Feature Selection in single-cell RNA-seq data via a Genetic Algorithm

Konstantinos I. Chatzilygeroudis[1,2], Aristidis G. Vrahatis[3], Sotiris K. Tasoulis[3], and Michael N. Vrahatis[2]

[1] Computer Engineering and Informatics Department (CEID),
University of Patras, Greece,
costashatz@upatras.gr

[2] Computational Intelligence Laboratory, Department of Mathematics, University of Patras, Greece,
vrahatis@math.upatras.gr

[3] Department of Computer Science and Biomedical Informatics,
University of Thessaly, Greece,
arisvrahatis@uth.gr, stasoulis@uth.gr

Abstract. Big data methods prevail in the biomedical domain leading to effective and scalable data-driven approaches. Biomedical data are known for their ultra-high dimensionality, especially the ones coming from molecular biology experiments. This property is also included in the emerging technique of single-cell RNA-sequencing (scRNA-seq), where we obtain sequence information from individual cells. A reliable way to uncover their complexity is by using Machine Learning approaches, including dimensional reduction and feature selection methods. Although the first choice has had remarkable progress in scRNA-seq data, only the latter can offer deeper interpretability at the gene level since it highlights the dominant gene features in the given data. Towards tackling this challenge, we propose a feature selection framework that utilizes genetic optimization principles and identifies low-dimensional combinations of gene lists in order to enhance classification performance of any off-the-shelf classifier (e.g., LDA or SVM). Our intuition is that by identifying an optimal genes subset, we can enhance the prediction power of scRNA-seq data even if these genes are unrelated to each other. We showcase our proposed framework's effectiveness in two real scRNA-seq experiments with gene dimensions up to 36708. Our framework can identify very low-dimensional subsets of genes (less than 200) while boosting the classifiers' performance. Finally, we provide a biological interpretation of the selected genes, thus providing evidence of our method's utility towards explainable artificial intelligence.

Keywords: Feature Selection · Optimization · single-cell RNA-seq · High-dimensional data.

1 Introduction

Almost two decades ago, the Human Genome Project [11] was completed, where the human genome was analyzed, offering the complete set of genetic information. The evolution of DNA sequencing from this point has undoubtedly brought about a great revolution in the field of biomedicine [29]. Although new technologies and analysis tools are constantly emerging, their experimental data have ultra-high dimensionality hindering success of traditional methods [8]. Hence, data mining approaches in such data create several computational challenges that novel or updated existing computational methodologies can address. Indicatively, a gene expression experiment includes each sample measurements for the entire genome, which contains tens of thousands of genes.

Meanwhile, classification in gene expression profiles is a longstanding research field with remarkable progress in complex disease identification, and treatment [13]. Started with data from the microarrays high-throughput technology [6] and continued with sequencing data [35]. We are now in the single-cell sequencing era, which allows biological information to be extracted from individual cells offering a deeper analysis at the cellular level. An indicative undercase transcriptomics study has gene measurements simultaneously for the entire genome isolating hundreds or thousands or even millions of cells in recent years. Given that we obtain measurements of tens of thousands of genes for each cell, in the computational perspective, we have to manage single-cell RNA-sequencing (scRNA-seq) data with ultra-high complexity.

Several single-cell RNA-seq data challenges are addressed through classification methods under the Machine Learning family [26]. These methods shed light on various biological issues such as the new cell types of identification [27], the cellular heterogeneity dissection [16], the cell cycle prediction [28], the cell sub-populations [7], the cells classification [25] and much more [33]. Despite the remarkable progress and promising results in the above challenges, the increasing scRNA-seq data generation, and the related technologies improvements creates new challenges and the need for novel classification methods under the perspective of supervised learning.

The nature of scRNA-seq technology, that is to examine individual cells from specific tissues, creates a quite sparse counts matrix since for every cell usually exists a high fraction of genes which are not informative [31]. Two appropriate ways to deal with this inherent particularity are dimensionality reduction techniques and feature selection methods. Dimension reduction techniques in scRNA-seq data aim to transfer the original $\mathbb{R}^D$ cell's space, where D is the genes expression profiles, to a lower-dimensional $\mathbb{R}^K$ space, with $K \ll D$. Such methods have gained ground in recent years with promising results in visualization [34] as in classification performance tasks [26]. Indicatively, the t -distributed stochastic neighbor embedding [21] and uniform manifold approximation and projection [5] techniques are usually applied in scRNA-seq data to obtain low-dimensional embedding offering a better visualization to uncover the relationship among cells and their categories.

However, their major drawback is that the reduced-dimensional projected space does not contain information about each gene since the original space has been transformed. It does not allow us a further biological analysis and a deeper interpretation of a given case under study (disease, biological process).

In gene expressions data, feature selection or variable selection is selecting a subset of genes for model construction or results interpretation. It has been shown that feature selection in such data improves the classification performance and offers the potential for a better data interpretation in a given study since we know its dominant and redundant genes (features), which are related to the various class separation. There are numerous feature selection algorithms with promising results in gene expression data [20, 1, 12]. In scRNA-seq data, feature selection methods aim to identify uninformative genes (features) with no meaningful biological variation across cells (samples) [31]. Identifying the appropriate set of marker genes interprets the scRNA-seq data at the gene level with a deeper biological meaning [3, 30].

Some studies considered the feature selection problem in such data as an optimization task from the mathematical perspective. In [19], the authors described a fitness function that incorporates both performance and feature size. Applying the Particle Swarm Optimization (PSO) method and the utilization of Convolutional Neural Networks, they offer promising results in classifying the different types of cancer based on tumor RNA-Seq gene expression data.

In [24], a feature selection approach is proposed for RNA-seq gene expression data. It reduces the irrelevant features by applying an ensemble L1-norm support vector machine methodology. Its classification performance in RNA-seq data shown promising results, especially in small n – large p problems, with n samples and p features (genes). scTIM [15] framework utilizes a multiobjective optimization technique aiming to maximize gene specificity by considering the gene-cell relationship while trying to minimize the number of selected genes. This model allows the new cell type discovery as well as the better cell categories separation. M3Drop [3], describes two feature selection methods for scRNA-seq data which isolate genes with the high proportion of zero values among their cells, also called the “dropouts” effect. It is a central feature of scRNA-seq due to the considerable technical and biological noise.

Despite the remarkable progress of feature selection methods in gene expressions, their adaptation in single-cell RNA-seq is at a very early stage. Given that these data have high complexity, dimensionality, and sparsity, lead us on the necessity of incorporating an optimization method for the appropriate feature (genes) selection. Our intuition here is that across the genome, several combinations of certain genes will be dominant in cell separation of a given experimental study.

In this paper, we propose a novel feature selection method and analysis, called Feature Selection via Genetic Algorithm (FSGA), that tackles the above challenges. FSGA utilizes genetic optimization principles and identifies low-dimensional sets of features. The aim here is to use a simple distance-based classifier (we use a KNN classifier) during the feature selection process in order to identify feature groups that are nicely separated in Euclidean space. This property is desirable in most classification methods, and thus we expect to boost the performance of any classifier. We showcase FSGA’s effectiveness in two real scRNA-seq experiments with gene dimensions up to 36708. Our framework can identify very low-dimensional subsets of genes (less than 200) while boosting the classifiers' performance. The obtained results offer new insights in the single-cell RNA-seq data analysis offering the potential that variants of the proposed method can work also in other data types.

2 Approach

2.1 Problem Formulation

scRNA-seq datasets have very high-dimensional feature spaces, with features spaces going up to $D = 30K$ dimensions. We would like to find a group of K features where $K \ll D$ and we can achieve similar or even better classification performance. We represent the feature space as $\boldsymbol{x} = [x_1, x_2, \dots, x_D]^T \in \mathbb{R}^D$ and define the problem of feature selection as (see also [12, 1]):

$$\boldsymbol{f}^* = \operatorname*{argmax}_{\boldsymbol{f}} J(\boldsymbol{x}_{\boldsymbol{f}}) \tag{1}$$

where $\boldsymbol{f} = [b_1, b_2, \dots, b_D]^T \in \mathbb{B}^D$ with b_i being a Boolean[1] value whether we select the dimension i, $\boldsymbol{x}_{\boldsymbol{f}} \in \mathbb{R}^K$ is the feature dimension vector where we keep only the dimensions as defined by $\boldsymbol{f}$, and J is training and evaluating the performance of a given feature vector.

2.2 Feature Selection via Genetic Algorithms

We choose to tackle this problem using a Genetic Algorithm (GA). GAs operate on a population of individuals and attempt to produce better individuals every generation. At each generation, a new population is created by selecting individuals according to their level of performance and recombining them together using operators borrowed from natural evolution. Offspring might also undergo a mutation operation. In more detail, any GA has the following generic steps:

1. Initialization of the population
2. Evaluation of the population
3. Selection of the fittest individuals
4. Crossover between some of the selected individuals
5. Mutation of some individuals
   The previous two steps produce a new population
6. Go back to step 2

There a few critical parameters to choose so that a GA can be effective: (1) gene representation, (2) selection pressure, (3) crossover operation, (4) mutation operation, (5) initialization of the population and (6) performance measure. Below we detail our choices.

[1] We define $\mathbb{B}$ as the space of Boolean variables.

Gene representation In order to be able to use GAs for solving the problem as defined in Eq. (1), we use the vectors $\boldsymbol{f} = [b_1, b_2, \ldots, b_D]^T \in \mathbb{B}^D$ for the gene representation. This is a natural choice for this task as changing the values in the gene will correspond in selecting different features for the classification [17, 14, 1, 12].

Selection operator We adopt a selection operator that selects the top-50% individuals of the population (according to the performance measure). More sophisticated selection operators can be used here to improve performance. We also always insert the best individual back into the new population (thus making the algorithm elitist).

Crossover operator The crossover operator consists of combining two (2) individuals (called parents) to produce a new one (offspring). We randomly determine parts of the gene parent vectors to be swapped.

Mutation operator Each offspring individual can undergo a mutation operator. For each individual we randomly switch any dimension of its gene vector. So, with some probability we change which features the offspring keeps for the classification.

Initialization of the population One crucial aspect of the initial population is to push for as little as possible number of selected features, but not hurt performance. For this reason, we produce the initial population where for each individual each feature dimension has an 1‰ chance of being selected. This procedure produces populations with small number of selected features, but keeps diversity in which feature dimensions are being selected.

Objective Function (Performance Measure) When optimizing for the best features, in each run of the algorithm we split the datasets into three sets: (a) training set, (b) validation set, and (c) test set. The sets are roughly 60%, 20% and 20% of the size of the original dataset respectively (keeping the percentage of classes similar in each dataset). At each evaluation, we use the training set to train the KNN-classifier, and create an objective function of the form:

$$J(\boldsymbol{x}_{\boldsymbol{f}}) = 0.6 * \mathrm{acc}_{\mathrm{val}} + 0.4 * \mathrm{acc}_{\mathrm{train}} - P_{\mathrm{sparseness}} \tag{2}$$

where $\mathrm{acc}_{\mathrm{val}}$ is the accuracy of the classifier in the validation set, $\mathrm{acc}_{\mathrm{train}}$ is the accuracy of the classifier in the training set, and $P_{\mathrm{sparseness}} = 10 * \sum_{i=1}^{D} b_i$ is a penalty score penalizing high dimensionality of the selected feature space. The proposed objective function is slightly different from the ones in the literature [14, 17, 1, 12]; we are doing the weighted average of the validation and the training set accuracy. The reasoning behind this weighted average is to not let the algorithm overfit a specific part of the dataset. At the end of each generation, we report the accuracy on the test set (see Sec. 3), but the algorithm never uses this.

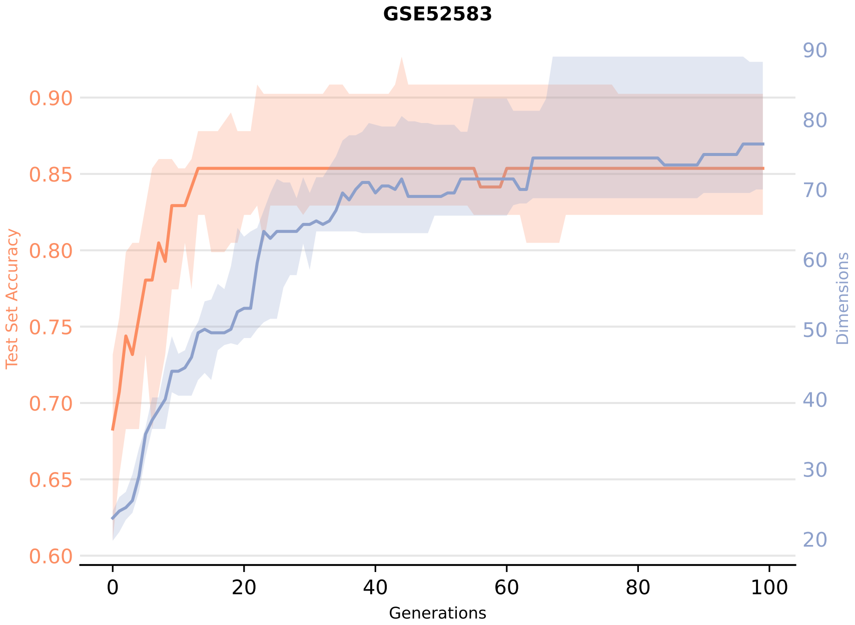


Fig. 1: Convergence of algorithm in GSE52583 dataset. Solid lines are the median over 20 replicates and the shaded regions are the regions between the 25-th and 75-th percentiles.

3 Experimental Analysis on scRNA-seq Datasets

We evaluated the classification performance of our FSGA method using two real transcriptomics datasets from single-cell RNA-seq studies. Datasets were obtained from Gene Expression Omnibus [10] and ArrayExpress [4]. More specific, the first dataset (accession number: GSE52583) [32] has transcriptomics experimental data profiles for 23, 228 genes. It is a transcriptome analysis of 201 distal mouse lung epithelial cells from four developmental stages. The second dataset (accession number: E-MTAB-2805) has studied expression patterns for 36078 genes at single cell level across the different cell cycle stages in 288 mouse embryonic stem cells [7].

The evaluation process was split into to three (3) parts: (a) evaluating the optimization process and whether our proposed scheme was converging to good individuals in all runs, (b) evaluating whether the produced features provide a good set of features for any classifier, and (c) try to determine whether the selected features have a biological meaning.

To tackle both challenges we chose to use a simple KNN classifier when optimizing for the best features. The KNN performs k-nearest-neighbor classification model [2] using the default parameters with Euclidean as distance measure as well KD-tree option as search method for $N = 5$ nearest neighbors. The rationale behind this choice is that a) KNN is fast, and b) a set of features that works well under KNN directly means that these features are nicely separated in Euclidean space. The first fact gave us the ability to run many replicates and have meaningful comparisons and statistics, while the second one makes it more likely for other classifiers to work well (see below).

If not mentioned otherwise, all plots are averaged (or taking median/percentiles) over 20 replicates.

3.1 Evaluation of Feature Selection Process

In order to evaluate the feature selection process, we keep track of the best individual of the optimization at each generation as well as the number of selected feature dimensions of the best individual.

The results show that the optimization is able to find high-performing individuals (see Fig. 1 and Fig. 2). In both datasets, we achieve a median accuracy score over 0.75 in the test set (this is the set that both the classifier and the optimizer have never seen). This showcases that our objective function is able to produce classifiers with nice generalization properties.

Moreover, the results demonstrate that the optimization process increases the dimensionality of the feature space as long as this helps the process get better performance. Once the performance stabilizes to a fixed value, the dimensionality of the feature space stops increasing. This is a desirable property of a feature selection process since we do not want it to keep adding dimensions if they do not help in the classification performance. The algorithm converges at around 77 dimensions for the GSE52583 dataset and around 165 dimensions for E-MTAB-2805 dataset (median values over 20 replicates). Our initialization process is crucial for achieving these results (see Sec. 2.2), as preliminary results with a population with individuals containing many dimensions did not manage to converge to low number of features.

3.2 Evaluation of Selected Features

In this section, we want to evaluate the quality of the selected features both quantitatively and qualitatively. For a principled analysis, we perform the following steps:

- For each run of our algorithm[2], we take the feature dimensions of the best individual at convergence;
- We take those feature dimensions and modify the datasets (i.e., we include only those input dimensions);
- Using the modified datasets we train three (3) different classification methods, namely KNN, LDA and SVM [9, 23];

[2] We have 20 runs/replicates.

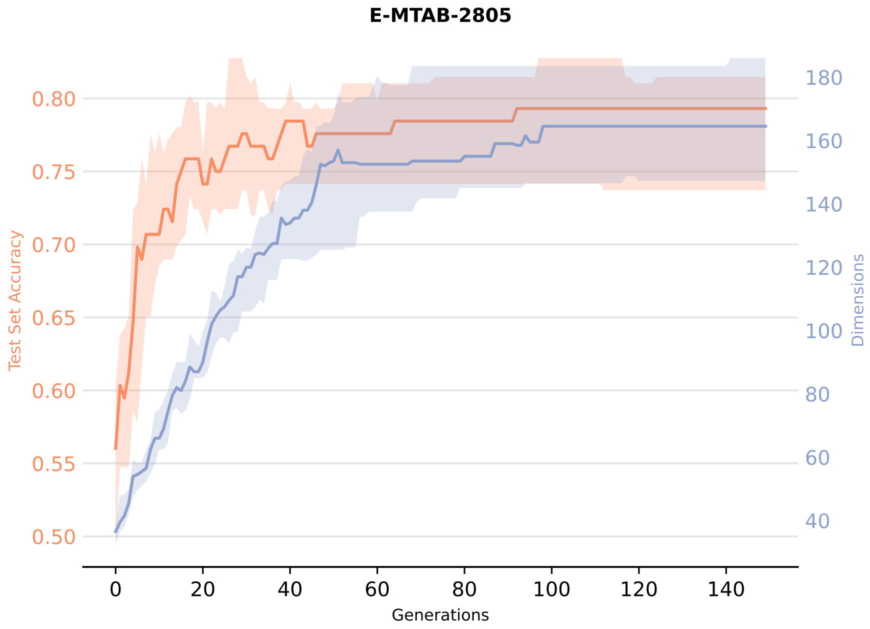


Fig. 2: Convergence of algorithm in E-MTAB-2805 dataset. Solid lines are the median over 20 replicates and the shaded regions are the regions between the 25-th and 75-th percentiles.

- We compare the performance of the algorithms using our selected features against training the same classifiers using all the feature dimensions (we use Accuracy and F1-score for comparisons);
- All executions are done using the 10-fold cross validation process in 20 independent trials.

Parameter setting for all methods was chosen based on a fitting procedure in order to optimize their performance. Minor variations for the selected values do not affect the results significantly and thus an extensive analysis is excluded. All algorithms were run with the corresponding default parameters. We exclude an extensive parameter analysis of all classifiers since our aim was to highlight for each classifier its difference between the classification performance in the original and in the reduced feature space.

The results showcase that in almost all cases the feature space produced by our algorithm increases the performance of any classification method (see Fig. 3). In all cases, except when using LDA on the E-MTAB-2805 dataset, our feature selection approach boosts significantly the performance of all classifiers. Even in the worst case (LDA/E-MTAB-2805), the result of the classification is comparable with training in the full feature dimensions by using only 180

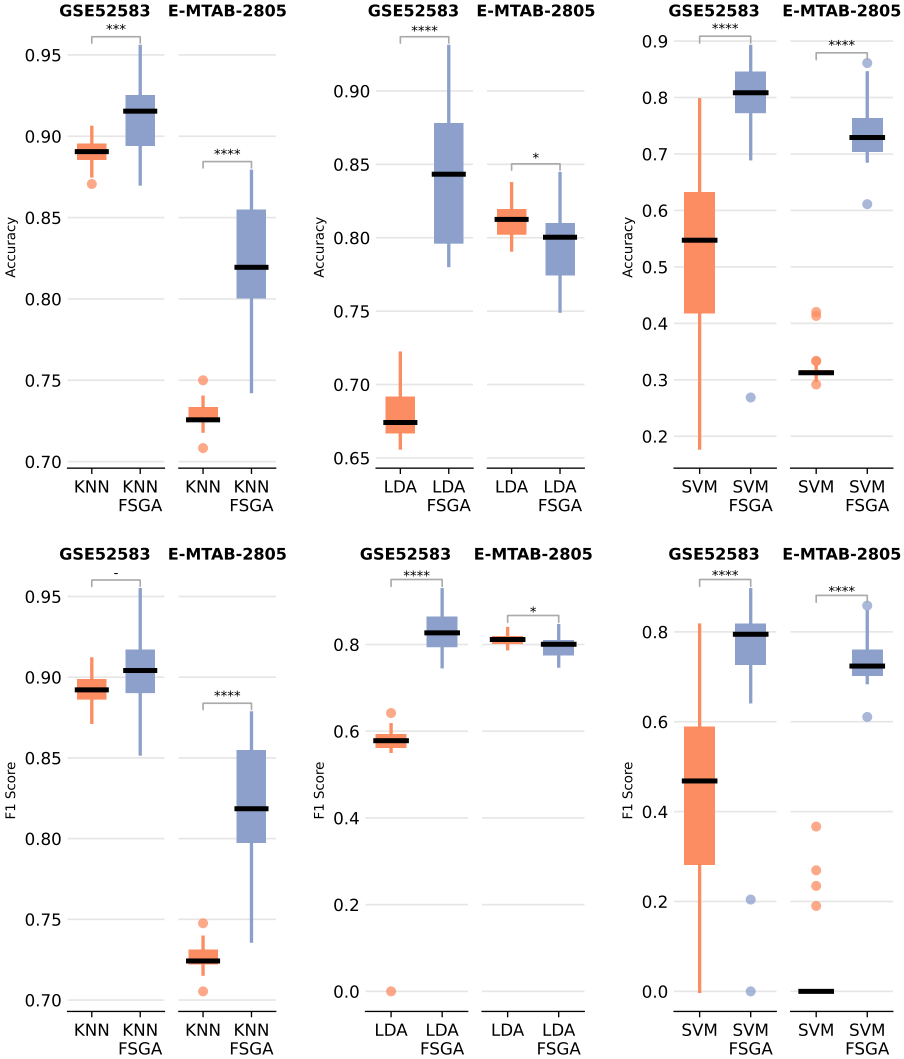


Fig. 3: Evaluation of selected features using different classifiers (20 replicates). We compare using two (2) metrics: Accuracy (top row) and F1-Score (bottom row). For each algorithm, we show results before and after the usage of our feature selection algorithm for both datasets. The box plots show the median (black line) and the interquartile range (25*th* and 75*th* percentiles); the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. The number of stars indicates that the p-value of the Mann-Whitney U test is less than 0.05, 0.01, 0.001 and 0.0001 respectively.

dimensions. The performance improvement to the SVM classifier is highlighting the effectiveness of our approach to generate separable feature spaces.

Fig. 4 shows tSNE plots [22] of one typical feature selection run of the algorithm in the GSE52583 and the E-MTAB-2805 datasets respectively for qualitative inspection/verification. The plots showcase the effectiveness of the proposed method to find feature dimensions that can separate the classes in different regions of the space.

3.3 Biological Analysis

We further examine the selected genes for each dataset concerning their enrichment in Gene Ontology terms for various Biological Processes (see Table 1) using the Functional Annotation Tool David [18]. Through this analysis, we aim to examine how our list of selected genes relates to terms corresponding to the respective biological case under-study. Both datasets extract genes which are related to cellular functions. Both studies are relevant with these functions since their studies are related with the developmental stages of distal mouse lung epithelial cells and the different cell cycle stages in Mouse Embryonic Stem Cells.

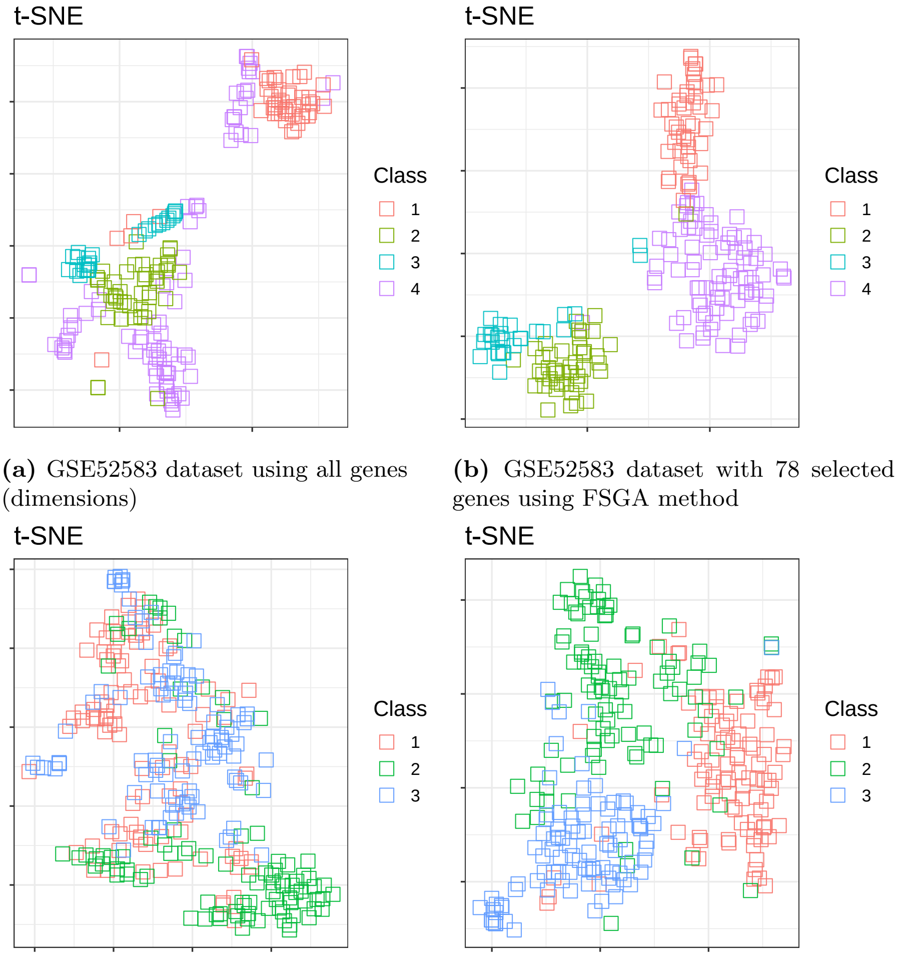
4 Discussion & Conclusion

Machine Learning tasks have become the first choice for gaining insight into large-scale and high-dimensional biomedical data. These approaches can tackle part of such data complexity offering a platform for effective and robust computational methods. Part of this complexity comes from a plethora of molecular biology experimental data having extremely high dimensionality. An indicative example is the single-cell RNA-seq (scRNA-seq), an emerging DNA sequencing technology with promising capabilities but significant computational challenges due to the large-scaled generated data.

Given that this technology offers the opportunity to understand various biological phenomena and diseases better, there is a need for novel computational methods to deal with this complexity and dimensionality. Dimensionality reduction methods are an appropriate choice, but they do not give us explanatory power at the gene level. A significant challenge here is identifying the feature list in terms of genes (dimensions), which will maintain or increase the performance of various machine learning tasks.

Highlighting the salient by eliminating the irrelevant features in a high dimensional dataset such as the high-throughput gene expression experiments, may lead to the strengthening of a traditional classifier's performance [24]. Also, given a features list which is dominant in terms of class separation in a classification process, we obtain a better understanding and interpretation of a given gene expressions dataset.

On the other hand, deep learning has gained ground in biomedical data mining methods. However, its inherent black-box feature offers a poor interpretability for a better understanding of such data. In the case of gene expressions where the data contain a record of tens of thousands of genes, it is crucial to find the genes and especially the combination of these genes, which will better



(a) GSE52583 dataset using all genes (dimensions)

(b) GSE52583 dataset with 78 selected genes using FSGA method

(c) E-MTAB-2805 dataset using all genes (dimensions)

(d) E-MTAB-2805 dataset with 180 selected genes using FSGA method

Fig. 4: 2D t-SNE visualizations are illustrated for comparisons between the original datasets and the datasets with reduced features using our FSGA method. Each point represents a cell sample, and each color represents a different cell type according to original data annotation. Our method shows its superiority by efficiently discriminating the cell classes in both datasets.

capture the information contained in the data set. Also, through the developing of interpretable ML approaches offers the opportunity not only for a better data interpretation but also for finding the dominant genes which may need to be considered individually or in combination for their potential effect on the under-study case (e.g. a disease, a biological process).

Through our proposed feature selection method using a Genetic Algorithm, we provided evidence about its potential in single-cell RNA-sew analysis re-

Term	Count	P-Value
GSE52583		
negative regulation of cellular process	55	9.0E-2
cellular macromolecule metabolic process	127	1.2E-10
cellular nitrogen compound metabolic process	96	1.9E-6
positive regulation of cellular metabolic process	54	3.4E-6
cellular catabolic process	31	8.7E-5
cellular response to chemical stimulus	44	1.0E-3
response to extracellular stimulus	6	3.8E-2
regulation of secretion by cell	7	6.2E-2
negative regulation of cell differentiation	7	9.4E-2
EM-TAB-2805		
intracellular transport	11	2.5E-2
establishment of localization in cell	13	2.8E-2
positive regulation of cell communication	11	5.4E-2
cell communication	31	7.0E-2
regulation of cellular component size	6	9.0E-2
circulatory system process	8	2.2E-3
response to extracellular stimulus	6	3.8E-2
regulation of secretion by cell	7	6.2E-2
negative regulation of cell differentiation	7	9.4E-2

Table 1: Enrichment analysis for GSE52583 and E-MTAB-2805 datasets using gene ontology terms of the selected features obtained from the proposed framework. The first column contains the gene ontology terms for various biological processes. The second column represents the number of genes that present enrich action in each term. The third column represents a modified Fisher's exact p-value.

garding the classification performance. Our intuition was that an optimal combination of genes could improve both the classification performance and the interpretability of a given data. The first is critical since even if this feature subset contains genes unrelated to each other, their combination might be highly correlated with the classification. The latter can contribute in the emerging explainable artificial intelligence field.

The obtained results offer new insights in the single-cell RNA-seq data analysis offering the potential that variants of the proposed method can work also in other data types. Our contribution, which is lies in the intuition that specific combinations of small gene groups have a key role in our scRNA-seq data, is partly confirmed by the above results.

References

1. Alba, E., Garcia-Nieto, J., Jourdan, L., Talbi, E.G.: Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. In: 2007 IEEE congress

on evolutionary computation. pp. 284–290. IEEE (2007)

2. Altman, N.S.: An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician **46**(3), 175–185 (1992)
3. Andrews, T.S., Hemberg, M.: M3drop: dropout-based feature selection for scrnaseq. Bioinformatics **35**(16), 2865–2867 (2019)
4. Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C., Fonseca, N.A., Petryszak, R., Papatheodorou, I., et al.: Arrayexpress update–from bulk to single-cell expression data. Nucleic acids research **47**(D1), D711–D715 (2019)
5. Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W., Ng, L.G., Ginhoux, F., Newell, E.W.: Dimensionality reduction for visualizing single-cell data using umap. Nature biotechnology **37**(1), 38 (2019)
6. Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Haussler, D.: Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences **97**(1), 262–267 (2000)
7. Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., Stegle, O.: Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. Nature biotechnology **33**(2), 155–160 (2015)
8. Chattopadhyay, A., Lu, T.P.: Gene-gene interaction: the curse of dimensionality. Annals of translational medicine **7**(24) (2019)
9. Chatzilygeroudis, K., Hatzilygeroudis, I., Perikos, I.: Machine learning basics. In: Intelligent Computing for Interactive System Design: Statistics, Digital Signal Processing, and Machine Learning in Practice, pp. 143–193 (2021)
10. Clough, E., Barrett, T.: The gene expression omnibus database. In: Statistical genomics, pp. 93–110. Springer (2016)
11. Collins, F.S., Morgan, M., Patrinos, A.: The human genome project: lessons from large-scale biology. Science **300**(5617), 286–290 (2003)
12. Dhaenens, C., Jourdan, L.: Metaheuristics for data mining. 4OR **17**(2), 115–139 (2019)
13. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American statistical association **97**(457), 77–87 (2002)
14. Estévez, P.A., Caballero, R.E.: A niching genetic algorithm for selecting features for neural network classifiers. In: International Conference on Artificial Neural Networks. pp. 311–316. Springer (1998)
15. Feng, Z., Ren, X., Fang, Y., Yin, Y., Huang, C., Zhao, Y., Wang, Y.: sctim: seeking cell-type-indicative marker from single cell rna-seq data by consensus optimization. Bioinformatics **36**(8), 2474–2485 (2020)
16. Hedlund, E., Deng, Q.: Single-cell rna sequencing: technical advancements and biological applications. Molecular aspects of medicine **59**, 36–46 (2018)
17. Hong, J.H., Cho, S.B.: Efficient huge-scale feature selection with speciated genetic algorithm. Pattern Recognition Letters **27**(2), 143–150 (2006)
18. Huang, X., Liu, S., Wu, L., Jiang, M., Hou, Y.: High throughput single cell rna sequencing, bioinformatics analysis and applications. In: Single Cell Biomedicine, pp. 33–43. Springer (2018)
19. Khalifa, N.E.M., Taha, M.H.N., Ali, D.E., Slowik, A., Hassanien, A.E.: Artificial intelligence technique for gene expression by tumor rna-seq data: a novel optimized deep learning approach. IEEE Access **8**, 22874–22883 (2020)

20. Liang, S., Ma, A., Yang, S., Wang, Y., Ma, Q.: A review of matched-pairs feature selection methods for gene expression data analysis. Computational and structural biotechnology journal **16**, 88–97 (2018)
21. Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S., Kluger, Y.: Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data. Nature methods **16**(3), 243–245 (2019)
22. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008)
23. McLachlan, G.J.: Discriminant analysis and statistical pattern recognition, vol. 544. John Wiley & Sons (2004)
24. Moon, M., Nakai, K.: Stable feature selection based on the ensemble l 1-norm support vector machine for biomarker discovery. BMC genomics **17**(13), 65–74 (2016)
25. Poirion, O.B., Zhu, X., Ching, T., Garmire, L.: Single-cell transcriptomics bioinformatics and computational challenges. Frontiers in genetics **7**, 163 (2016)
26. Qi, R., Ma, A., Ma, Q., Zou, Q.: Clustering and classification methods for single-cell rna-sequencing data. Briefings in bioinformatics **21**(4), 1196–1208 (2020)
27. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al.: Science forum: the human cell atlas. Elife **6**, e27041 (2017)
28. Scialdone, A., Natarajan, K.N., Saraiva, L.R., Proserpio, V., Teichmann, S.A., Stegle, O., Marioni, J.C., Buettner, F.: Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods **85**, 54–61 (2015)
29. Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., Waterston, R.H.: Dna sequencing at 40: past, present and future. Nature **550**(7676), 345 (2017)
30. Taguchi, Y.: Principal component analysis-based unsupervised feature extraction applied to single-cell gene expression analysis. In: International Conference on Intelligent Computing. pp. 816–826. Springer (2018)
31. Townes, F.W., Hicks, S.C., Aryee, M.J., Irizarry, R.A.: Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. Genome biology **20**(1), 1–16 (2019)
32. Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., Quake, S.R.: Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. Nature **509**(7500), 371 (2014)
33. Vrahatis, A.G., Tasoulis, S.K., Maglogiannis, I., Plagianakos, V.P.: Recent machine learning approaches for single-cell rna-seq data analysis. In: Advanced Computational Intelligence in Healthcare-7, pp. 65–79. Springer (2020)
34. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., Batzoglou, S.: Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. Nature methods **14**(4), 414 (2017)
35. Witten, D.M., et al.: Classification and clustering of sequencing data using a poisson model. The Annals of Applied Statistics **5**(4), 2493–2518 (2011)