations, resulting in the partitioning of the dendrogram tree into 40 height-windows and the production of 40 distinct clustering outcomes. The most stable clustering outcome was determined as the optimal one among a series of successive tree-height values. The authors conducted pairwise differential expression analysis utilizing a negative binomial test and a general linear model, as outlined in the DESeq package, to delineate the identified clusters. Reactome functional interaction analysis was utilized to conduct network analysis on the differentially expressed genes that were deemed significant.

2.3.3 Results

The research discovered noticeable gene expression patterns that suggest varying degrees of maturation and differentiation. The genes associated with neural cell adhesion molecule signaling and Hedgehog pathway were found to be upregulated in subpopulation one, indicating a possible state of progenitor or early differentiation. The second subpopulation exhibited an increase in gene expression related to the regulation of Notch protein, neuronal function and development, and DNA repair, suggesting a more specialized phenotype of retinal ganglion cells. The third subpopulation exhibited an upregulation of genes related to axon guidance and extracellular matrix proteoglycans, while also displaying a significant downregulation of genes associated with the cell cycle, indicating a more mature neuronal phenotype. The investigation additionally recognized genes that are expressed in diverse subcategories of retinal ganglion cells (RGCs), which may indicate the existence of a minimum of nine distinct RGC subtypes.

3. Motivation of additional analysis

The retina is a complex and specialized tissue that plays a crucial role in the process of vision. The retinal ganglion cells are a crucial cellular subtype in the retina that performs a vital function of relaying visual signals from the eye to the brain. Comprehending the molecular mechanisms that regulate the development and operation of retinal ganglion cells can offer valuable insights into diverse retinal pathologies, including glaucoma and age-related macular degeneration. Single-cell RNA sequencing is a frequently employed technique to delineate the molecular diversity of cell populations. Nevertheless, this method can be both costly and time intensive. Hence, the creation of a classifier capable of precisely forecasting the cell type of novel cells founded on their gene expression profiles, which can also be obtained through various methods such as bulk RNA sequencing, microarray analysis, or other techniques, may serve as a cost-effective and efficient substitute for scRNAseq. This is particularly relevant in scenarios where conducting scRNAseq on each individual cell is impractical or unnecessary. The aforementioned methodology has the potential to optimize the procedure of characterizing cellular populations and furnish a robust instrument for investigating retinal ganglion cells and other cellular subtypes within the retina. In addition, the development of a classifier aimed at identifying said subpopulations may aid in the examination of forthcoming experiments and enhance our comprehension of the biology of retinal ganglion cells, potentially impacting the diagnosis and management of ocular ailments.

4. Replication of results

Prior to presenting the experimental findings related to the classification problem, we outline our attempt to reproduce the outcomes of the clustering analysis as reported in the original paper. The process of replication holds significant importance in scientific research as it enables the validation and verification of the outcomes obtained from prior studies. In order to ensure the reliability and accuracy of our subsequent analyses, we tried to conduct a thorough replication of the clustering results using the same dataset and clustering approach employed by the original authors. In this section, we provide a comprehensive description of our replication methodology and analyze its potential impacts.

With the goal to reproduce the clustering outcomes presented in the original research paper, we initiated the process by acquiring the normalized results file that was shared by the authors. Given the computational requirements and time limitations involved in reproducing the entire process from the beginning, we decided to employ the normalized outcomes as a foundation. Following the methodology proposed by the authors, we conducted a filtering procedure, wherein we selected and retained the top 1500 genes exhibiting the highest degree of variability. This selection was made with the intention of prioritizing the most pertinent and informative characteristics of the dataset.

In the subsequent phase of replication, similar to the initial study, Principal Component Analysis (PCA) was employed to diminish the dimensionality of the feature space. Given the considerable number of features, the initial 20 principal components were extracted in order to preserve the most essential information while simultaneously mitigating computational complexity. The resulting clustering analyses were based on these 20 principal components.

Subsequently, we proceeded to the clustering process utilizing the agglomerative algorithm Ward, as suggested in the original research article. In order to investigate the influence of various dimensionality reduction methods on the performance of clustering, a series of clustering experiments were conducted using three distinct approaches: UMAP, t-SNE (initialized with the 20 principal components), and plain PCA.

The UMAP and t-SNE algorithms are widely acknowledged for their efficacy in visualizing data with high dimensionality, while simultaneously maintaining the integrity of both local and global structure. In contrast, we utilized plain Principal Component Analysis (PCA) as a simple method for reducing dimensionality, serving as a baseline for comparison.

In this study, our objective was to assess the clustering performance of different approaches, particularly the Ward algorithm in conjunction with UMAP, t-SNE, and plain PCA. We sought to identify the approach that produces the most optimal clustering outcomes for our dataset. The replication of the clustering results from the original paper is of great importance as it serves as a fundamental reference point for the subsequent analyses. Presented below are the obtained results, our attention was directed towards the clusterings that shared the same number of clusters as the ones we utilized in our classification experiments.

4.1 2 clusters results

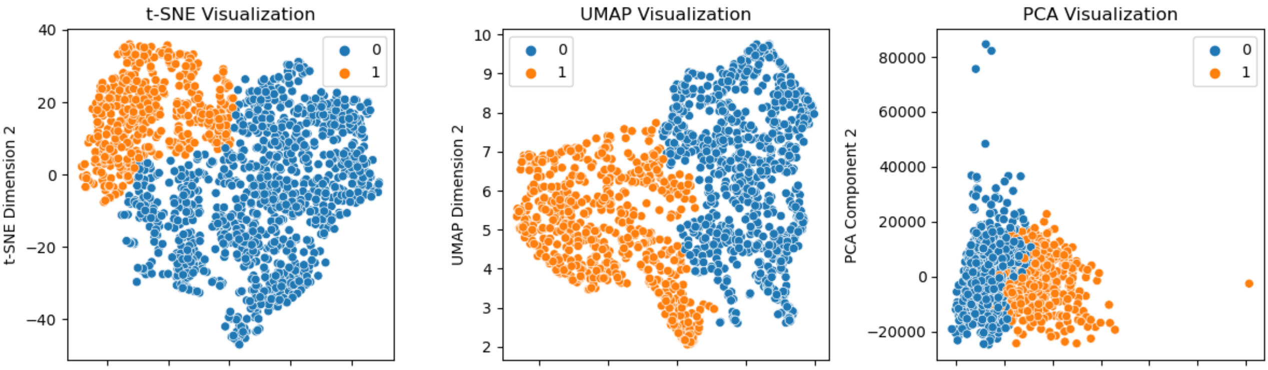


Figure 1. 2D visualizations of clustering results using Ward algorithm and different dimensionality techniques (2 clusters)

COMPARISON WITH AUTHORS' RESULTS (2 clusters)		
t-SNE	**UMAP**	**PCA**
Number of common elements in ***cluster 0****:* 988 out of 1332	Number of common elements in ***cluster 0***: 786 out of 1332	Number of common elements in ***cluster 0***: 956 out of 1332
Number of common elements in ***cluster 1***: 71 out of 95	Number of common elements in ***cluster 1***: 74 out of 95	Number of common elements in ***cluster 1***: 74 out of 95

Table 1. Comparison of our study's results with the authors' results, based on common observation inside the formed clusters

4.2 4 clusters results

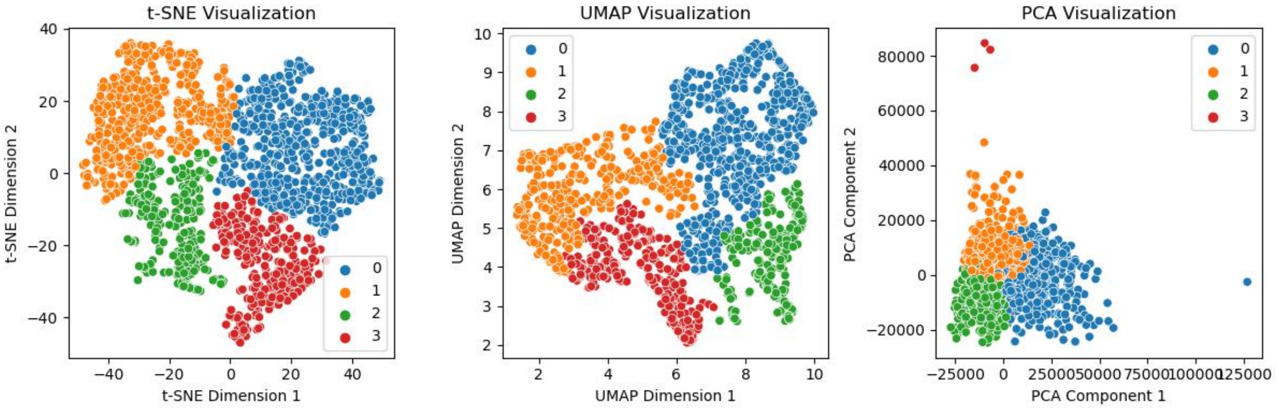


Figure 2.. 2D visualizations of clustering results using Ward algorithm and different dimensionality techniques (4 clusters)

COMPARISON WITH AUTHORS' RESULTS (4 clusters)		
t-SNE	**UMAP**	**PCA**
Number of common elements in ***cluster 0***: 286 out of 655	Number of common elements in **cluster 0**: 332 out of 655	Number of common elements in ***cluster 0***: 144 out of 655
Number of common elements in ***cluster 1***: 142 out of 270	Number of common elements in ***cluster 1***: 130 out of 270	Number of common elements in ***cluster 1***: 46 out of 270
Number of common elements in ***cluster 2***: 24 out of 269	Number of common elements in ***cluster 2***: 50 out of 269	Number of common elements in ***cluster 2***: 88 out of 269
Number of common elements in ***cluster 3***: 4 out of 233	Number of common elements in ***cluster 3***: 139 out of 233	Number of common elements in ***cluster 3***: 0 out of 233

Table 2. Comparison of our study's results with the authors' results, based on common observation inside the formed clusters (4 clusters)

4.1 11 clusters results

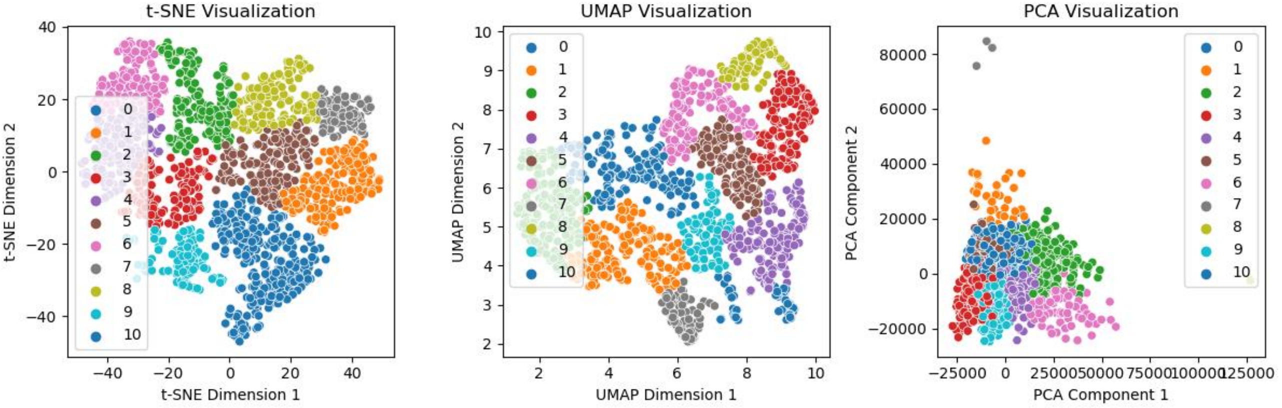


Figure 3. 2D visualizations of clustering results using Ward algorithm and different dimensionality techniques (11 clusters)

COMPARISON WITH AUTHORS’ RESULTS (11 clusters)		
t-SNE	**UMAP**	**PCA**
Number of common elements in ***cluster 0***: 30 out of 281	Number of common elements in ***cluster 0***: 32 out of 281	Number of common elements in ***cluster 0***: 42 out of 280
Number of common elements in ***cluster 1***: 51 out of 243	Number of common elements in ***cluster 1***: 35 out of 243	Number of common elements in ***cluster 1***: 1 out of 243
Number of common elements in ***cluster 2***: 17 out of 168	Number of common elements in ***cluster 2***: 33 out of 168	Number of common elements in ***cluster 2***: 35 out of 168
Number of common elements in ***cluster 3***: 1 out of 145	Number of common elements in ***cluster 3***: 20 out of 145	Number of common elements in ***cluster 3***: 24 out of 145
Number of common elements in ***cluster 4***: 3 out of 117	Number of common elements in ***cluster 4***: 1 out of 117	Number of common elements in ***cluster 4***: 20 out of 117
Number of common elements in ***cluster 5***: 11 out of 110	Number of common elements in ***cluster 5***: 15 out of 110	Number of common elements in ***cluster 5***: 3 out of 110
Number of common elements in ***cluster 6***: 29 out of 108	Number of common elements in ***cluster 6***: 9 out of 108	Number of common elements in ***cluster 6***: 41 out of 108
Number of common elements in *cluster 7*: 3 out of 96	Number of common elements in *cluster 7*: 0 out of 96	Number of common elements in ***cluster 7***: 0 out of 96
Number of common elements in ***cluster 8***: 0 out of 84	Number of common elements in ***cluster 8***: 0 out of 84	Number of common elements in ***cluster 9***: 0 out of 84
Number of common elements in ***cluster 9***: 7 out of 38	Number of common elements in ***cluster 9***: 6 out of 38	Number of common elements in ***cluster 9***: 0 out of 38
Number of common elements in ***cluster 10***: 2 out of 39	Number of common elements in ***cluster 10***: 27 out of 38	Number of common elements in ***cluster 10***: 0 out of 38

Table 3. Comparison of our study's results with the authors' results, based on common observation inside the formed clusters (11 clusters)

The clustering analysis did not achieve a perfect success rate, particularly in the case of the 11-cluster category. Based on the obtained results, it can be inferred that the most likely scenario involves the presence of 2 to 4 clusters, rather than a larger number.

5. Classification Approach

In this section, we will explore the analysis of our classification methodology for scrns data, with the objective of effectively identifying and characterizing unique cell populations. Through the implementation of a thorough pipeline and the utilization of cutting-edge methodologies, our objective was to enhance the efficacy of classification and acquire a deeper understanding of the fundamental biological processes that contribute to cellular heterogeneity. The scope of our analysis includes the examination of feature selection techniques, optimization methods, and the assessment of various classifiers. Additionally, we emphasize the metrics employed for evaluating classification performance and draw comparisons between our findings on feature importance and the results obtained from marker genes. The purpose of this analysis, as mentioned earlier, is to showcase the effectiveness and potential of our classification method in comprehending the intricacies of single-cell gene expression profiles.

5.1 Dataset Description – Preprocessing

This section presents a comprehensive overview of the data exploration and preprocessing procedures implemented in our study to facilitate the classification of single-cell RNA sequencing (scRNA-seq) data. The data preprocessing workflow encompassed the collection of clustering results acquired from distinct TSV files. These results were then integrated with gene expression profiles obtained from scRNA-seq analysis, which were stored in an MTX file format. The ultimate objective was to generate a comprehensive dataset suitable for subsequent classification tasks. The following steps were followed in order to do so:

1. *Consolidation of Clustering Results*: The clustering results obtained from the initial research were dispersed among numerous TSV files. Each file corresponds to a different number of clusters. The aforementioned files encompassed cellular designations and their corresponding assigned groupings. In order to achieve data uniformity, we performed individual parsing of the TSV files for each experiment and subsequently merged them into a cohesive entity using Python. The process of consolidation facilitated the acquisition of essential cluster data, which was subsequently utilized for analysis purposes.

2. *Integration of Gene Expression Profiles*: The authors included the gene expression profiles of the cells obtained from single-cell RNA sequencing (scRNA-seq) analysis in an MTX file format, alongside the clustering results. The provided file contained a matrix that represented the gene expression values. The matrix had dimensions of 1,427 rows, which corresponded to the cells, and 19,595 columns, which represented the genes. Normalization of gene expression values was performed on a per-cell basis to enable subsequent analysis.

3. *Development of an Extensive Dataset*: In order to establish a comprehensive dataset suitable for classification purposes, we combined the consolidated clustering outcomes with the normalized gene expression profiles. The dataset obtained comprised of cells arranged as rows, with their corresponding gene expression profiles as columns. Additionally, there were nine extra columns representing the clusters assigned to each cell based on the various conducted experiments. The amalgamated entity functioned as the reference standard for our classification phase, allowing us to effectively train and assess our classifiers.

The integration of clustering results with gene expression profiles in our dataset yielded a comprehensive resource for classification. This integration allowed us to incorporate both cellular identity information and gene expression signatures, enhancing the richness of our dataset. This dataset offered a comprehensive basis for our subsequent classification pipeline.

5.2 Pipeline Structure

The classification pipeline was developed with the intention of taking advantage of the dataset that was collected and preprocessed, as outlined in the preceding sections. The pipeline encompassed multiple stages, including dataset filtration, division into training and testing sets, optimization of features, tuning of hyperparameters, and ultimate evaluation. This section presents a comprehensive overview of the pipeline structure, delineating each individual step and elucidating its significance within the classification process.

In accordance with the methodology employed in the original study, we applied a filtering process to the dataset in order to select and retain the 1500 genes that exhibited the highest degree of variation. The purpose of this filtering step was to prioritize genes that demonstrated substantial variation in expression, employing a methodology akin to the authors' clustering analysis. As a result, the dataset was expanded to include 1427 rows that represent cells and 1500 columns that represent gene expression profiles. Additionally, there were extra columns that contained the ground truth labels obtained from the authors' clustering results.

In order to ensure an impartial evaluation and validation of the classifiers, a train-test split was conducted on the filtered dataset. The dataset was partitioned into a training set and a concealed test set. The training set underwent multiple splits to facilitate subsequent optimization processes, while the test set was kept concealed to ensure its integrity for the final evaluation of the classifiers. The implementation of this separation ensured a fair and unbiased evaluation of the classifiers' performance on data that had not been previously encountered.

The implemented pipeline utilized feature optimization techniques in order to determine the most optimal subset of features for the purpose of classification. During this stage, several feature selection methods, including ANOVA and mRMR, were utilized to identify the gene expression profiles that are most informative and discriminative. The objective of the optimization process was to enhance the efficiency and performance of the classifiers by prioritizing the most pertinent features.

Hyperparameter tuning was performed in order to optimize the performance of the classifiers. The process of hyperparameter optimization, which involves adjusting parameters such as the regularization parameter in logistic regression or the number of estimators in random forests, was conducted using a stratified 5-fold cross-validation technique. The utilization of this approach successfully mitigated the issue of class imbalance present in our dataset, thereby enabling the establishment of a more resilient and precise configuration for the classifier.

After conducting the optimization procedures, the classifiers were assessed by employing the concealed test set. The evaluation metrics employed encompassed balanced accuracy, precision, recall, and F1-score. The performance of the classifiers was more comprehensively evaluated by utilizing the balanced accuracy metric, taking into account the class imbalance present in our dataset. The classifiers' performance in accurately classifying each class was assessed using precision, recall, and F1-score, which took into account both false positives and false negatives.

The diagram presented below depicts the configuration of our classification pipeline, which encompasses several stages including dataset filtering, train-test split, feature optimization, hyperparameter tuning, and final evaluation.

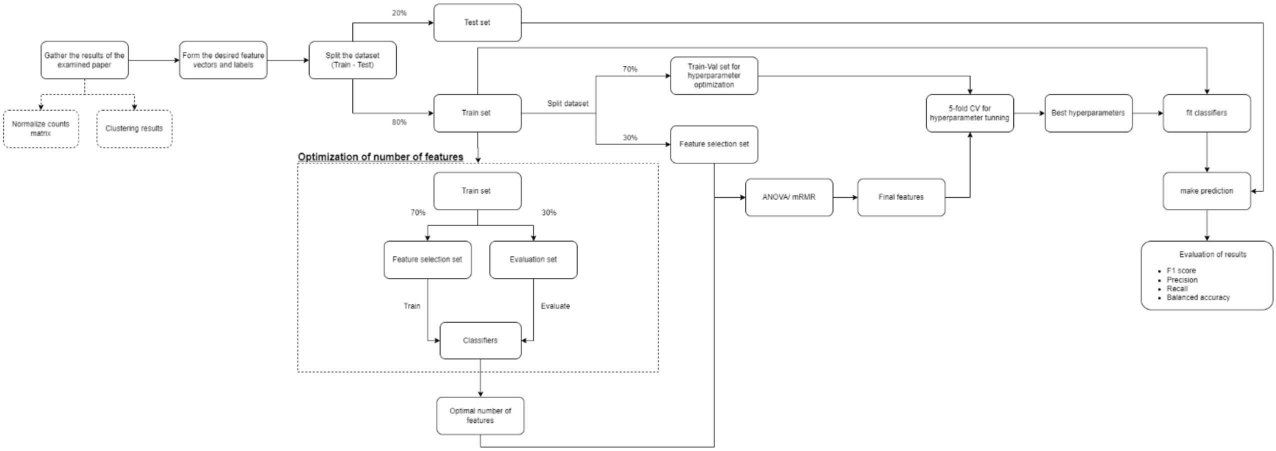


Figure 4. Classification pipeline scheme

The objective of implementing this pipeline framework was to enhance the efficiency of the classification procedure and achieve precise forecasts of cell populations by leveraging their gene expression profiles. The following sections will provide an in-depth analysis of the specific details and outcomes of each stage, emphasizing the performance and insights acquired from our classification pipeline.

5.3 Pipeline Steps-Description

5.3.1 Optimization of number of features to keep

Following the filtration of the dataset to preserve the 1500 genes exhibiting the highest variability, the subsequent stage in our pipeline entailed ascertaining the ideal quantity of features to retain. In order to mitigate the potential challenge posed by a feature space with a high number of dimensions, our objective was to identify a subset of features that would preserve classification accuracy while simultaneously decreasing computational complexity. The inclusion of this step played a pivotal role in improving the effectiveness and comprehensibility of the classifiers.

In order to achieve this objective, the training dataset was divided into two separate sets. The initial set was employed for the purpose of feature selection, wherein various iterations were conducted to explore different quantities of features. The iterative procedure encompassed the training of five classifiers, namely Random Forest, XGBoost, Logistic Regression, Support Vector Classifier (SVC), and Gaussian Naive Bayes, on the dataset containing the chosen features. The second set, referred to as the evaluation set, was utilized to generate predictions utilizing the trained classifiers, and the Matthews Correlation Coefficient (MCC) score was employed as the metric for evaluation. Through the iterative repetition of this procedure, with the manipulation of the number of features in each iteration, a comprehensive set of Matthews Correlation Coefficient (MCC) scores was acquired for each classifier. In order to ascertain the most suitable number of features, we computed the mean Matthews Correlation Coefficient (MCC) score across all classifiers. The MCC score, calculated as an average, was utilized as the optimization metric.

In order to enhance the optimization process, we utilized Optuna, a Python library designed for hyperparameter optimization. The utilization of Optuna facilitated a methodical examination of the feature space, enabling the search for the optimal number of features that would yield the highest average Matthews Correlation Coefficient (MCC) score. Through a process of iterative evaluation involving varying numbers of features, we successfully determined the optimal subset that exhibited the highest performance across all classifiers.

The implementation of this optimization procedure enabled us to achieve a harmonious equilibrium between the reduction of dimensionality and the accuracy of classification. Our objective was to reduce the likelihood of overfitting and maintain the essential informative attributes required for precise classification by choosing a subset of features that consistently demonstrated strong performance across the classifiers. In our pipeline, the optimization of feature selection played a pivotal role in determining the optimal number of features to retain for subsequent classification tasks. The feature subset that was optimized resulted in improved computational efficiency, interpretability, and robust classification performance.

5.3.2 Main split of train dataset

In an effort to enhance the reliability of model training and mitigate the risk of data leakage, we proceeded to partition the initial training dataset (derived from the the first split of the dataset) into two separate subsets. One subset was allocated for the purpose of feature selection, while the other subset was designated for hyperparameter tuning. The purpose of this separation was to mitigate the risk of inadvertent information leakage, which could potentially result in overfitting and biased performance assessment.

The smaller subset of the train dataset was specifically allocated for the purpose of feature selection. This subset was employed to investigate different combinations of features and identify the most optimal subset of features for the purpose of classification. By performing feature selection on a reduced dataset, we were able to reduce computational complexity while maintaining the integrity of the results.

The remaining portion of the train dataset was allocated for the purpose of hyperparameter tuning. Hyperparameters refer to parameters that are not subject to optimization during the training phase and exert a substantial influence on the performance of the model. The separation of the hyperparameter tuning and feature selection phases was implemented to ensure that the optimization of hyperparameters was conducted using a dataset that was entirely distinct from the feature selection process, in an attempt to reduce the risk of inadvertent information leakage as we previously mentioned.

5.3.3 Feature selection techniques

In our classification pipeline, we incorporated two feature selection methodologies: Analysis of Variance (ANOVA) and minimum Redundancy Maximum Relevance (mRMR). These methodologies enabled us to discern the most informative and distinguishing characteristics from the dataset, which enhanced the classification efficacy and comprehensibility of our models.

The Analysis of Variance (ANOVA) is a statistical technique utilized to evaluate the statistical significance of variations among means of different groups. In the domain of feature selection, ANOVA assesses the extent of variability exhibited by each feature across distinct classes or clusters. Features that exhibit a high degree of variability between different classes and a low degree of variability within each class are regarded as more discriminative and are consequently chosen for subsequent analysis. The objective of our study was to utilize analysis of variance (ANOVA) on our dataset in order to identify more genes that displayed notable variations in expression patterns across distinct cell populations.

The concept of minimum redundancy refers to the principle of minimizing unnecessary repetition or duplication in a given context. The Maximum Relevance (mRMR) algorithm is a feature selection technique designed to identify a subset of features that optimizes the relevance to the target variable, while simultaneously minimizing redundancy among the selected features. The algorithm computes the significance of each feature with respect to the target variable and evaluates the overlap or duplication among the features. The mRMR algorithm aims to identify a subset of features that effectively captures relevant and diverse information for the purpose of classification, while minimizing redundancy. The utilization of this technique facilitated the identification of features that exhibited strong correlations with the target labels, while concurrently mitigating the likelihood of selected features exhibiting substantial overlap.

In our feature selection process, we aimed to encompass the discriminative capability and the variety of gene expression profiles linked to distinct cell populations by experimenting with ANOVA and mRMR methodologies. The utilization of these techniques facilitated our ability to concentrate on the most informative features while minimizing the impact of redundant or irrelevant features.

5.3.4 Hyperparameter tunning

The objective of the hyperparameter tuning stage in our pipeline was to optimize the parameters of each classifier in order to maximize their performance on the chosen features. In order to achieve this objective, we utilized Optuna once again.

In order to obtain accurate performance estimates and address the issue of class imbalance in our dataset, we employed stratified 5-fold cross-validation during the hyperparameter tuning phase. The utilization of this technique ensured that the distribution of class labels remained consistent across each fold, thereby facilitating a reliable assessment of the classifiers' efficacy.

Through the utilization of Optuna's optimization capabilities, we conducted a systematic exploration of various hyperparameter configurations for each classifier. The Optuna framework employed intelligent sampling and evaluation techniques to explore a range of hyperparameter combinations. Its objective was to identify the optimal set of hyperparameters that produced the highest performance, as determined by the selected evaluation metric, MCC score. The aim of the study was to determine the hyperparameters that would yield the highest performance metric, which would allow the achievement of optimal classification outcomes and robustness of the classifiers.

6. Results – Discussion

The authors of the aforementioned paper discussed the concept of three clusters. However, in their results files, they provide labels for 2, 4, 7, 8, 11, 18, 23, 31 and 37 clusters. Taken this into account, this study examined the problem from three different perspectives: a 2-class problem, a 4-class problem, and an 11-class problem, which the authors asserted to be the most effective approaches.

It is imperative to acknowledge that our methodology prioritized data-driven analysis, utilizing the top 1500 genes with high variability, and tried to reduce this amount, rather than exclusively depending on marker genes. The aim of our study was to streamline the issue at hand and enhance comprehension, given the authors' lack of a comprehensive biological rationale for their clustering outcomes.

The results will be presented using the optimized hyperparameters acquired through the hyperparameter tuning process of our pipeline, along with the 300 features suggested by our feature selection methods. In addition, we will illustrate the significant attributes of our classifiers in comparison to the marker genes provided by the authors for each clustering scenario in their results files.

6.1 Two class problem and comparison with marker genes

In this section, we will proceed to present the outcomes obtained for the binary classification task. The subsequent tables and figures exhibit the outcomes for our binary classification problem, encompassing the evaluation metrics and the significant features identified by each classifier.

ANOVA technique used for feature selection

Classifier	Balanced Accuracy	Precision	Recall	F1-score
Random Forest	0.972	1.0	0.94	0.971
XGBoost	0.916	1.0	0.83	0.909
Logistic Regression	0.968	0.894	0.94	0.918
Support Vector Machine	0.966	0.85	0.94	0.8947
Gaussian Naive Bayes	0.831	0.923	0.66	0.774

Table 4. Classification results, 2 class problem, ANOVA technique used for feature selection

mRMR technique used for feature selection

Classifier	Balanced Accuracy	Precision	Recall	F1-score
Random Forest	1.0	1.0	1.0	1.0
XGBoost	0.972	1.0	0.944	0.971
Logistic Regression	0.94	0.888	0.888	0.888
Support Vector Machine	0.94	0.88	0.888	0.888
Gaussian Naive Bayes	0.972	0.842	0.888	0.864

Table 5. Classification results, 2 class problem, mRMR technique used for feature selection

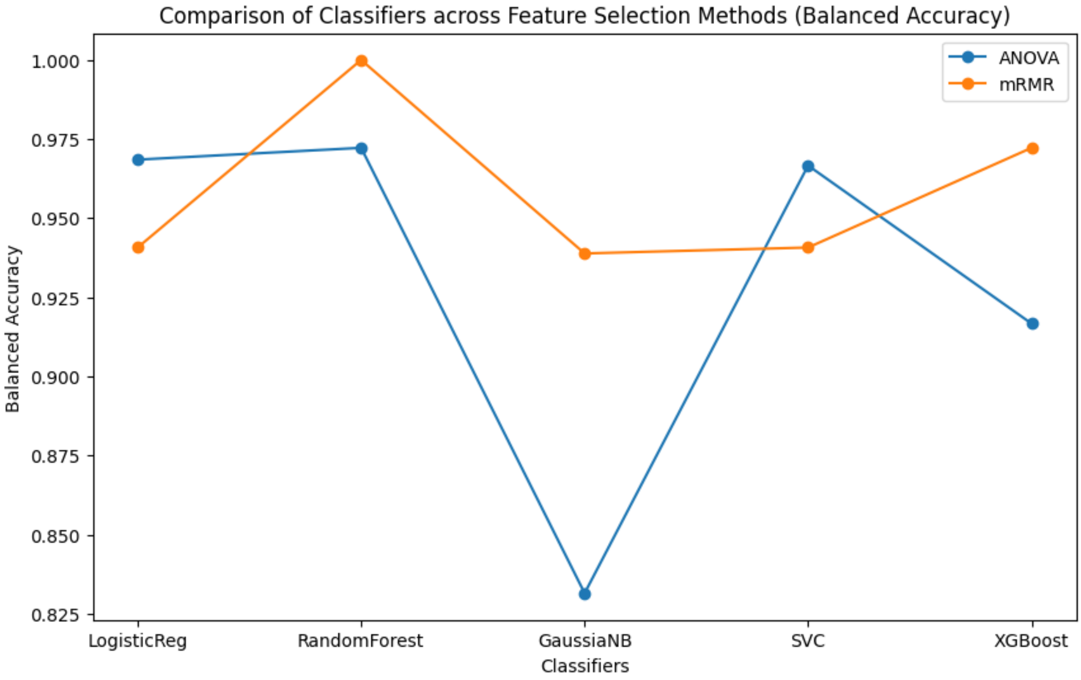


Figure 5. Balanced accuracy across different classifiers with different feature selection techniques (2class problem)

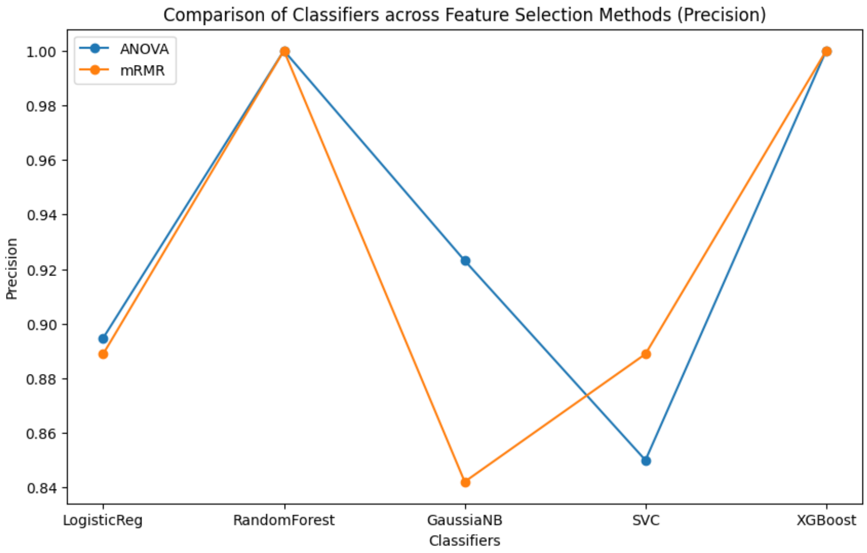


Figure 6. Precision across different classifiers with different feature selection techniques (2class problem)

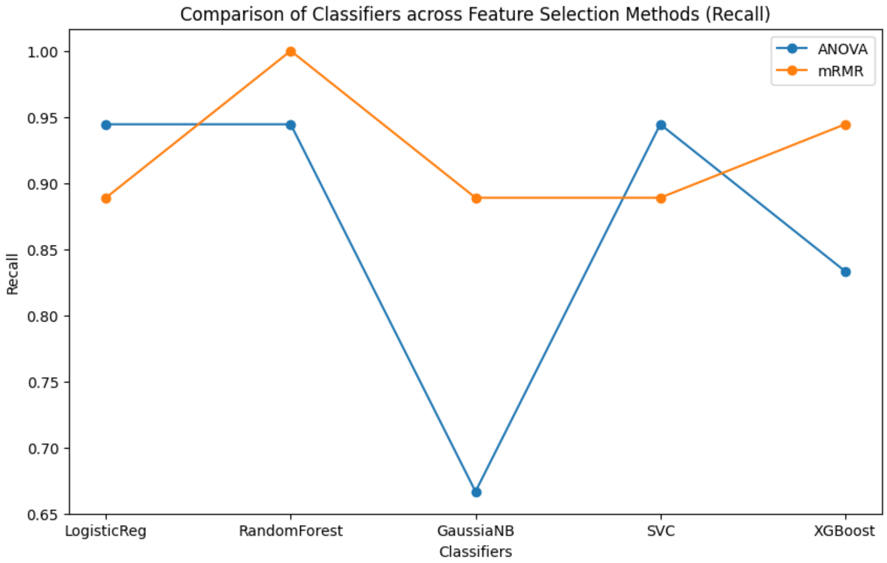


Figure 7. Recall across different classifiers with different feature selection techniques (2class problem)

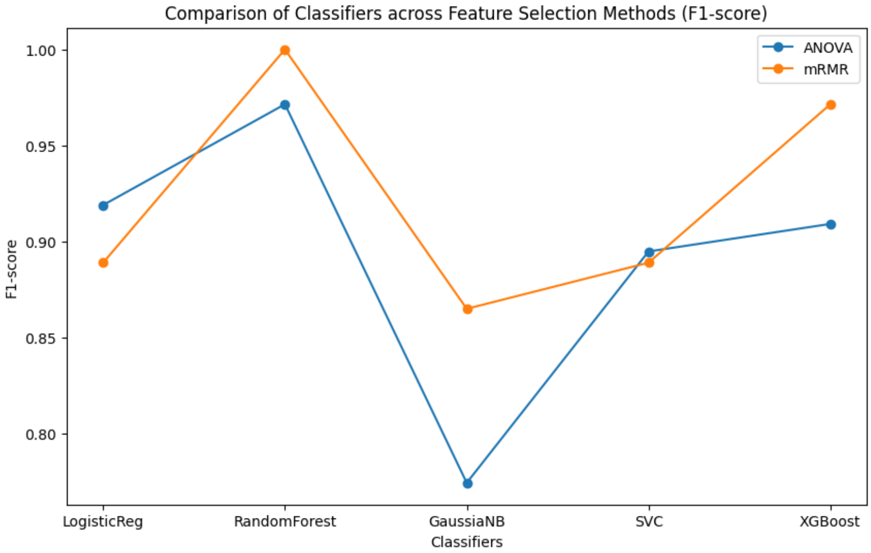


Figure 8.F1-score across different classifiers with different feature selection techniques (2class problem)

For the purpose of ascertaining the key attributes in our classifiers, we utilized various classifier's attributes . In the context of tree-based algorithms such as Random Forest and XGBoost, we employed the "feature_importances_" attribute to assess the significance of individual features in the classification procedure. We utilized the "coef_ "attribute to extract the feature weights for the remaining classifiers, namely Logistic Regression and SVC. GaussianNB was excluded from this analysis due to its comparatively lower performance. In order to maintain relevance, a threshold of 0.01 was established for the tree-based classifiers, resulting in the extraction of the most significant features. Conversely, for the remaining classifiers, the 20 most important features were obtained.

Feature importance threshold: 0.01 (for tree classifiers) **20 most important features based on their weights** (rest of classifiers)		
Marker genes given by the authors for two clusters experiment: 118		
Feature selection**: ANOVA**	**RF:**	returned 22 genes, **all of them were marker genes from *class 1***
	XGBoost:	returned 30 genes **17/30: marker genes from *class 1*** 1/30 : marker gene from *class 2* 12/30 : no marker genes
	LR:	7/20: marker genes from *class 1* 4/20: marker genes from *class 2* **9/20: no marker genes**
	SVC:	7/20: marker genes from *class 1* 4/20: marker genes from *class 2* **9/20: no marker genes**
Feature selection**: mRMR**	**RF:**	returned 20 genes **17/20: marker genes from *class 1*** 2/20: marker genes from *class 2* 1/20: no marker genes
	XGBoost:	returned 17 genes **11/17: marker genes from *class 1*** 2/17 : marker gene for *class 2* 4/17: no marker genes
	LR:	2/20: marker genes from *class 1* 7/20: marker genes from *class 2* **11/20: no marker genes**
	SVC:	1/20: marker genes from *class 1* 8/20: marker genes from *class 2* **11/20: no marker genes**

Table 6. Comparison of important features of classifiers with marker genes identified by authors for 2class problem

6.2 Four class problem and comparison with marker genes

In this section, we will present the outcomes obtained for the fourth-class classification approach. The following tables and figures present the results of classification problem, including the evaluation metrics and the significant features identified by each classifier.

ANOVA technique used for feature selection

Classifier	Balanced Accuracy	Precision	Recall	F1-score
Random Forest	0.902	0.914	0.913	0.912
XGBoost	0.906	0.916	0.916	0.915
Logistic Regression	0.880	0.891	0.891	0.891
Support Vector Machine	0.843	0.861	0.860	0.859
Gaussian Naive Bayes	0.895	0.888	0.881	0.882

Table 7. Classification results, 4 class problem, ANOVA technique used for feature selection

mRMR technique used for feature selection

Classifier	Balanced Accuracy	Precision	Recall	F1-score
Random Forest	0.914	0.924	0.923	0.922
XGBoost	0.905	0.917	0.916	0.915
Logistic Regression	0.876	0.888	0.888	0.888
Support Vector Machine	0.802	0.833	0.846	0.844
Gaussian Naive Bayes	0.851	0.867	0.864	0.862

Table 8. Classification results, 4 class problem, mRMR technique used for feature selection

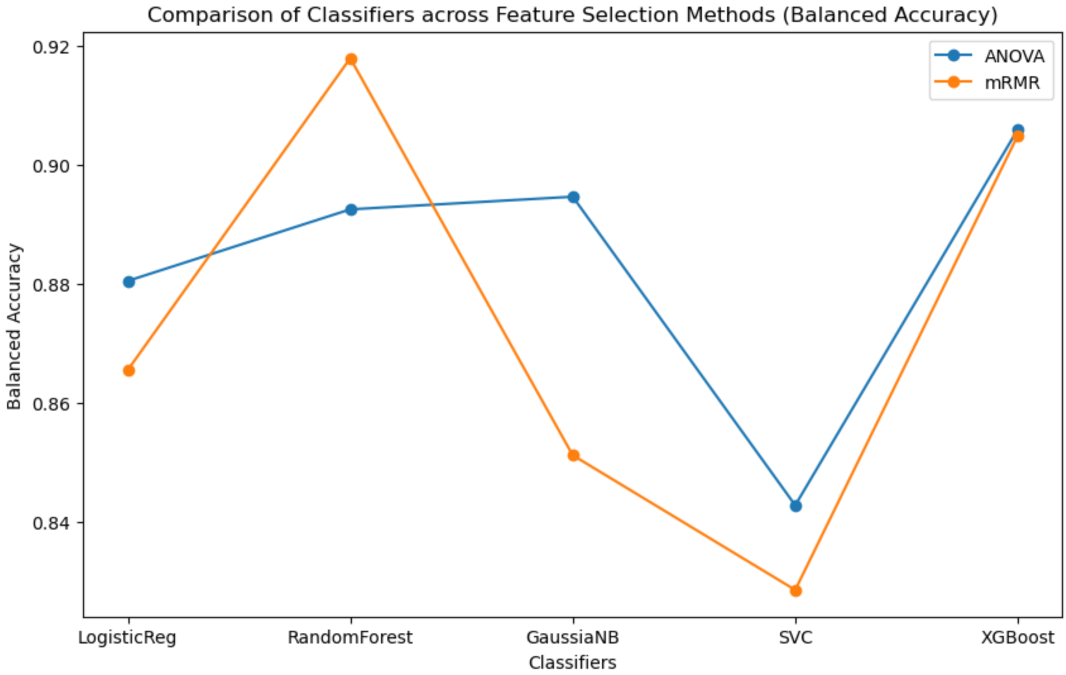


Figure 9. Balanced accuracy across different classifiers with different feature selection techniques (4class problem)

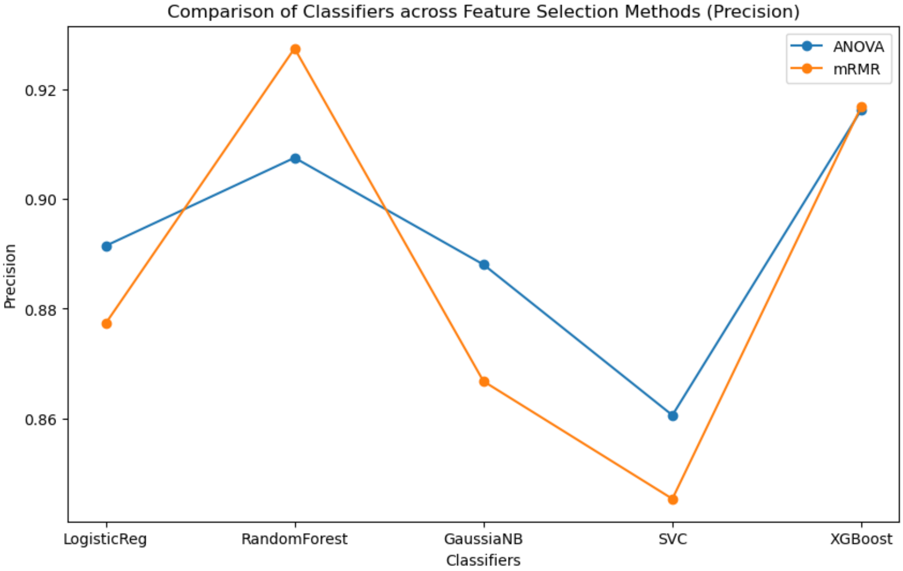


Figure 10. Precision across different classifiers with different feature selection techniques (4class problem)

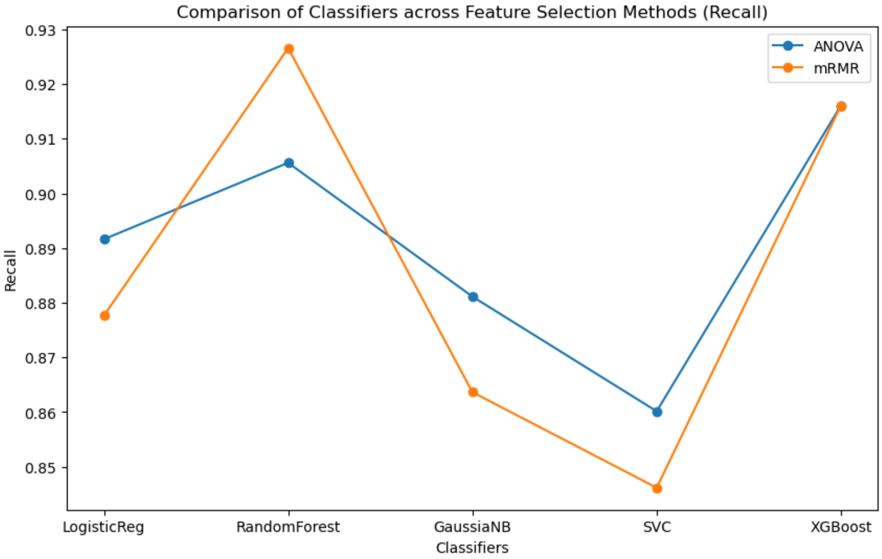


Figure 11. Recall across different classifiers with different feature selection techniques (4class problem)

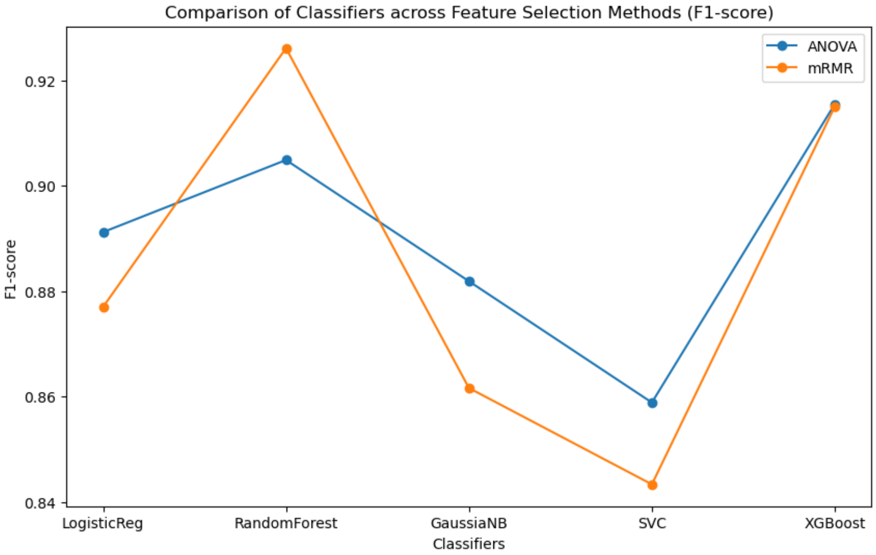


Figure 12. F1-score across different classifiers with different feature selection techniques (4class problem)

The relevant features for our classifiers were identified once more, employing the identical methodology utilized in the classification of the two-class problem. Additionally, we conducted further analysis on our Logistic Regression classifier and identified the significant genes associated with each class. The findings are displayed in the table provided.

Feature importance threshold: 0.01 (for tree classifiers)

20 most important features based on their weights (rest of classifiers)

Marker genes given by the authors for two clusters experiment: 302

Feature selection**: ANOVA:** **RF:** returned 25 genes
14/25: marker genes from *class 1*
9/25 : marker gene from *class 3*
2/25 : marker genes from *class 4*

XGBoost: returned 12 genes
9/12: marker genes from *class 1*
2/12 : marker gene from *class 3*
1/12 : marker genes from *class 4*

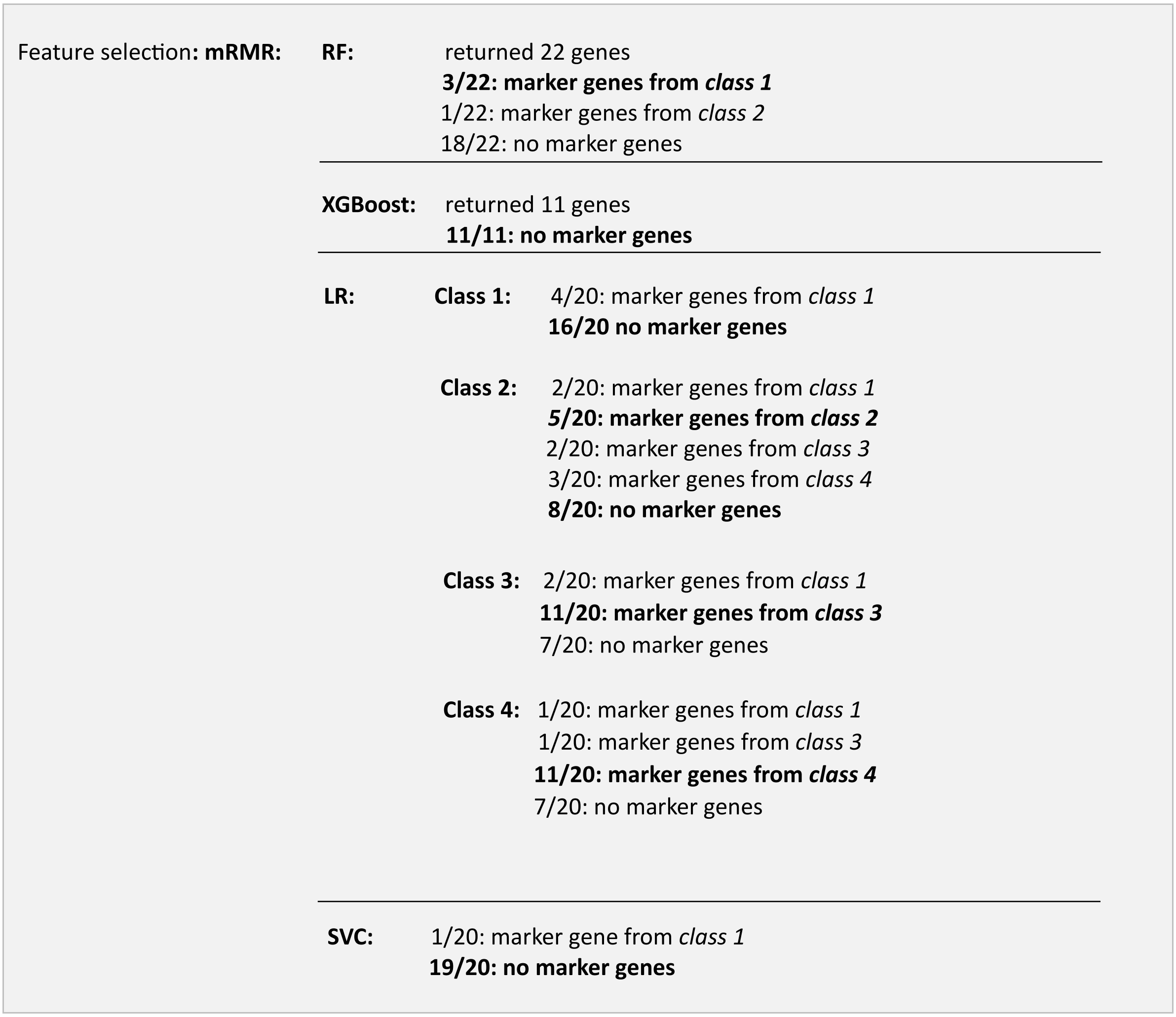
LR: **Class 1:** **3/20: marker genes from *class 1***
17/20 no marker genes

Class 2: 2/20: marker genes from *class 1*
6/20: marker genes from *class 2*
2/20: marker genes from *class 3*
2/20: marker genes from *class 4*
8/20: no marker genes

Class 3: 2/20: marker genes from *class 1*
11/20: marker genes from *class 3*
7/20: no marker genes

Class 4: 2/20: marker genes from *class 3*
11/20: marker genes from *class 4*
7/20: no marker genes

SVC: 1/20: marker gene from *class 1*
1/20: marker gene from *class 4*
18/20: no marker genes



Feature selection**: mRMR:**	**RF:**		returned 22 genes **3/22: marker genes from *class 1*** 1/22: marker genes from *class 2* 18/22: no marker genes
	XGBoost:		returned 11 genes **11/11: no marker genes**
	LR:	**Class 1:**	4/20: marker genes from *class 1* **16/20 no marker genes**
		Class 2:	2/20: marker genes from *class 1* **5/20: marker genes from *class 2*** 2/20: marker genes from *class 3* 3/20: marker genes from *class 4* **8/20: no marker genes**
		Class 3:	2/20: marker genes from *class 1* **11/20: marker genes from *class 3*** 7/20: no marker genes
		Class 4:	1/20: marker genes from *class 1* 1/20: marker genes from *class 3* **11/20: marker genes from *class 4*** 7/20: no marker genes
	SVC:		1/20: marker gene from *class 1* **19/20: no marker genes**

Table 9. Comparison of important features of classifiers with marker genes identified by authors for 4class problem

6.3 Eleven class problem and comparison with marker genes

Construction of an eleven-class classifier was our ultimate strategy. The results, including the assessment metrics and the relevant features determined by each classifier, are shown in the tables and figures that follow.

ANOVA technique used for feature selection

Classifier	Balanced Accuracy	Precision	Recall	F1-score
Random Forest	0.703	0.795	0.790	0.782
XGBoost	0.741	0.794	0.811	0.809
Logistic Regression	0.759	0.773	0.766	0.764
Support Vector Machine	0.742	0.771	0.759	0.759
Gaussian Naive Bayes	0.692	0.755	0.706	0.704

Table 10. Classification results, 11 class problem, ANOVA technique used for feature selection

mRMR technique used for feature selection

Classifier	Balanced Accuracy	Precision	Recall	F1-score
Random Forest	0.727	0.809	0.804	0.794
XGBoost	0.732	0.794	0.790	0.782
Logistic Regression	0.800	0.808	0.801	0.799
Support Vector Machine	0.769	0.771	0.762	0.758
Gaussian Naive Bayes	0.730	0.755	0.731	0.732

Table 11. Classification results, 11 class problem, mRMR technique used for feature selection

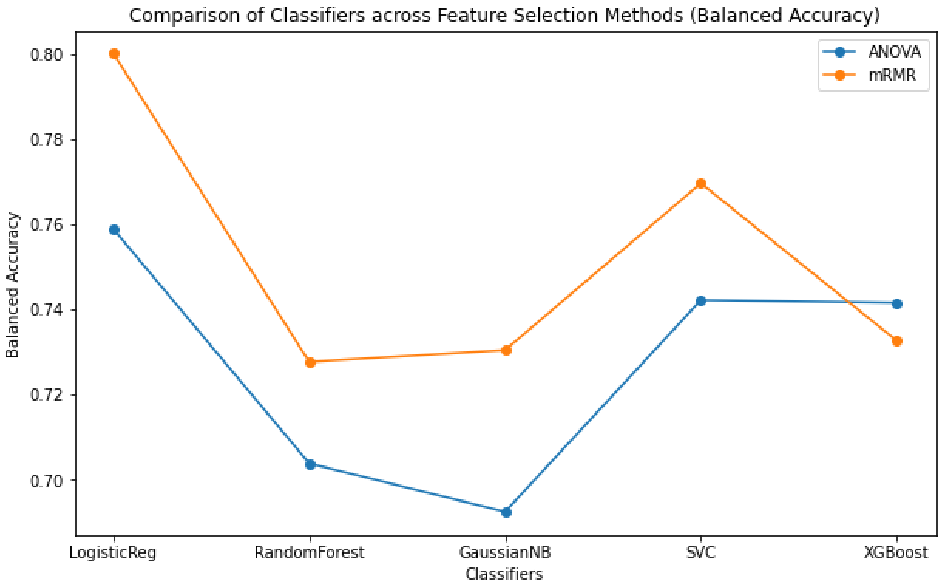


Figure 13. Balanced accuracy across different classifiers with different feature selection techniques (11class problem)

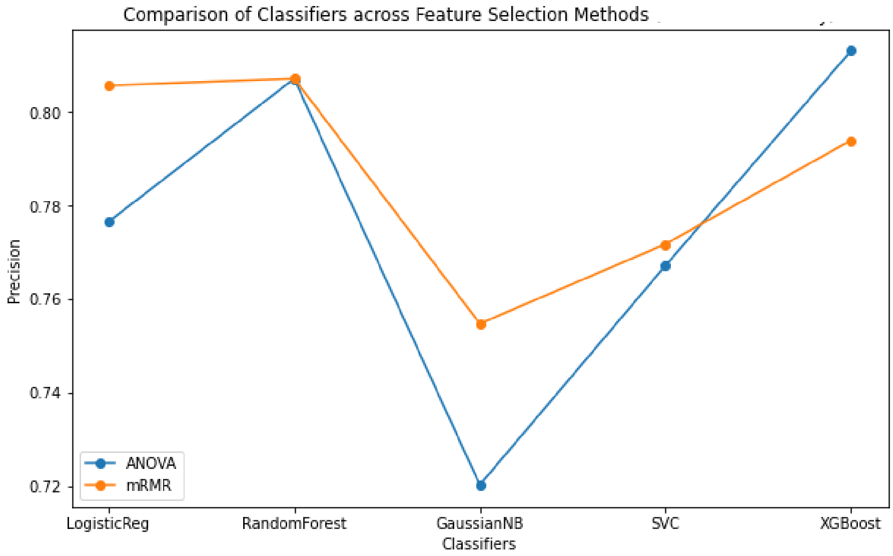


Figure 14. Precision across different classifiers with different feature selection techniques (11class problem)

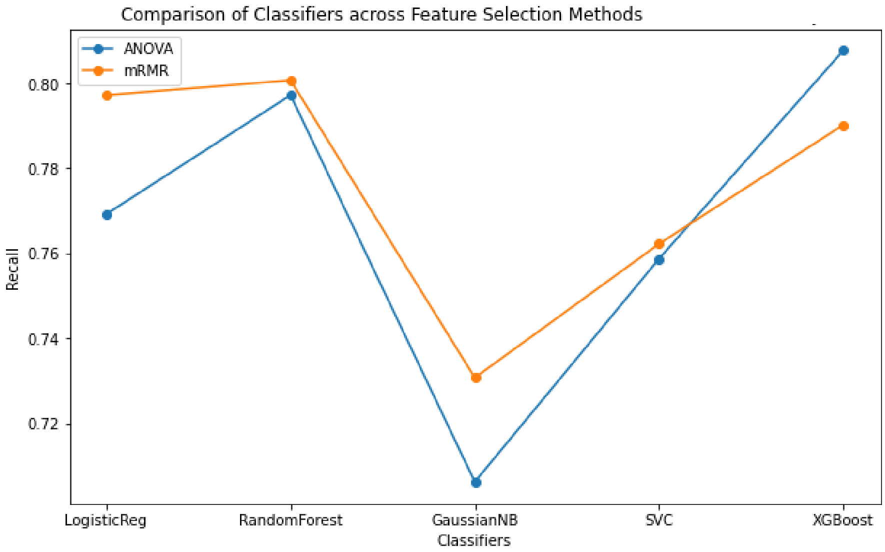


Figure 15. Recall across different classifiers with different feature selection techniques (11class problem)

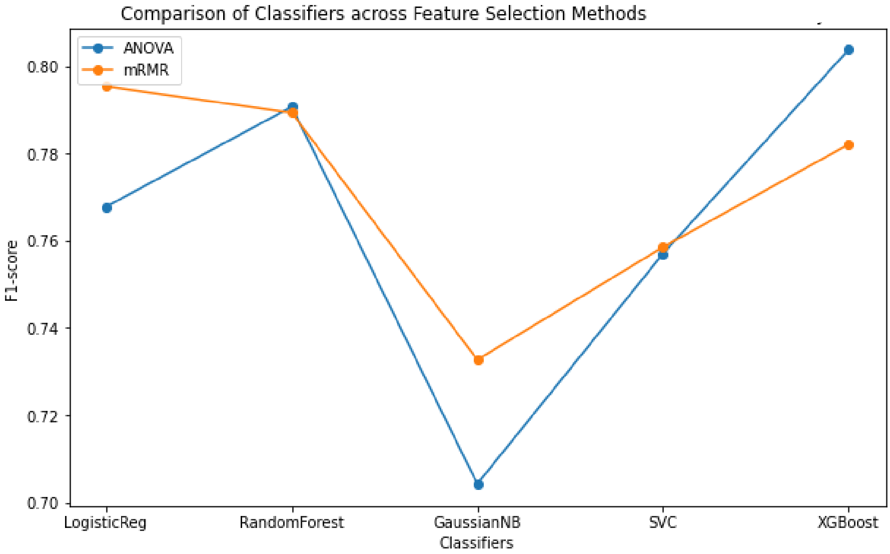
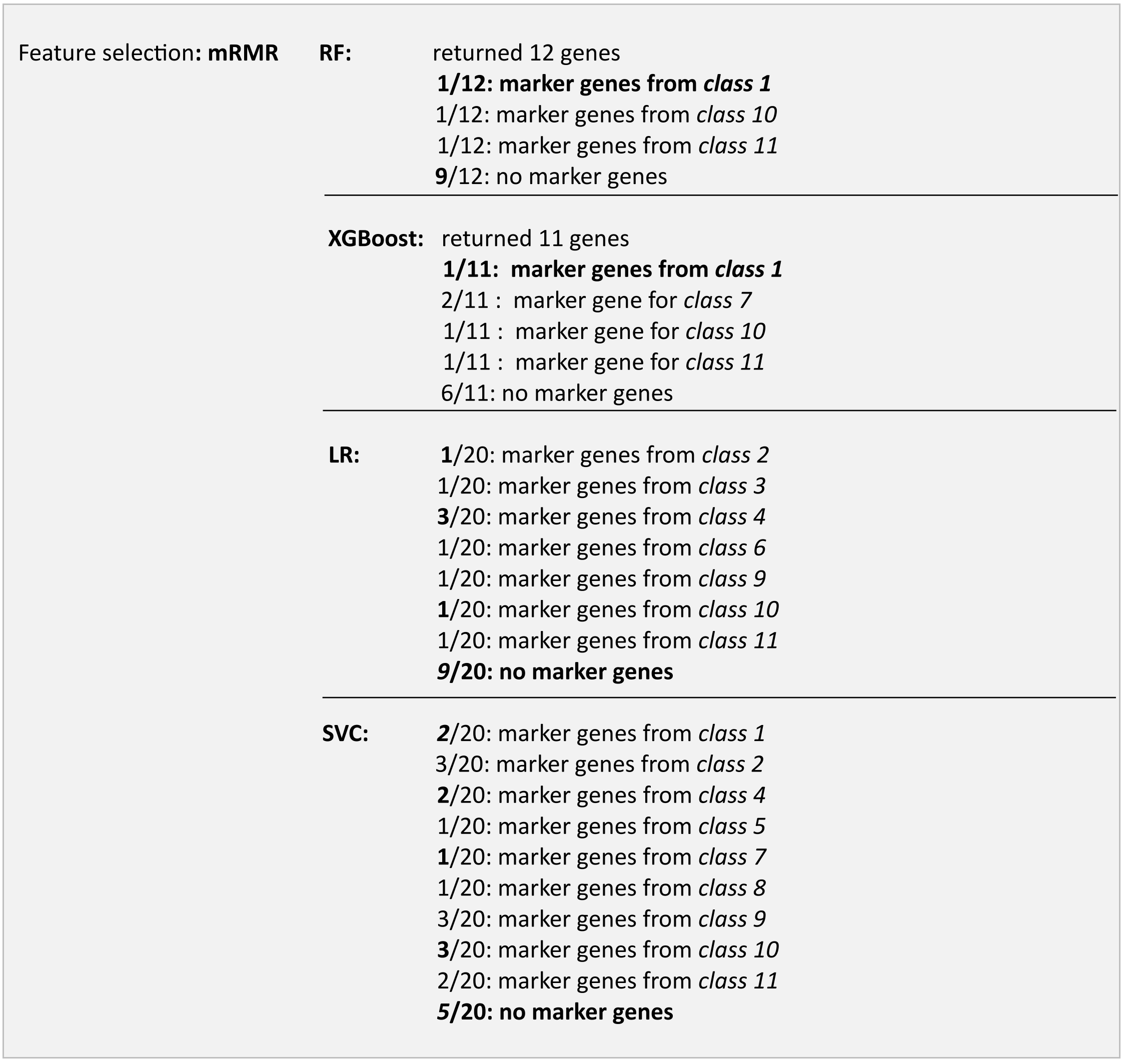


Figure 16. F1-score across different classifiers with different feature selection techniques (11class problem)

Feature importance threshold: 0.01 (for tree classifiers) **20 most important features based on their weights** (rest of classifiers)		
Marker genes given by the authors for two clusters experiment: 118		
Feature selection**: ANOVA**	**RF:**	returned 13 genes 2/13: marker genes from *class 4* **1**/13: marker genes from *class 5* 1/13: marker genes from *class 6* 2/13: marker genes from *class 7* **1**/13: marker genes from *class 8* 1/13: marker genes from *class 9* ***5*/13: no marker genes**
	XGBoost:	returned 17 genes **3/17: marker genes from *class 4*** 1/17: marker gene from *class 6* 2/17: marker genes from *class 7* **1**/17: marker genes from *class 8* 1/17: marker genes from *class 9* 1/17: marker genes from *class 10* ***8*/17: no marker genes**
	LR:	**2**/20: marker genes from *class 4* 2/20: marker genes from *class 6* 1/20: marker genes from *class 9* 2/20: marker genes from *class 10* 2/20: marker genes from *class 11* ***11*/20: no marker genes**
	SVC:	***3***/20: marker genes from *class 1* 2/20: marker genes from *class 4* 1/20: marker genes from *class 7* 2/20: marker genes from *class 8* 1/20: marker genes from *class 9* 2/20: marker genes from *class 10* 1/20: marker genes from *class 11* **8/20: no marker genes**



Feature selection**: mRMR**	**RF:**	returned 12 genes **1/12: marker genes from *class 1*** 1/12: marker genes from *class 10* 1/12: marker genes from *class 11* **9**/12: no marker genes
	XGBoost:	returned 11 genes **1/11: marker genes from *class 1*** 2/11 : marker gene for *class 7* 1/11 : marker gene for *class 10* 1/11 : marker gene for *class 11* 6/11: no marker genes
	LR:	**1**/20: marker genes from *class 2* 1/20: marker genes from *class 3* **3**/20: marker genes from *class 4* 1/20: marker genes from *class 6* 1/20: marker genes from *class 9* **1**/20: marker genes from *class 10* 1/20: marker genes from *class 11* ***9*/20: no marker genes**
	SVC:	***2***/20: marker genes from *class 1* 3/20: marker genes from *class 2* **2**/20: marker genes from *class 4* 1/20: marker genes from *class 5* **1**/20: marker genes from *class 7* 1/20: marker genes from *class 8* 3/20: marker genes from *class 9* **3**/20: marker genes from *class 10* 2/20: marker genes from *class 11* ***5*/20: no marker genes**

Table 12. Comparison of important features of classifiers with marker genes identified by authors for 11class problem

6.4 Classifiers' performance comments

There is a negative correlation between the number of classes and the classifier's scores, whereby an increase in the former tends to result in a decrease in the latter. In the context of the two-class problem, it was consistently observed that mRMR feature selection yielded superior performance compared to ANOVA, as indicated by the scores obtained. Among the various classifiers considered, Random Forest exhibited the most favorable performance, as evidenced by achieving the highest scores across all evaluation metrics, namely Balanced Accuracy, Precision, Recall, and F1-score, all of which were equal to 1.

In the context of the four-class problem, it was observed that the disparity between the mRMR and ANOVA scores was relatively diminished. Nonetheless, the Random Forest algorithm, when combined with mRMR feature selection, exhibited superior performance compared to other methods. It achieved a Balanced Accuracy of 0.914, Precision of 0.924, Recall of 0.923, and F1-score of 0.922.

As for the eleven-class problem, it was observed that Random Forest with mRMR feature selection consistently achieved the highest scores. The model attained a Balanced Accuracy score of 0.727, Precision score of 0.809, Recall score of 0.804, and F1-score of 0.794.

In general, the findings of this study demonstrate the superior performance of Random Forest with mRMR feature selection in comparison to ANOVA across various classification tasks. The scores serve as a measure of the classifier's capacity to effectively categorize data across multiple classes, thereby showcasing its potential as a dependable and resilient method for classification endeavors.

6.5 Important Features Analysis - Identification of Non-Marker Genes

Following the completion of the classification process, our attention was redirected towards the identification of significant features for each classifier. In order to achieve this objective, we utilized methodologies that were specific to the attributes under consideration. The feature_importances_ attribute was employed in tree-based algorithms such as Random Forest and XGBoost. In contrast, the coef_ attribute was utilized for classifiers such as Logistic Regression and SVC. This methodology facilitated the identification of the most influential features that govern the classification procedure.

After conducting a comparative analysis between our findings and the marker genes identified by the authors using DESeq, it was observed that certain aspects of our results were consistent with theirs. However, it was observed that specific genes that we considered significant for all three of our classification issues were not encompassed in the list provided by the authors. The observed disparity served as a catalyst for our further investigation into the significance of these genetic factors.

Unfortunately, there were instances where the precise class affiliations of these significant genes could not be ascertained. Therefore, it is not possible to make definitive inferences regarding class attributes solely from these genetic factors. However, a research study was conducted to investigate the molecular function of these entities in order to obtain additional understanding of their importance in our classifiers.

The genes deemed significant in various classification problems, as indicated in the subsequent table, include RPS24, TMSB4X, TMSB10, and GAPDH. We conducted an examination of the Gene Ontology (GO) annotations pertaining to molecular functions associated with them. The analysis conducted yielded a wide array of functions that are associated with these genes.

The ribosomal protein RPS24 has been identified as having a role in the binding of RNA molecules. Through extensive research, it has been discovered that this particular gene is closely linked to Müller glia (MG) cells and reactive Müller glia (MG) cells [2][3]. These specific types of glial cells are highly prevalent in the neural retina and exhibit a reactive response following injury or in the presence of disease. Another noteworthy publication in which this gene is referenced is documented in [4].

TMSB4X and TMSB10, which belong to the thymosin beta family, possess significant annotations related to actin monomer binding and actin binding. These annotations suggest their participation in the regulation of actin cytoskeleton dynamics. Furthermore, the aforementioned genes have been subjected to annotation for protein binding, indicating their potential for interacting with other proteins. It is noteworthy to observe that both TMSB4X and TMSB10 have been identified as significant genes specifically for class 1 in the context of the 4-class classification problem. The aforementioned discovery implies that these genes might have a significant impact on differentiating class 1 from the remaining classes, potentially indicating their involvement in distinct biological processes or molecular pathways linked to that particular class.

The most intriguing findings were observed for the TMSB4X gene. The gene in question was examined in the aforementioned study [5], which produced a transcriptome atlas of the mature human retina. This paper aims to identify certain subpopulations and examine the high expression of a specific gene referred to as the glioma-related gene [6]. Furthermore, in a scholarly article [7], a study was conducted to investigate the genetic regulation in primary open-angle glaucoma. Within this study, a particular gene was identified within a cluster that was characterized by genes associated with cell growth and differentiation. It is worth noting that these genes are not exclusively specific to the retina, but rather have broader implications. This gene is also identified in the same scholarly article as markers linked to the organization of the cytoskeleton.

Finally, GAPDH which is annotated with Gene Ontology (GO) terms indicating its oxidoreductase activity, implies its potential participation in redox reactions or the facilitation of oxidation-reduction processes.

The wide range of molecular functions exhibited by these genes offers valuable insights into their roles within our classifiers. Moreover, these findings serve as a foundation for future investigations into the functional significance of these genes within the specific context of our classification problems.

Gene name	Gene code	Class identified significant. (4-class problem)	Class identified significant. (11-class problem)	Molecular Functions
RPS24	**ENSG00000138326**	**2**	**5**	RNA binding, structural constituent of ribosome, translation initiation factor binding
TMSB4X	**ENSG00000205542**	**1**	**8**	RNA binding, actin binding, protein binding, actin monomer binding, enzyme binding,
TMSB10	**ENSG000000034510**	**1**	**1**	actin monomer binding, actin binding, protein binding
GAPDH	**ENSG00000111640**			disordered domain specific binding, NAD binding, NADP binding, identical protein binding, peptidyl-cysteine S-nitrosylase activity, aspartic-type endopeptidase inhibitor activity, transferase activity, oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor, oxidoreductase activity, microtubule binding, protein binding, glyceraldehyde-3-phosphate dehydrogenase (NAD+) (phosphorylating) activity, nucleotide binding

Table 13. Functions of important to our classification experiments genes

7. Conclusions

The present study introduces a comprehensive classification pipeline for the analysis of single-cell gene expression profiles. By utilizing single-cell RNA sequencing (scRNA-seq) data obtained from the primary research article, our objective was to investigate the diversity within cell populations and ascertain the crucial characteristics that influence the classification procedure. The methodology employed in this study included several essential stages, which involved the exploration and preprocessing of data, the selection of relevant features, the fine-tuning of hyperparameters, and the implementation of classification using multiple classifiers.

In the initial stage of data exploration and preprocessing, we merged the clustering outcomes derived from various experiments into a cohesive dataset. A ground truth dataset for classification was generated by integrating cell labels and gene expression profiles. This procedure facilitated the establishment of a solid foundation for subsequent analysis.

Subsequently, we employed two feature selection methodologies, namely ANOVA and mRMR, in order to ascertain the most informative and discriminative features from the gene expression profiles. This approach allowed us to concentrate on pertinent genes while minimizing the impact of superfluous or inconsequential characteristics.

During the hyperparameter tuning phase, Optuna was employed to optimize the parameters of each classifier. The utilization of stratified 5-fold cross-validation was implemented to ensure a reliable assessment of performance and to address the issue of class imbalance present in the dataset. The enhanced hyperparameters yielded substantial improvements in the accuracy, precision, recall, and F1-score of the classifiers.

By considering the problem from the perspectives of a two-class, four-class, and eleven-class scenario, we were able to examine the resilience of our classification methodology. The classifiers exhibited favorable performance when applied to various cluster configurations, providing valuable insights into the potential biological implications of cellular heterogeneity.

In order to ascertain the most significant attributes in our classifiers, we utilized attribute-specific methodologies. The feature_importances_ attribute was employed for tree-based algorithms such as Random Forest and XGBoost, whereas the coef_ attribute was utilized for other classifiers such as Logistic Regression and SVC. This facilitated the extraction of the most influential features that drive the classification process.

The utilization of a data-driven approach enabled a more streamlined and comprehensible analysis in contrast to conventional marker gene-centric methodologies. Furthermore, it was observed that certain prominent characteristics identified by our classifiers did not necessarily align with the marker genes specified by the original authors. We were able to identify four important genes that were consistently influential across our classification tasks. These genes, namely RPS24, TMSB4X, TMSB10, and GAPDH, were further investigated, and we extracted valuable information about their molecular functions. This novel technique allows us to explore marker genes based on their functional significance rather than solely relying on differential expression analysis. This approach offers a broader perspective on the biological mechanisms driving classification outcomes.

In summary, our classification pipeline demonstrated efficacy in accurately categorizing individual cells by utilizing gene expression profiles. Promising results were obtained across various cluster configurations through the optimization of hyperparameters, feature selection, and the utilization of multiple classifiers. Our research makes a valuable contribution to the comprehension of cellular heterogeneity and establishes a foundation for future inquiries in the field of single-cell analysis. The integration of data-driven methodologies and machine learning algorithms presents fresh possibilities for investigating complex biological structures and the development of diseases at the individual cell level.

References

[1] Daniszewski M, Senabouth A, Nguyen QH, et al. Single cell RNA sequencing of stem cell-derived retinal ganglion cells. Scientific Data. 2018 Feb;5:180013. DOI: 10.1038/sdata.2018.13. PMID: 29437159; PMCID: PMC5810423.

[2] Tran NM, Shekhar K, Whitney IE, Jacobi A, Benhar I, Hong G, Yan W, Adiconis X, Arnold ME, Lee JM, Levin JZ, Lin D, Wang C, Lieber CM, Regev A, He Z, Sanes JR. Single-Cell Profiles of Retinal Ganglion Cells Differing in Resilience to Injury Reveal Neuroprotective Genes. Neuron. 2019 Dec 18;104(6):1039-1055.e12. doi: 10.1016/j.neuron.2019.11.006. Epub 2019 Nov 26. PMID: 31784286; PMCID: PMC6923571.

[3] Celotto, L.,Technische Universität Dresden: Deciphering the transcriptional states of Müller glia and their progeny in the regenerating zebrafish retina. Qucosa - Technische Universität Dresden: Deciphering the Transcriptional States of Müller Glia and Their Progeny in the Regenerating Zebrafish Retina. https://tud.qucosa.de/landing-page/?tx_dlf[id]=https%3A%2F%2Ftud.qucosa.de%2Fapi%2Fqucosa%253A86228%2Fmets

[4] Gramlich OW, Godwin CR, Wadkins D, Elwood BW, Kuehn MH. Early Functional Impairment in Experimental Glaucoma Is Accompanied by Disruption of the GABAergic System and Inceptive Neuroinflammation. Int J Mol Sci. 2021 Jul 15;22(14):7581. doi: 10.3390/ijms22147581. PMID: 34299211; PMCID: PMC8306430.

[5] Lukowski SW, Lo CY, Sharov AA, Nguyen Q, Fang L, Hung SS, Zhu L, Zhang T, Grünert U, Nguyen T, Senabouth A, Jabbari JS, Welby E, Sowden JC, Waugh HS, Mackey A, Pollock G, Lamb TD, Wang PY, Hewitt AW, Gillies MC, Powell JE, Wong RC. A single-cell transcriptome atlas of the adult human retina. EMBO J. 2019 Sep 16;38(18):e100811. doi: 10.15252/embj.2018100811. Epub 2019 Aug 22. PMID: 31436334; PMCID: PMC6745503.

[6] Whitmore SS, Wagner AH, DeLuca AP, Drack AV, Stone EM, Tucker BA, Zeng S, Braun TA, Mullins RF, Scheetz TE. Transcriptomic analysis across nasal, temporal, and macular regions of human neural retina and RPE/choroid by RNA-Seq. Exp Eye Res. 2014 Dec;129:93-106. doi: 10.1016/j.exer.2014.11.001. Epub 2014 Nov 5. PMID: 25446321; PMCID: PMC4259842.

[7] Maciej Daniszewski, Anne Senabouth, Helena H. Liang, Xikun Han, Grace E. Lidgerwood, Damián Hernández, Priyadharshini Sivakumaran, Jordan E. Clarke, Shiang Y. Lim, Jarmon G. Lees, Louise Rooney, Lerna Gulluyan, Emmanuelle Souzeau, Stuart L. Graham, Chia-Ling Chan, Uyen Nguyen, Nona Farbehi, Vikkitharan Gnanasambandapillai, Rachael A. McCloy, Linda Clarke, Lisa S. Kearns, David A. Mackey, Jamie E. Craig, Stuart MacGregor, Joseph E. Powell, Alice Pébay, Alex W. Hewitt,
Retinal ganglion cell-specific genetic regulation in primary open-angle glaucoma, Cell Genomics,Volume 2, Issue 6,2022,100142,ISSN 2666-979X, https://doi.org/10.1016/j.xgen.2022.100142.

[8] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

[9] Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. John Wiley & Sons.

[10] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Supplementary Material

Appendix

It is noteworthy that the Python code utilized to generate the outcomes was composed in Google Colab, which is a cloud-based development platform that offers free access to GPUs and TPUs. The utilization of Google Colab proved to be beneficial due to its facilitation of expeditious and consequently, it is advisable to utilize Google Colab in order to replicate the findings outlined in this manuscript.
To achieve accurate replication of the results, it is recommended to carefully follow the steps presented in the following demonstration:

1. *Open the notebook provided using Google Colab*
2. *Access the data section of the environment by clicking on the folder icon located on the left side. Then, upload the files mentioned in the start of every notebook. All files needed are provided within the exercise's zip file. To upload them, simply drag and drop the files into the designated area.*
3. *Run the cells in the notebook in the order they appear, making sure to follow the instructions provided in each cell.*

When executing the notebook in Jupyter, it is essential to ensure that the necessary packages have been installed. Additionally, it is important to specify the path to the files an input parameter to all the functions needed.