applied sciences

MDPI

Article

An Ensemble Feature Selection Approach for Analysis and Modeling of Transcriptome Data in Alzheimer's Disease

Petros Paplomatas *, Marios G. Krokidis *, Panagiotis Vlamos and Aristidis G. Vrahatis

Bioinformatics and Human Electrophysiology Laboratory, Department of Informatics, Ionian University, 49100 Corfu, Greece

* Correspondence: p.paplomatas@hotmail.com (P.P.); mkrokidis@ionio.gr (M.G.K.)

Abstract: Data-driven analysis and characterization of molecular phenotypes comprises an efficient way to decipher complex disease mechanisms. Using emerging next generation sequencing technologies, important disease-relevant outcomes are extracted, offering the potential for precision diagnosis and therapeutics in progressive disorders. Single-cell RNA sequencing (scRNA-seq) allows the inherent heterogeneity between individual cellular environments to be exploited and provides one of the most promising platforms for quantifying cell-to-cell gene expression variability. However, the high-dimensional nature of scRNA-seq data poses a significant challenge for downstream analysis, particularly in identifying genes that are dominant across cell populations. Feature selection is a crucial step in scRNA-seq data analysis, reducing the dimensionality of data and facilitating the identification of genes most relevant to the biological question. Herein, we present a need for an ensemble feature selection methodology for scRNA-seq data, specifically in the context of Alzheimer's disease (AD). We combined various feature selection strategies to obtain the most dominant differentially expressed genes (DEGs) in an AD scRNA-seq dataset, providing a promising approach to identify potential transcriptome biomarkers through scRNA-seq data analysis, which can be applied to other diseases. We anticipate that feature selection techniques, such as our ensemble methodology, will dominate analysis options for transcriptome data, especially as datasets increase in volume and complexity, leading to more accurate classification and the generation of differentially significant features.

Keywords: ensemble method; big data; dimensionality reduction; feature selection; Alzheimer's disease

Citation: Paplomatas, P.; Krokidis, M.G.; Vlamos, P.; Vrahatis, A.G. An Ensemble Feature Selection Approach for Analysis and Modeling of Transcriptome Data in Alzheimer's Disease. *Appl. Sci.* **2023**, *13*, 2353. https://doi.org/10.3390/app13042353

Academic Editor: Je-Keun Rhee

Received: 12 December 2022
Revised: 7 February 2023
Accepted: 9 February 2023
Published: 11 February 2023

1. Introduction

High throughput molecular biology technologies, such as whole-genome sequencing, have been used extensively in a research capacity to investigate disease mechanisms, allowing a deeper analysis at the cellular level and more reliable results [1]. In the era of personalized medicine there is a constant need to develop robust computational approaches for big data analysis in order to explore molecular entities and hidden patterns in a more realistic manner [2]. Towards overcoming these challenges, remarkable progress has been observed through different generations of sequencing technology providing explosive growth in omics data [3]. Single-cell sequencing, as an emerging technique, has revolutionized the way diseases are studied at the cellular level. This technology can accurately study individual cells and explore pathological mechanisms at the single-cell level to indicate diagnostic biomarkers or potent therapeutic targets [4]. The single-cell RNA sequencing (scRNA-seq) technique allows detection and quantitative analysis of mRNA molecules at a single cell resolution instead of bulk RNAseq studies which investigate global gene expression. It is a complex process, which involves single cell isolation and capture, lysis of cells, reverse transcription, amplification, and library preparation [5,6]. However, although experimental approaches are rapidly increasing, in silico pipelines for handling raw data files remain limited. A typical scRNA-seq dataset includes thousands of cells and their

corresponding transcripts, therefore, from a computational perspective, we have to deal with a large amount of complex data (features) for each cell with ultra-high dimensionality and huge volume [7].

To effectively analyze these complex data, feature selection is a critical aspect of machine learning data preprocessing, which has gained significant importance in recent years due to the exponential growth in the size of real-world datasets [8]. The feature selection process aims to identify and select relevant features to improve model performance and efficiency. By removing irrelevant features, it can also improve the efficiency of the algorithm. In the context of scRNA-seq data, feature selection is particularly important as it can help identify the most important genes and cellular pathways, leading to a better understanding of the cellular functionality [9].

Machine learning (ML) processes are the first choice for mining information from such datasets as they can tackle part of this data complexity. In terms of high dimensionality due to the large number of feature spaces (genes), ML approaches to feature extraction can address this challenge by removing as much noisy and redundant information from the extracted features as possible [10]. In some cases, feature importance scores generated by ML methods may be unreliable, which can lead to instability and make it challenging to distinguish between important and unimportant features [11]. While filter and wrapper feature selection methods are commonly used, they also have some limitations. One common issue with filter approaches is that they tend to overlook how the classifier interacts with each feature and evaluate each feature independently, neglecting feature correlations. Additionally, it is difficult to determine the appropriate threshold for selecting only relevant features and excluding noise [12]. Wrapper feature selection also has several drawbacks, including a greater risk of overfitting, longer training times, reliance on a specific classifier, and higher computational cost and discriminatory power [13].

Ensemble feature selection frameworks provide a dependable method to decrease the dimensionality of high-dimensional data and pinpoint the most meaningful features in a case study [14,15]. An ensemble method that combines the results of multiple feature selection techniques can overcome the limitations of using single feature selection methods. This approach can lead to improved performance and robustness compared to any individual method. Therefore, it is important to utilize specific feature selection methods to minimize irrelevant and redundant features, which will decrease complexity in the analysis and models and improve the overall modeling process efficiency [16]. In a recent study, a gene selection pipeline combining filter, wrapper, and unsupervised methods was developed to select the relevant features in causing Alzheimer's disease (AD) [17]. An ensemble method based on consensus-guided unsupervised feature selection was proposed to identify Huntington's disease-associated genes [18].

In this study, we present the scEFS (ensemble feature selection for single-cell RNA-seq data) method, an ensemble framework that takes advantages of three different feature selection methods for single-cell RNA-seq data. These methods are combined using a ranking strategy to determine the best genes. We evaluate the effectiveness of various feature selection strategies in a prediction task using single-cell RNA-sequencing data from a mouse model with Alzheimer's-like pathology and phenotypes (CK-p25 model). Our goal is to demonstrate the complexity of scRNA-seq datasets and investigate the impact of ensemble feature selection on the prediction accuracy using traditional classifiers.

2. Materials and Methods

Our scEFS approach is summarized in two main steps: (a) the application of three main feature selection methods; (b) the genes prioritization through an ensemble voting technique, called Borda.

2.1. Feature Selection Techniques

DUBStepR (determining the underlying basis using stepwise regression) is a feature selection algorithm that is used to identify important genes for a given phenotype [19]. It uses a novel measure called the density index (DI) to identify gene–gene correlations and to evaluate the inhomogeneity in feature space. It leverages these correlations to perform stepwise regression analysis to select the most informative genes. The algorithm is designed to work with high-dimensional datasets and can be applied to a wide range of biological and medical applications. This tool can be useful in identifying key genes and pathways associated with a specific phenotype or disease and can aid the development of new diagnostic or therapeutic strategies. One limitation of the DUBStepR algorithm is that it is based on gene–gene correlations, which may not always accurately reflect the underlying biology of a phenotype or disease. Additionally, the algorithm is based on stepwise regression, which can be sensitive to the choice of initial genes and may not always converge to the optimal solution.

Random forest is a popular machine learning algorithm that is commonly used for classification and regression tasks. One of the key strengths of random forest is its ability to provide insight into the relative importance of each feature (or variable) in the dataset. Variable importance, as calculated by random forest, can be used as a feature selection method. The idea is to rank the features based on their importance and select the top-ranking features for further analysis or use in a predictive model. This method has several advantages over traditional feature selection methods since random forest can handle non-linear relationships and interactions between features, and it does not require the assumption of independence between features. However, there are also some limitations to using random forest for feature selection. One main limitation is that the feature importance scores can be affected by the correlation between features. It should be used in combination with other methods and domain knowledge to come to a more informed decision.

Furthermore, we applied the statistical strategy using the BPSC package [20]. BPSC (Beta-Poisson model for single-cell RNA-seq data analyses) is an R tool for the analysis of single-cell RNA-seq data. It uses a statistical model called the Beta-Poisson model to analyze the expression of genes in individual cells. The Beta-Poisson model is a flexible model that can handle over-dispersed count data, which are a common characteristic of single-cell RNA-seq data. One of the major advantages of BPSC is that it is able to handle high-dimensional single-cell data and it can be used to identify biologically meaningful gene expression patterns. However, as with any computational method, it also has some limitations. The performance of BPSC depends on the quality of the input data and the specific characteristics of the sample.

2.2. Important Genes Prioritization

The three main FS methods are combined, creating a consensus score regarding the importance of genes under the perspective of predicting process accuracy. Our aim here is to export the optimal list of genes which offer a capable separation among the cell classes in our single-cell RNA-seq experimental dataset. We employed the Borda count, a single-winner election technique in which voters rank candidates or choices according to preference [21]. By awarding each candidate a certain number of points for each ballot depending on the number of lowest-rated contestants, the Borda count determines the outcome of a debate or the winner of an election. The choice or candidate with the most points after all votes have been tallied is declared the winner. A ranking list was constructed for each FS method, and using Borda count, we received a list of ranked genes that include all the values from the variable importance scores from the three main FS methods. As a result, the genes with the highest score are the key players for the better separability of our cell samples regarding their classes.

3. Results and Discussion

We evaluated the classification performance of our scEFS with the three main feature selection methods using expression profiles from a single-cell RNA-seq analysis. All methods were applied to the GSE103334 dataset from Gene Expression Omnibus [22]. This study tracks the microglia activation in neurodegeneration, examining their phenotypic heterogeneity and transcriptional dynamic. The experimental approach utilized a neurodegeneration mouse model with Alzheimer's-like pathology and phenotypes (CK-p25 model), discovering novel microglia cell states and uncovering the underlying transcriptional programs. In our analysis, this case is considered a classification task for the accurate prediction of selected datasets derived from two different timepoints (0 weeks and 2 weeks) to assess microglia activation during the progression of neurodegeneration between CK-p25 animals and controls.

The implementation process involved calculating the *p*-value of each feature using the BSPC package and retaining only the top 200 features with the lowest *p*-values. Additionally, the varImp function from the random forest classifier was used to determine the significance of each variable in predicting the class attribute of the dataset. Only the top 200 most important variables were kept. Similarly, the DUBStepR algorithm was executed by isolating the same number of the most dominant genes.

A Venn diagram was used as a visual approach to organize and compare differentially expressed genes obtained from the various applied feature selection methods. It comprises a useful tool for illustrating the relationship and overlap between different datasets as well as highlighting the percentage of genes selected by each method. As depicted in Figure 1, the results of each technique yielded a set of characteristics that exhibited a maximum of 31% unique representation. There was minimal overlap between the feature selection methods, with a total of 2% overlap observed among all the different methods, and even the overlap between pairs of methods did not exceed 8%. Given this limited overlap, an ensemble method was proposed to identify potential gene markers by utilizing a ranking of all of these distinct gene groups. This limited overlap can be attributed to the dissimilar perspectives of each technique, since each feature selection method utilizes a distinct strategy to find the most dominant genes.

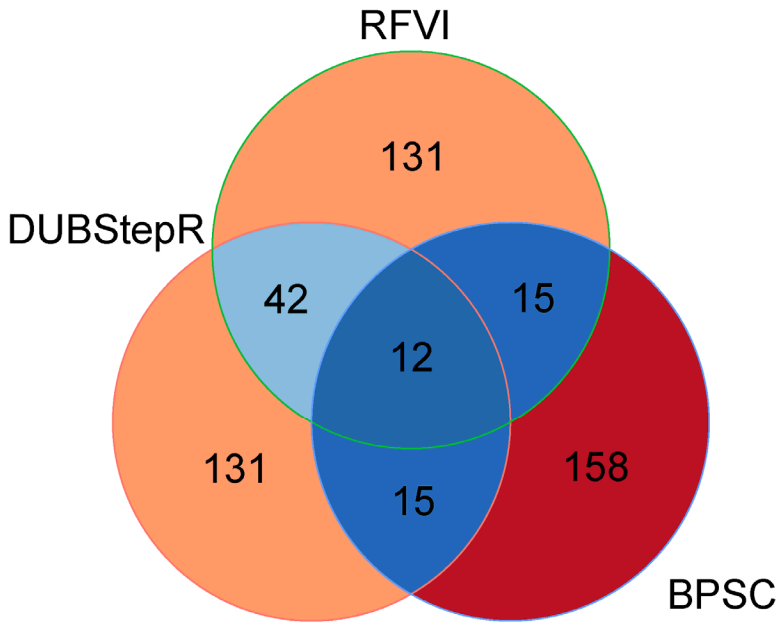

Figure 1. A Venn diagram depicting the three main future selection methods, highlighting their low rate of overlap.

The kNN classifier was used to provide a comparison between the three different techniques and the initial data. The k-nearest neighbor (kNN) method is a well-known classification approach in the fields of data mining and statistics due to the ease with which it can be implemented and the high classification performance it offers [23]. In each case study, we initially trained each classification algorithm by providing it with the set of 200 informative genes selected using each gene selection method. The goal was to assess how well the aforementioned groups can predict the class feature and compare them based on metrics such as accuracy, sensitivity, specificity, and F1 score, as Table 1 shows. In Figure 2, the receiver operating characteristics (ROC) curve is provided and the area under the curve (AUC) is used as a measure of performance between all groups.

Table 1. Prediction accuracy with kNN classifier prediction accuracy based on 200 informative genes by each gene selection method and initial data.

Filter Data	Accuracy	Kappa	Precision	Recall	F1	Elapsed Time (m)
Initial Data	0.77	0.55	0.73	0.86	0.79	4.25
Variable Data	0.69	0.39	0.72	0.63	0.67	1.03
Importance Data	0.81	0.63	0.83	0.78	0.81	1.3
Statistical Data	0.60	0.21	0.72	0.34	0.46	1.01
EFSM	0.82	0.65	0.83	0.81	0.82	1

In order to understand the high-level functions and utilities of the biological system according to the isolated genes, we conducted an enrichment analysis using Enrichr https://maayanlab.cloud/Enrichr/ (accessed on 6 November 2022) [24,25]. GO function and Reactome https://reactome.org/ (accessed on 6 November 2022) [26] pathway enrichment analyses were performed for the twelve DEGs, as Figure 1 shows. Following enrichment analysis, we concluded that AD is proven. The enriched GO terms were divided into biological process (BP), molecular function (MF), and cellular component (CC) ontologies (Tables S1–S4). The results of the GO analysis indicated that DEGs were mainly enriched in BP, including mRNA metabolic process, oxidative RNA demethylation, regulation of core promoter binding, and glandular epithelial cell development (Figure S1A). MF analysis revealed that the DEGs were significantly enriched endoribonuclease activity, producing 3′-phosphomoesters, guanylate kinase activity, cAMP binding, and microtubule plus-end binding (Figure S1B). For CC, the DEGs were enriched in a collage-containing extracellular matrix, a basement membrane and endosome lumen (Figure S1C). The results of the Reactome pathway analysis showed that DEGs were mainly enriched in pathways such as in assembly of collagen fibrils extracellular matrix organization, collagen formation, and cell–cell communication (Figure S1D). More precisely, a *Dst* (dystonin) gene encodes a member of the plakin protein family of adhesion junction plaque proteins acting as an integrator of intermediate filaments and microtubule cytoskeleton networks. A *Dst* regulates the organization and stability of the microtubule network of sensory neurons to allow axonal transport and mediates docking of the dynein/dynactin motor complex to vesicle cargos for retrograde axonal transport through its interaction with TMEM108 and DCTN1, while loss of function can cause hereditary sensory, autonomic neuropathy type 6, and epidermolysis bullosa simplex [27]. Biological process analysis indicated that a *Dst* was significantly enriched in hemidesmosome assembly and retrograde axonal transport (Figure S2A, Table S6). Deficits in the latter are associated with the pathogenesis of multiple neurodegenerative diseases, including amyotrophic lateral sclerosis [28]. Pathway enrichment analysis showed that a *Dst* was mainly enriched in the RHO family of GTPases (Figure S2D, Table S8) which are involved in the regulation of cell migration and cell adhesion and play important roles in neuronal development [29].

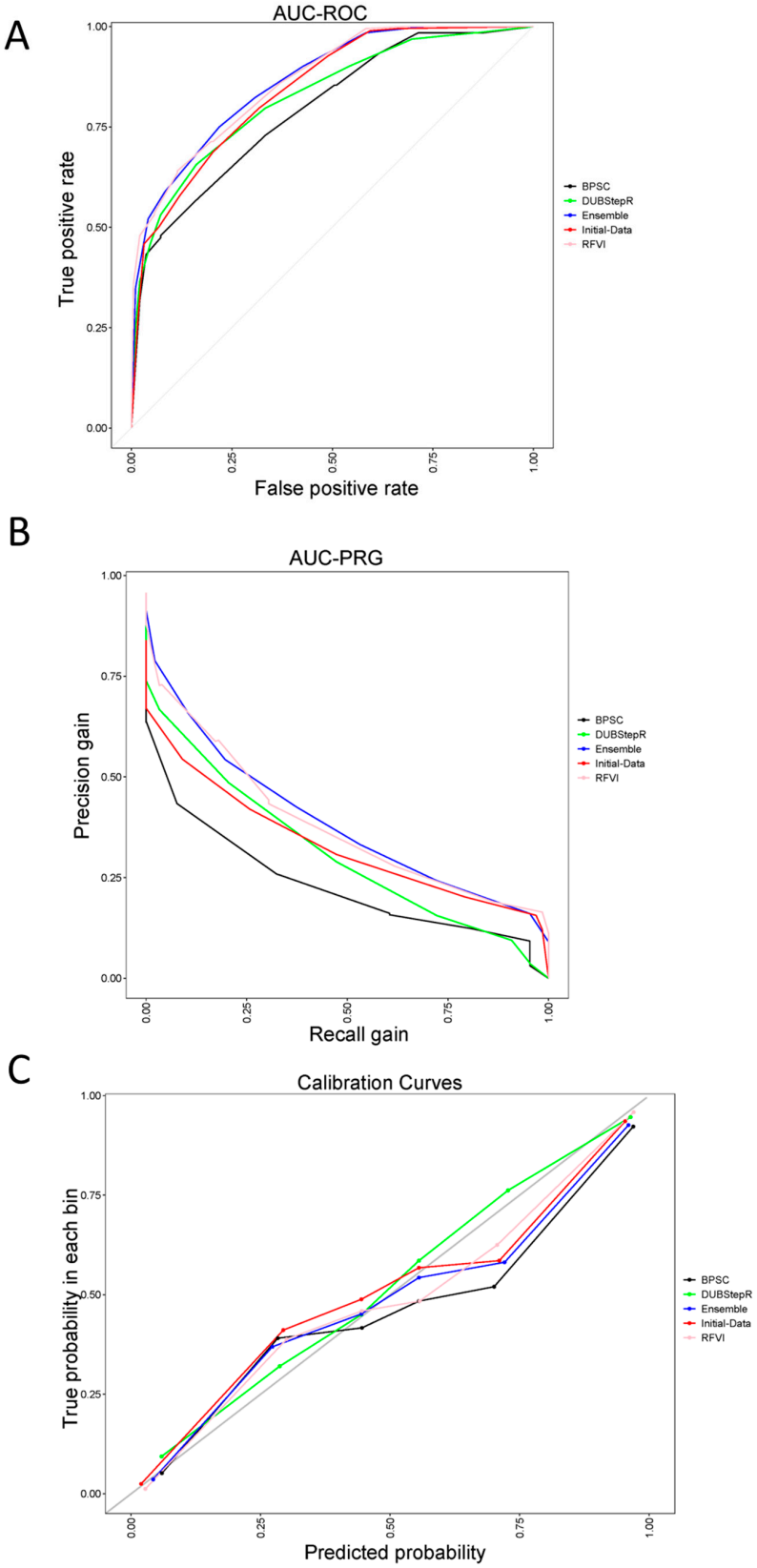

Figure 2. A representation of the receiver operating characteristics (ROC) indicating that the classifier can be organized and its performance better understood. (**A**) AUC-ROC, (**B**) AUC-PRG, (**C**) calibration curves.

A *Rapgef4* (Rap guanine nucleotide exchange factor 4) gene is involved in the regulation of neuronal action potential and the development of the nervous system and plays a role in postsynaptic density and glutamatergic synapse. For BF, *Rapgef4* was enriched in the regulation of peptide hormone secretion, the regulation of insulin and protein secretion, respectively (Figure S3A, Table S9). Peptide hormones are also the hypothalamic peptide hormones such as CRH, GHIGH, and TRH and their function is implicated in the regulation of peptide-containing secretory neurons [30]. The results of MF analysis revealed that *Rapgef4* was mainly enriched cyclic adenosine monophosphate (cAMP) binding, cyclic nucleotide binding, and guanyl-nucleotide exchange factor activity (Figure S3B, Table S10). It should be noticed that cAMP is involved in neuronal functionality and is known to activate and integrate a variety of downstream pathways. cAMP-dependent signaling takes place in neuronal metabolism, growth cone motility, and neuroprotection in the central nervous system [31]. Reactome pathway analysis indicated that *Rapgef4* was significantly enriched in Rap1 and the integrin signaling R-HAS pathway, platelet aggregation, and glycagon-like peptide (Figure S3C, Table S11). Neuronal Rap1 regulates energy balance, glucose homeostasis, and leptin actions [32]. Moreover, the results of GO analysis indicated that *Pax6*, a gene which the encoded highly conserved transcription factor is essential in the formation of tissues and organs during embryonic development, was significantly enriched in BP, including regulation of core promoter binding, glandular epithelial cell development, and neuron fate commitment (Figure 3A). *Pax6* plays a critical role for neural stem cell proliferation and neurogenesis in many regions of the central nervous system, including the cerebral cortex [33]. It also controls neuronal development and targets a large number of promoters in neural progenitor cells [34]. Furthermore, the results of the Reactome pathway analysis showed that *Cask* was mainly enriched in dopamine neurotransmitter release cycle R-HAS and assembly and cell surface presentation of NMDA receptors R-HAS (Figure 3B). The encoded multidomain scaffolding protein is highly expressed in the mammalian nervous system and participates in brain development [35]. It was originally identified as a binding partner of neurexins, transmembrane proteins expressed in neurons, and neuroendocrine cells, while loss of its action affects synaptic function in cortical excitatory neurons [36]. Recent studies indicated that CASK is localized to the nucleus in both mouse neuronal cultures and brain tissues [37].

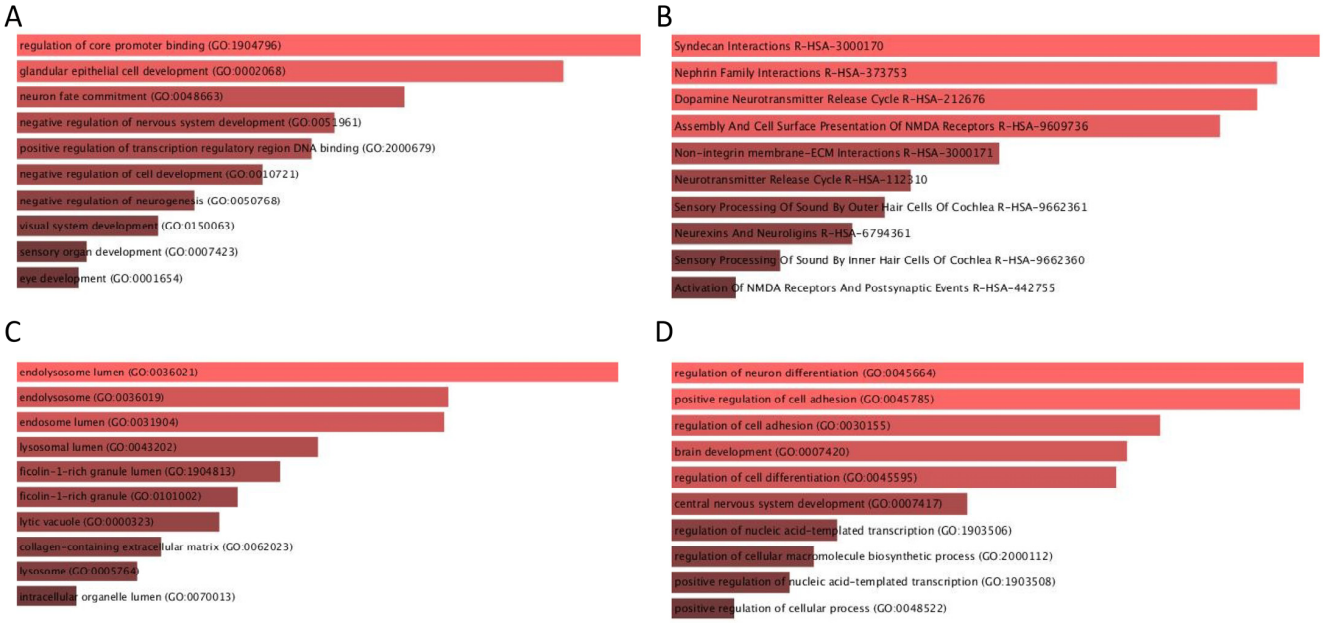

Figure 3. GO term and pathway enrichment analysis performed using Enrichr on DEGs. (**A**) The top 10 enriched biological process for *Pax6*. (**B**) The top 10 enriched Reactome pathway for *Cask*. (**C**) The top 10 enriched cellular component for *Ctsb*. (**D**) The top 10 enriched biological process for *ZFHX3*.

Cathepsin B (*CTSB*), another high-scored gene in our analysis, plays a neuroprotective role in AD and elevated levels have been associated with targeting intervention against the disease. It is considered as a candidate protease for the generation of N-terminally truncated Aβ in astroglial cell cultures [38]. Enhanced hippocampal and cortical amyloid depositions were also observed in the cathepsin B-deficient mouse model of AD overexpressing human amyloid protein precursors (hAPP) [39]. According to our GO analysis, MF showed that *Ctsb* was enriched in cysteine-type peptidase activity (Figure S6B, Table S20). New insights into the role of cysteine protease inhibitor, such as cysteine cathepsins and calpain 1, in neuroinflammation have been recently reported [40,41]. For CC, *Ctsb* was mainly enriched in endolysosome lumen, endolysosome, and endosome lumen (Figure 3C). Alterations in endolysosomal trafficking can induce neurodevelopmental progression [42]. In particular, the accumulation of protease-deficient LAMP1-positive organelles has been observed in axonal distensions near extracellular Aβ plaques in AD modes, indicating defects in lysosome maturation with AD pathogenesis [43]. Lastly, the GO analysis showed that *Zfhx3* in BF was significantly enriched in the regulation of neuronal differentiation, adhesion, and brain development (Figure 3D). *ZFHX3* (zinc finger homeobox 3) encodes a transcription factor with multiple homeodomains and zinc finger motifs which regulates myogenic and neuronal differentiation. The survival of neurons by inducing platelet-derived growth factor receptor β expression is promoted through the signaling pathway involving cAMP-responsive element-binding protein (CREB) and ZFHX3 [44]. The KEGG pathway map (Figure S8) shows significant expressed genes identified in the discrimination of the cell classes [45].

Given a dataset with ultra-high dimensionality we have two main pillars to deal with its complexity: dimensionality reduction and feature selection methods. The first pillar transforms features into a lower dimension trying to keep the pairwise sample distances as efficiently as possible based on the original feature space. The latter pillar includes a simpler method that isolates a list of features that are the most dominant features in the entire dataset. In this study, we focus on the second way since we are also interested in identifying dominant genes in scRNA-seq data that can improve the performance of various supervised learning tasks such as the classification process. An indicative feature selection categorization includes three main strategies: filter-based, wrapper-based, and ensemble-based strategies [46]. More specifically, filter methods select features based on a performance measure independent of the data modeling algorithm used, while they can classify individual features or evaluate entire subsets of features. Indicative measures for selecting the best features through such methods are information, distance, consistency, similarity, and statistical measures. The reason is that it is computationally more feasible and is performed by selecting the variables that have a higher value in a predefined numerical function that estimates the weight of the variable for the classification task. Functions incorporate various criteria such as information gain, mutual information, chi-square, odds ratio, relevance score, and correlation coefficient [47]. Wrappers are feature selection algorithms that assess a subset of features based on the accuracy of a predictive model developed with them. The evaluation is carried out with the assistance of a classifier that provides an estimate of the importance of a certain group of attributes [48]. This class of approaches has proven effective; however, their high computational cost limits their use. In terms of ensemble methods, we refer to a compilation of subsets of features derived from a variety of different base classifiers.

The list of ranked genes obtained by the present approach includes all values from the variable importance scores derived from the three feature selection methods. Using Borda count, the aim of this score is to identify genes that have high scores across all feature selection methods, so as to obtain robust and remarkable gene markers that drive the structure of the given transcriptomics dataset [49]. A gene will rank low, for example, if it presents the best value in one feature selection method and is somewhere in the middle of the ranking list for the other methods. On the other hand, a gene will earn a high Borda score if it ranks third in all scores (even it is not in the top positions) because its function

was significant in all feature selection methods. According to this process, we ensure that we obtain a fairer genes hierarchy and we do not miss any important gene due to the limitations and weaknesses of each individual method. Population heterogeneity, spatial heterogeneity, and temporal heterogeneity are three distinct groups of biologically relevant heterogeneity. Cellular heterogeneity provides insights into the network connectivity and plays an important role in regulating intrinsic cell fate decisions. Data distribution derives the selection of the optimal visualization tool such as methods for classification and pathway modeling [50]. Dimensionality is the typical problem of maintaining several features in a heterogeneous dataset with a number of instances from an analytics perspective, such as in scRNA-seq [51]. The main purpose of this research is to develop an ensemble feature selection approach for transcriptome data using established feature selection techniques. The evaluation of the results is based on both how effectively the classification models respond and how adequately the enrichment analysis works.

4. Conclusions

Rapid advances in next-generation sequencing technologies are providing important insights into complex biological systems. scRNA-seq is a particularly promising transcriptomics technology; however, the high-dimensional nature of its data poses a significant challenge for downstream analysis. In this study, we proposed an ensemble feature selection methodology for scRNA-seq data, specifically in the context of Alzheimer's disease. Our ensemble approach combines various feature selection strategies to obtain the most dominant genes in an AD scRNA-seq dataset, which can be considered as potential regulators in cellular mechanisms. An enrichment analysis was performed on the highest-scoring genes, which showed profound alterations in biological process, molecular function, and cellular component (CC) ontologies related to the disorder. We anticipate that this approach is promising for identifying potential transcriptome biomarkers through AD scRNA-seq data analysis and can be applied to other disease contexts. The encouraging results provided by this study justify the significance and suggest further implementation on datasets with high dimensionality.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app13042353/s1, Figure S1: GO term and Reactome pathway enrichment analyses performed using Enrichr on DEGs; Figure S2: GO term and Reactome pathway enrichment analyses performed using Enrichr on *Dst*; Figure S3: GO term and Reactome pathway enrichment analyses performed using Enrichr on *Rapgef4*; Figure S4: GO term and Reactome pathway enrichment analyses performed using Enrichr on *Pax6*; Figure S5: GO term and Reactome pathway enrichment analyses performed using Enrichr on *Cask*; Figure S6: GO term and Reactome pathway enrichment analyses performed using Enrichr on *Ctsb*; Figure S7: GO term and Reactome pathway enrichment analyses performed using Enrichr on *ZFHX3*; Figure S8: KEGG pathway of Alzheimer's disease; Figure S9: The results of a k-nearest neighbors model; Figure S10: Performance of various methods in terms of accuracy, F1 score, Kappa, Precision, and Recall; Tables S1–S4: The top 10 enriched biological process, molecular function, cellular component, Reactome pathway for DEGs; Tables S5–S8: The top 10 enriched biological process, molecular function, cellular component, Reactome pathway for *Dst*; Tables S9–S11: The top 3 enriched biological process, molecular function, Reactome pathway for *Rapgef4*; Tables S12–S14: The top 10 enriched biological process, molecular function, Reactome pathway for *Pax6*; Tables S15–S18: The top 10 enriched biological process, molecular function, cellular component, Reactome pathway for *Cask*; Tables S19–S22: The top 10 enriched biological process, molecular function, cellular component, Reactome pathway for *Ctsb*; Tables S23–S26: The top 10 enriched biological process, molecular function, cellular component, Reactome pathway for *ZFHX3*.

Author Contributions: Conceptualization, M.G.K. and A.G.V.; methodology, P.P. and A.G.V.; software, P.P. and A.G.V.; validation, M.G.K. and A.G.V.; data curation, M.G.K. and A.G.V. writing—original draft preparation, P.P., M.G.K. and A.G.V.; writing—review and editing, P.V.; supervision, M.G.K. and A.G.V.; funding acquisition, P.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call Regional Excellence (Research Activity in the Ionian University, for the study of protein folding in neurodegenerative diseases) (FOLDIT) MIS 5047144.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Reuter, J.A.; Spacek, D.V.; Snyder, M.P. High-throughput sequencing technologies. *Mol. Cell* **2015**, *58*, 586–597. [CrossRef] [PubMed]
2. Cirillo, D.; Valencia, A. Big data analytics for personalized medicine. *Curr. Opin. Biotechnol.* **2019**, *58*, 161–167. [CrossRef] [PubMed]
3. Heather, J.M.; Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **2016**, *107*, 1–8. [CrossRef]
4. Tang, X.; Huang, Y.; Lei, J.; Luo, H.; Zhu, X. The single-cell sequencing: New developments and medical applications. *Cell Biosci.* **2019**, *9*, 53. [CrossRef]
5. Choi, Y.H.; Kim, J.K. Dissecting cellular heterogeneity using single-cell RNA sequencing. *Mol. Cells* **2019**, *42*, 189.
6. Jovic, D.; Liang, X.; Zeng, H.; Lin, L.; Xu, F.; Luo, Y. Single-cell RNA sequencing technologies and applications: A brief overview. *Clin. Transl. Med.* **2022**, *12*, e694. [CrossRef]
7. Wang, R.; Peng, G.; Tam, P.P.; Jing, N. Integration of computational analysis and spatial transcriptomics in single-cell study. *Genom. Proteom. Bioinform.* **2022**, *in press*. [CrossRef] [PubMed]
8. Dokeroglu, T.; Deniz, A.; Kiziloz, H.E. A comprehensive survey on recent metaheuristics for feature selection. *Neurocomputing* **2022**, *494*, 269–296. [CrossRef]
9. Xue, B.; Zhang, M.; Browne, W.N.; Yao, X. A Survey on Evolutionary Computation Approaches to Feature Selection. *IEEE Trans. Evol. Comput.* **2016**, *20*, 606–626. [CrossRef]
10. Mahendran, N.; Durai Raj Vincent, P.M.; Srinivasan, K.; Chang, C.Y. Machine learning based computational gene selection models: A survey, performance evaluation, open issues, and future research directions. *Front. Genet.* **2020**, *11*, 603808. [CrossRef]
11. Rengasamy, D.; Rothwell, B.C.; Figueredo, G.P. Towards a more reliable interpretation of machine learning outputs for safety-critical systems using feature importance fusion. *Appl. Sci.* **2021**, *11*, 11854. [CrossRef]
12. Chen, C.W.; Tsai, Y.H.; Chang, F.R.; Lin, W.C. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Syst.* **2020**, *37*, e12553. [CrossRef]
13. Aziz, R.; Verma, C.K.; Srivastava, N. Dimension reduction methods for microarray data: A review. *AIMS Bioeng.* **2017**, *4*, 179–197. [CrossRef]
14. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef] [PubMed]
15. Mera-Gaona, M.; López, D.M.; Vargas-Canas, R.; Neumann, U. Framework for the ensemble of feature selection methods. *Appl. Sci.* **2021**, *11*, 8122. [CrossRef]
16. Alhenawi, E.A.; Al-Sayyed, R.; Hudaib, A.; Mirjalili, S. Feature selection methods on gene expression microarray data for cancer classification: A systematic review. *Comput. Biol. Med.* **2022**, *140*, 105051. [CrossRef]
17. Mahendran, N.; Vincent, P.D.R.; Srinivasan, K.; Chang, C.Y. Improving the classification of alzheimer's disease using hybrid gene selection pipeline and deep learning. *Front. Genet.* **2021**, *12*, 784814. [CrossRef] [PubMed]
18. Guo, X.; Jiang, X.; Xu, J.; Quan, X.; Wu, M.; Zhang, H. Ensemble consensus-guided unsupervised feature selection to identify Huntington's disease-associated genes. *Genes* **2018**, *9*, 350. [CrossRef] [PubMed]
19. Ranjan, B.; Sun, W.; Park, J.; Mishra, K.; Schmidt, F.; Xie, R.; Alipour, F.; Singhal, V.; Joanito, I.; Honardoost, M.A.; et al. DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data. *Nat. Commun.* **2021**, *12*, 5849. [CrossRef] [PubMed]
20. Vu, T.N.; Wills, Q.F.; Kalari, K.R.; Niu, N.; Wang, L.; Rantalainen, M.; Pawitan, Y. Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **2016**, *32*, 2128–2135. [CrossRef]
21. Drotár, P.; Gazda, M.; Vokorokos, L. Ensemble feature selection using election methods and ranker clustering. *Inf. Sci.* **2019**, *480*, 365–380. [CrossRef]
22. Mathys, H.; Adaikkan, C.; Gao, F.; Young, J.Z.; Manet, E.; Hemberg, M.; De Jager, P.L.; Ransohoff, R.M.; Regev, A.; Tsai, L.H. Temporal Tracking of Microglia Activation in Neurodegeneration at Single-Cell Resolution. *Cell Rep.* **2017**, *21*, 366–380. [CrossRef] [PubMed]
23. Parry, R.M.; Jones, W.; Stokes, T.H.; Phan, J.H.; Moffitt, R.A.; Fang, H.; Shi, L.; Oberthuer, A.; Fischer, M.; Tong, W.; et al. k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *Pharm. J.* **2010**, *10*, 292–309. [CrossRef]
24. Kuleshov, M.V.; Jones, M.R.; Rouillard, A.D.; Fernandez, N.F.; Duan, Q.; Wang, Z.; Koplev, S.; Jenkins, S.L.; Jagodnik, K.M.; Lachmann, A.; et al. Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **2016**, *44*, W90–W97. [CrossRef]

25. Xie, Z.; Bailey, A.; Kuleshov, M.V.; Clarke, D.J.; Evangelista, J.E.; Jenkins, S.L.; Lachmann, A.; Wojciechowicz, M.L.; Kropiwnicki, E.; Jagodnik, K.M.; et al. Gene set knowledge discovery with Enrichr. *Curr. Protoc.* **2021**, *1*, e90. [CrossRef] [PubMed]
26. Fabregat, A.; Sidiropoulos, K.; Viteri, G.; Forner, O.; Marin-Garcia, P.; Arnau, V.; D'Eustachio, P.; Stein, L.; Hermjakob, H. Reactome pathway analysis: A high-performance in-memory approach. *BMC Bioinform.* **2017**, *18*, 142. [CrossRef]
27. Motley, W.W.; Züchner, S.; Scherer, S.S. Isoform-specific loss of dystonin causes hereditary motor and sensory neuropathy. *Neurol. Genet.* **2020**, *6*, e496. [CrossRef] [PubMed]
28. Ström, A.L.; Gal, J.; Shi, P.; Kasarskis, E.J.; Hayward, L.J.; Zhu, H. Retrograde axonal transport and motor neuron disease. *J. Neurochem.* **2008**, *106*, 495–505. [CrossRef]
29. Stankiewicz, T.R.; Linseman, D.A. Rho family GTPases: Key players in neuronal development, neuronal survival, and neurodegeneration. *Front. Cell. Neurosci.* **2014**, *8*, 314. [CrossRef]
30. Sadow, T.F.; Rubin, R.T. Effects of hypothalamic peptides on the aging brain. *Psychoneuroendocrinology* **1992**, *17*, 293–314. [CrossRef]
31. Boczek, T.; Kapiloff, M.S. Compartmentalization of local cAMP signaling in neuronal growth and survival. *Neural Regen. Res.* **2020**, *15*, 453. [PubMed]
32. Kaneko, K.; Xu, P.; Cordonier, E.L.; Chen, S.S.; Ng, A.; Xu, Y.; Morozov, A.; Fukuda, M. Neuronal Rap1 regulates energy balance, glucose homeostasis, and leptin actions. *Cell Rep.* **2016**, *16*, 3003–3015. [CrossRef]
33. Sansom, S.N.; Griffiths, D.S.; Faedo, A.; Kleinjan, D.J.; Ruan, Y.; Smith, J.; Van Heyningen, V.; Rubenstein, J.L.; Livesey, F.J. The level of the transcription factor Pax6 is essential for controlling the balance between neural stem cell self-renewal and neurogenesis. *PLoS Genet.* **2009**, *5*, e1000511. [CrossRef]
34. Thakurela, S.; Tiwari, N.; Schick, S.; Garding, A.; Ivanek, R.; Berninger, B.; Tiwari, V.K. Mapping gene regulatory circuitry of Pax6 during neurogenesis. *Cell Discov.* **2016**, *2*, 15045. [CrossRef]
35. Hsueh, Y.P. The role of the MAGUK protein CASK in neural development and synaptic function. *Curr. Med. Chem.* **2006**, *13*, 1915–1927. [CrossRef]
36. McSweeney, D.; Gabriel, R.; Jin, K.; Pang, Z.P.; Aronow, B.; Pak, C. CASK loss of function differentially regulates neuronal maturation and synaptic function in human induced cortical excitatory neurons. *Iscience* **2022**, *25*, 105187. [CrossRef] [PubMed]
37. Meng, Y.; Montilla-Perez, P.; Hillary, R.; Wen, J.; Benner, C.; Telese, F. The Function of CASK in Transcriptional Regulation in Neurons. *FASEB J.* **2020**, *34*, 1. [CrossRef]
38. Oberstein, T.J.; Utz, J.; Spitzer, P.; Klafki, H.W.; Wiltfang, J.; Lewczuk, P.; Kornhuber, J.; Maler, J.M. The role of Cathepsin B in the degradation of Aβ and in the production of Aβ peptides starting with Ala2 in cultured astrocytes. *Front. Mol. Neurosci.* **2021**, *13*, 615740. [CrossRef]
39. Hook, V.Y.; Kindy, M.; Reinheckel, T.; Peters, C.; Hook, G. Genetic cathepsin B deficiency reduces β-amyloid in transgenic mice expressing human wild-type amyloid precursor protein. *Biochem. Biophys. Res. Commun.* **2009**, *386*, 284–288. [CrossRef]
40. Pišlar, A.; Bolčina, L.; Kos, J. New insights into the role of cysteine cathepsins in neuroinflammation. *Biomolecules* **2021**, *11*, 1796. [CrossRef]
41. Siklos, M.; BenAissa, M.; Thatcher, G.R. Cysteine proteases as therapeutic targets: Does selectivity matter? A systematic review of calpain and cathepsin inhibitors. *Acta Pharm. Sin. B* **2015**, *5*, 506–519. [CrossRef] [PubMed]
42. Kulkarni, V.V.; Maday, S. Neuronal endosomes to lysosomes: A journey to the soma. *J. Cell Biol.* **2018**, *217*, 2977. [CrossRef] [PubMed]
43. Gowrishankar, S.; Yuan, P.; Wu, Y.; Schrag, M.; Paradise, S.; Grutzendler, J.; De Camilli, P.; Ferguson, S.M. Massive accumulation of luminal protease-deficient axonal lysosomes at Alzheimer's disease amyloid plaques. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E3699–E3708. [CrossRef] [PubMed]
44. Kim, T.S.; Kawaguchi, M.; Suzuki, M.; Jung, C.G.; Asai, K.; Shibamoto, Y.; Lavin, M.F.; Khanna, K.K.; Miura, Y. The ZFHX3 (ATBF1) transcription factor induces PDGFRB, which activates ATM in the cytoplasm to protect cerebellar neurons from oxidative stress. *Dis. Model. Mech.* **2010**, *3*, 752–762. [CrossRef]
45. Kanehisa, M.; Furumichi, M.; Sato, Y.; Kawashima, M.; Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **2023**, *51*, D587–D592. [CrossRef]
46. Santana, L.E.A.D.S.; de Paula Canuto, A.M. Filter-based optimization techniques for selection of feature subsets in ensemble systems. *Expert Syst. Appl.* **2014**, *41*, 1622–1631. [CrossRef]
47. Tadist, K.; Najah, S.; Nikolov, N.S.; Mrabti, F.; Zahi, A. Feature selection methods and genomic big data: A systematic review. *J. Big Data* **2019**, *6*, 79. [CrossRef]
48. Uncu, Ö.; Türkşen, I.B. A novel feature selection approach: Combining feature wrappers and filters. *Inf. Sci.* **2007**, *177*, 449–466. [CrossRef]
49. Sarkar, C.; Cooley, S.; Srivastava, J. Robust feature selection technique using rank aggregation. *Appl. Artif. Intell.* **2014**, *28*, 243–257. [CrossRef]
50. Gough, A.; Stern, A.; Maier, J.; Lezon, T.; Shun, T.-Y.; Chennubhotla, C.; Schurdak, M.; Haney, S.; Taylor, D. Biologically Relevant Heterogeneity: Metrics and Practical Insights. *SLAS Discov.* **2017**, *22*, 213–237. [CrossRef]
51. Xiang, R.; Wang, W.; Yang, L.; Wang, S.; Xu, C.; Chen, X. A comparison for dimensionality reduction methods of single-cell RNA-seq data. *Front. Genet.* **2021**, *12*, 646936. [CrossRef] [PubMed]