Ensemble-based Gene Selection and an Enhanced Deep Multi-Layer Perceptron-based Classification Model for Classifying Alzheimer's disease

Nivedhitha Mahendran
Vellore Institute of Techology, Vellore
Durai Raj Vincent PM (pmvincent@vit.ac.in)
Vellore Institute of Techology, Vellore

Article

Keywords: Gene selection, Ensemble-based feature selection, Artificial intelligence, Deep learning, Microarray data, Gene expression, Alzheimer's disease

Posted Date: May 20th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1628234/v1

Abstract

Memory and cognitive disabilities have been known to humankind for a long time. However, the pathology and symptoms associated with Alzheimer's Disease (AD) are documented during the current century's first decade. Over the years, AD became the common form of Dementia that has complex pathology and is termed a heterogeneous disorder. The advancements in microarray technology made capturing hundreds and thousands of gene sequences possible. However, the data generated is complex and beyond the understanding of the human brain. AD symptoms are slow but fatal, and the diagnosis is made only through autopsy. Thus, early and accurate diagnosis of the disease is critical. The significant difficulty in handling the gene expression data is the curse of dimensionality or the High Dimensional Low Sample Size (HDLSS) issue. The HDLSS issue demands interdisciplinary research, such as Artificial intelligence, machine learning, etc. This study proposed an ensemble-based feature selection to isolate the necessary genes responsible for causing AD. After selecting the relevant features, the Deep Multi-layer Perceptron (DMLP) is used to classify the AD and non-AD patients. The results are compared with other state-of-the-art feature selection techniques and classification algorithms.

1. Introduction

AD is the most dreaded and heterogeneous form of Dementia [1][2]. It is widely found during mid-to-late in life, affecting cognitive ability. AD is regarded as a slow and progressive disorder. The significant symptoms of AD are lack of judgment, irregularities in memory, difficulty in speaking, and abnormalities in visuospatial perceptions and behavior [3][4]. When AD progresses, individuals will experience difficulty carrying out their daily living. They will usually go mute, bedridden, and challenging to handle [5]. AD gradually provides fatal results from prolonged medical illness. In rare cases, it will cause delusions and hallucinations [6]. It often begins with a vague memory loss, which slowly interrupts the quality of life, destroying memory and ability to think [1].

AD selectively damages the neurons in specific regions in the neural systems and the brain. The damage includes the nerve cells in the hippocampus, cortex, anterior thalamus, amygdala, and basal forebrain. The damaged nerve cells are found to have an accumulation of amyloid plaques (Aβ), and neurofibrillary tangles called the Tau tangle [7]. The amyloid plaques interrupt the communication between the neurons. They affect the cells, cause their death, and obstruct intercellular transport. The plagues and Tau tangles are formed in the hippocampus early and later spread to other brain regions [7], [8]. The risk factors causing AD include lifestyle choices (exercise, diet, smoking, etc.), genetics, biology (gender and age), and accidents (head trauma) [9].

AD is considered one of the orphan diseases, diseases with no cure [10]. The diagnosis is made only after the autopsy. Brain imaging, neurocognitive tests, and cerebrospinal fluid (CSF) analysis are the diagnostic approaches widely used [11][12][13]. AD is one of the forms of Dementia. Dementia is anticipated in 50 million people worldwide, and among them, about 67% have AD [9]. Unfortunately, there is no cure, and there seem to be only four drugs approved for slowing down the symptoms. Treatments can only slow down progress and do not eliminate the symptoms. A better understanding of the clinical pathologies, mechanism of the disease, and genetic risk factors may help treat AD effectively [14].

As discussed above, AD is a progressive neurological disorder that primarily affects memory and learning. AD being an orphan disease combined with a high dimensional complex data and uncertainty demands advanced approaches, such as Artificial Intelligence (AI) [15]. AD aids in revealing patterns in vast complex gene expression data, which leads to the discovery of disease-related genes. The difficulty with AD is that more neurons have died by the time it is diagnosed, making it irreversible. The etiology of AD is uncertain, but it is estimated that 70% of the

cases are associated with genetic factors [16]. Currently, for AD, there is no reversible or prevention treatment. Also, it is challenging to build a simplified model because of complex interactions among various factors and the complexity of humans. Fortunately, the recent and rapid developments in AI offered solutions to these problems, which involve ultra-complex massive data [17]. AI carries out an integrative approach and models the functional neurobiological components that influence neuropsychiatric disorders. This study implemented an AI-based strategy on two AD microarray datasets. We analyzed the data using a standard preprocessing routine and normalized it to make it comparable across all the platforms. Then, we performed the ensemble-based feature selection to choose the relevant genes that possibly cause AD. Once the required genes are selected, AD and non-AD classification is carried out using the Deep learning approach.

The remaining section of the paper is split into materials and methods, which discuss the background and the approaches used, Results and Discussion, discuss the results of the applied techniques and their performance evaluation, and then the conclusion of the work done.

2. Materials And Methods

This section discuss the background about the dataset, DNA Microarray, preprocessing feature selection and classification of AD. The Fig. 1 represents the overall process flow of the implemented framework.

2.1. Dataset

We applied the proposed approaches individually on two gene expression datasets (GSE33000 and GSE44770) downloaded from the GEO omnibus database. Both the datasets are extracted from the Prefrontal Cortex of the brain. In GSE33000, there are 310 cases and 157 controls with a sample size of 467. The GSE44770 dataset has 128 cases and 102 controls with a sample size of 230. The number of features in both datasets is 39,280. The datasets are extracted using a custom-made Agilent 44k array from the autopsied tissues of the brain. For most of the preprocessing routine, we used the GEO2R tool, and for normalization, feature selection, and classification, we used RStudio packages.

Table 1
Dataset Information

Dataset	Source	Dataset Size	No. of features
GSE33000 [18]	Geo Omni bus	467 (310 cases and 157 controls)	39,280
GSE44770 [19]	Geo Omni bus	230 (128 cases and 102 controls)	39,280

2.2. DNA Microarray

DNA microarrays are microscopic slide that holds a massive amount of thousands of gene sequences; robotic machines capture them [20]. More than 40,000 gene sequences are available for research in the database [21]. The process starts when a gene is activated and a few DNA segments are copied. The copied segments are called the mRNA, the template for building up the proteins. To identify the genes' turned on and turned off, the researchers must gather the mRNA molecules. These microarray chips aid the researchers in analyzing the gene expressions.

2.3. Preprocessing

Preprocessing is critical to any data analysis to transform the unintelligible raw data into a functional and understandable format. The massive and gene expression datasets tend to have missing values, skewed distributions, background noise, and non-biological variations [22]. This study used AD's two microarray gene expression datasets and performed the standard preprocessing routine to make the data ready for further processing.

The preprocessing routine includes handling missing values, quality control, normalization, and identifying the differentially expressed genes.

- Handling missing values: Missing values in the gene expression dataset will reduce the model's classification accuracy [23]. We used the kNN imputation method to impute the missing values in the datasets.
- Quality control: The data is put through log transformation, and the adjusted p-values are estimated. Quality control is done to identify the uncertain or poorly performing samples [22].
- Normalization: The irregularities in the gene expression levels make the data distribution often skewed. Normalizing the data will help remove the noise, technical variations, and other non-biological differences [24]. We used Interquartile Range (IQR) normalization. IQR is effective in identifying the outliers than the commonly used Z-Score normalization. IQR is given by,

$$g_{mn} = \frac{g_{mn} - Median_n}{GIQR_j / 1.35}$$

Where g_{mn} is the expression value of the gene n from the sample m, $Median_n$ is the median of the gene n over all the available samples. $GIQR_n$ is the IQR of the specific gene n.

- Differentially Expressed Genes: The common goal in any microarray is to identify the differentially expressed genes (DEGs). The DEGs are selected through two steps. The first step is to identify the p-values using statistical methods, such as t-test, M-statistics, etc. Then, the genes are ranked based on their respective p-values. The second step is finding out the significance level of the genes with the help of the p-value cutoff threshold [25]. The cutoff threshold we used in this study is 0.05 and selected the top 200 genes.

2.4. Feature Selection

The DNA Microarray technology assists the researchers in analyzing the vast amount of gene expression profiles in parallel belonging to different medical fields. However, the dimension is critical in handling the gene expression data [26]. The dimensionality problem makes the work tedious for the classification algorithms in studying the characteristics of the data. The gene expression profiles represent the molecular level of the cells. Compared to the genes involved, the number of samples is relatively small. To overcome the issue mentioned above, feature selection techniques are applied to extract useful information from the dataset. The feature selection approaches select only those genes that influence the implemented classification models' accuracy by eliminating the irrelevant genes [27].

Feature selection is an active and critical research area in data mining, pattern recognition, and statistics. The improper or irrelevant genes in the dataset significantly affect the classification model's performance. The main idea behind the feature selection approach is to isolate the best subset from the input features and ignore the features with no or little contribution to the final prediction. Feature selection aids in achieving minimized feature

space, which improves the accuracy of the classification model. There are four feature selection types: filter, wrapper, embedded, and ensemble [28].

- Filter: The filter methods are majorly used for filtering or data preprocessing. The features are ranked based on their discriminative power or relevance to the target class. They select features based on the general characteristics of the dataset [21]. They are statistical approaches that are not dependent on the classification models. Example: Information theory and entropy.
- Wrapper: The learning model and feature selection are wrapped together in wrapper methods. The features are evaluated with the help of the accuracy of the learning model. It is computationally extensive, depending on the learning model [21]. Example: SVM-RFE.
- Embedded: It is performed along with training the learning models embedded within them [29]. They are specific to the implemented learning model. Example: LASSO.
- Ensemble: The ensemble-based feature selection approaches combine the output of different feature selectors through aggregation [30]. The result usually is ranked features or a subset of the features. Example: Random forest, ADA boost.

2.5. Ensemble Feature Selection:

The ensemble-based feature selection brings together a group of different learning algorithms and combines their output using appropriate combination techniques [31]. There are two steps involved in building an ensemble [30]. In the first step, the feature selection algorithms are chosen, and in the second step, the outputs of the selected feature selectors are aggregated and returned as a single decision. There are two types of ensemble feature selection approaches. They are homogeneous and heterogeneous [30]. In the homogeneous approach, the feature selection algorithm remains the same, whereas the data subsets are different, for example, Bagging (Random Forest). In heterogeneous, the training data is the same with different feature selection algorithms, such as boosting (Adaboost). The output of an ensemble feature selection is obtained in two ways, a subset of features and ordered ranking. A threshold is needed to reduce the feature space for an ordered ranking [32].

In recent years, apart from the accuracy, the stability of the model is also given equal importance. Stability helps measure the sensitivity of the selected features when there are variations in the dataset [31]. The primary purpose of ensemble-based feature selection is to improve the stability of the selected features. The focus on stability eventually enhances the classification accuracy and generalization of the model.

2.6. Deep Learning and Multi-layer Perceptron

Deep learning is one of the sub-areas of machine learning inspired by Artificial Neural Networks with two or more hidden layers [33]. The Deep Neural Network (DNN) aids in approximating a function, say, 'f,' where the approximated function, 'f,' is used to predict or form new representations. The ANN's are modeled from the biological neural networks. It is a group of neurons with weighted connections [34]. Each neuron tries to convert its input to output in the network using an appropriate activation function. The two widely implemented neural networks are feedforward and feed backward. In this study, we used Deep MLP (DMLP), which works under forwarding feed networks.

Basically, in the feedforward network, the inputs are propagated in a forward pass, processing through the hidden layers. The output of each layer is fed as the input for the next hidden layer and so on until it produces an output.

Finally, gradient descent optimization is applied to update the weights to minimize the loss function. The loss function is the error difference between the actual and expected output [35].

Multi-layer Perceptron (MLP) is the conventional and simplest form of neural network. It consists of three layers: input, hidden, and output [34]. MLP is converted into DMLP by adding more hidden layers, decreasing the number of nodes per hidden layer. The shallow MLPs are designed with a backpropagation algorithm using gradient descent with random allocation of weights. The DMLP is implemented with Rectified Linear Unit (ReLU) as an activation function, which allows the network to learn faster [36].

3. Results And Discussion

In this study, we used two datasets (GSE33000 and GSE44770), one for training the model and testing. As the sample size is less than the feature space, we have used two datasets to compensate for the issue. Both have the same number of features and are extracted from the brain's prefrontal cortex. Both the datasets are preprocessed using the same routine. We performed the kNN imputation to handle all the missing values in the data. After imputing the missing values, the datasets are subjected to log transformation. Figure 2 and Fig. 3 show the volcano plot of both datasets. The volcano plot shows the biological and statistical significance over many genes. It is a plot between the p-value and the fold change. The p-value threshold used is 0.05, and the fold change is greater than 2. The red points in the volcano plot represent the up-regulated genes, and the blue points represent the down-regulated genes. The adjusted p-value of both the datasets is shown in Figs. 4 and 5. The datasets are normalized using IQR normalization. Figure 6 and Fig. 7 show the density distribution of the samples after performing log transformation and normalization. After preprocessing, the top 250 genes are identified as the most significant genes causing AD.

The most significant genes are further processed using ensemble-based feature selection. We used the function variation method, where the training data is the same for different feature selectors. Filters are commonly used in function variation methods, as wrappers are computationally complex. In this study, we used five filter techniques. They are signal-to-noise ratio, ReliefF, chi-squared, mutual information, and information gain. Each feature selector outputs a ranked list of features aggregated into a single ranked list of features. This is achieved with the help of mean aggregation, which finds the mean of every feature's rank across all the feature selectors and considers that as the final rank. The ranked features are used for the classification AD. The classification model we implemented is the Deep MLP, enhancing stopping criteria for fast convergence and avoiding overfitting.

The top 100 features from the ranked list are used as portions in the classification model. The features are continuously added until there is no improvement in the model's accuracy. Thus, 24 final features are selected and used for the final classification of AD. The parameters of the DMLP are tuned using the Grid search algorithm. The parameters are run through five-fold cross-validation, and the average values of each parameter are used for the final classification. Table 2 shows the best performance percentage of the implemented approaches from the five-fold cross-validation. The parameters we chose for DMLP are the number of epochs, activation function (ReLU), dropout regularization (0.7), number of neurons in the hidden layer, and number of hidden layers (3 hidden layers). We added a condition to stop the training for the overfitting issue once the convergence is reached. The test and training accuracy after 15 epochs are compared. When the convergence is reached, the training and testing are stopped. From Table 1 and Fig. 8, we can see that the proposed ensemble approach and the DMLP offer better results. The shallow MLP shows better performance than the other classification models. Also, it provides improved

outcomes when implemented with the proposed feature selection approach. When depth is introduced in the shallow MLP, the results are enhanced, as seen from the table.

Table 2
Performance Evaluation

Feature Selection	Classification Approach	Accuracy	Sensitivity	Specificity	F-score	AU-ROC
mRmR	LDA	0.715	0.721	0.718	0.719	0.725
	SVM	0.747	0.731	0.744	0.735	0.730
	MLP	0.851	0.847	0.846	0.859	0.84
	DMLP	0.912	0.917	0.907	0.904	0.919
SVM-RFE	LDA	0.768	0.761	0.751	0.769	0.75
	SVM	0.794	0.79	0.782	0.786	0.797
	MLP	0.887	0.883	0.876	0.871	0.88
	DMLP	0.921	0.917	0.919	0.928	0.923
LASSO	LDA	0.807	0.817	0.809	0.811	0.81
	SVM	0.827	0.814	0.819	0.824	0.815
	MLP	0.871	0.875	0.88	0.877	0.881
	DMLP	0.937	0.931	0.925	0.933	0.92
Proposed Ensemble Approach	LDA	0.854	0.864	0.869	0.841	0.866
	SVM	0.891	0.897	0.88	0.897	0.892
	MLP	0.958	0.959	0.947	0.949	0.95
	DMLP	0.972	0.979	0.970	0.981	0.987

4. Conclusion

There are many ailments that people suffer over the years due to aging. AD is such slow and exhausting for both patients and their family members. The symptoms of AD have lingered over a long time. The onset of AD damages the brain tissues to an irrecoverable stage if diagnosed late. It severely affects the individual's daily life, causing permanent damage to the cognitive abilities. Modern data technologies, such as microarray, aids in capturing thousands of gene expressions. It assists the researchers exceptionally in analyzing and finding the cure for incurable orphan diseases, such as AD. However, the data generated is incomprehensible. For example, the dataset we used in this study has 467 sample sizes, but the feature size is 39,280. The uneven sample to feature ratio creates the HDLSS problem. To overcome the HDLSS issue, machine learning techniques are proposed. The feature selection techniques reduce dimensionality by isolating only the relevant features directly affecting the target class. We proposed an ensemble-based feature selection approach and an enhanced DMLP for classifying the AD and non-AD patients. The datasets are cleaned and transformed before applying the feature selection, and classification approaches. We applied log transformation and normalized the data using IQR normalization. Then, the most

significant differentially expressed genes are identified using the adjusted p-values. The threshold we used for the p-value is 0.05. The most important, top 250 genes are selected and processed further.

The feature selection approaches, such as the filter, wrapper, and embedded techniques, focus majorly on the diversity of the model. The ensemble-based approaches concentrate on improving the stability and generalization of the model. The proposed approach implements five filter-based feature selection models and combines the results of each feature selector using an aggregation technique. The ranked features are then used for classification. We implemented the DMLP to classify the AD and non-AD patients. The DMLP is enhanced with stopping criteria from avoiding overfitting and minimizing the computational complexity. The ensemble-based feature selection is compared with mRmR (filter), SVM-RFE (wrapper), and LASSO (embedded) approaches with four different classifiers (LDA, SVM, MLP, and DMLP). We used two datasets, one for training and testing, and classification. The results are tabulated and show that the proposed ensemble approach and enhanced DMLP perform better than the implemented approaches. We used Accuracy, Sensitivity, Specificity, F-score, and ROC measures to validate the models. The ensemble approach we used is function variation, and there is one other approach called the data variation. In data variation, the feature selector is the same, and the dataset used is replaced. There is scope for researchers on data variation-based ensemble approaches on microarray datasets to classify AD in the future.

Declarations

Acknowledgement

Not applicable

Author's Contributions

This research specifies below individual contributions. "Conceptualization – D.R.V and N.M; Data curation – D.R.V; Formal analysis – N.M; Methodology – N.M; Project administration – D.R.V; Resources - D.R.V; Software – N.M; Supervision – D.R.V; Validation – D.R.V and N.M; Visualisation – N.M; Writing, review and editing – D.R.V and N.M.

Competing Interest

The author(s) declare no competing interests.

Data Availability

The datasets generated during and/or analysed during the current study are available in the Geo Omnibus repository, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33000 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44770

References

1. C. A. Lane, J. Hardy, and J. M. Schott, "Alzheimer's disease," *Eur. J. Neurol.*, vol. 25, no. 1, pp. 59–70, Jan. 2018, doi: 10.1111/ENE.13439.
2. T. Wang, R. G. Qiu, and M. Yu, "Predictive Modeling of the Progression of Alzheimer's Disease with Recurrent Neural Networks," *Sci. Reports* 2018 *81*, vol. 8, no. 1, pp. 1–12, Jun. 2018, doi: 10.1038/s41598-018-27337-w.

3. S. S. Sisodia, “Series Introduction: Alzheimer’s disease: perspectives for the new millennium,” J. Clin. Invest., vol. 104, no. 9, p. 1169, 1999, doi: 10.1172/JCI8508.
4. A. Ng *et al.*, “IL-1β, IL-6, TNF- α and CRP in Elderly Patients with Depression or Alzheimer’s disease: Systematic Review and Meta-Analysis,” *Sci. Reports* 2018 *81*, vol. 8, no. 1, pp. 1–12, Aug. 2018, doi: 10.1038/s41598-018-30487-6.
5. Z. S. Khachaturian, “Diagnosis of Alzheimer’s Disease,” *Arch. Neurol.*, vol. 42, no. 11, pp. 1097–1105, Nov. 1985, doi: 10.1001/ARCHNEUR.1985.04060100083029.
6. L. Mucke, “Alzheimer’s disease,” *Nat.* 2009 *4617266*, vol. 461, no. 7266, pp. 895–897, Oct. 2009, doi: 10.1038/461895a.
7. T. C. Gamblin *et al.*, “Caspase cleavage of tau: Linking amyloid and neurofibrillary tangles in Alzheimer’s disease,” *Proc. Natl. Acad. Sci.*, vol. 100, no. 17, pp. 10032–10037, Aug. 2003, doi: 10.1073/PNAS.1630428100.
8. J. A. Hardy and G. A. Higgins, “Alzheimer’s disease: the amyloid cascade hypothesis,” *Science (80-.).*, vol. 256, no. 5054, pp. 184–186, Apr. 1992, Accessed: Dec. 27, 2021. [Online]. Available: https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=00368075&v=2.1&it=r&id=GALE%7CA12207965&sid=googleScholar&linkaccess=fulltext.
9. J. A. Potashkin, V. Bottero, J. A. Santiago, and J. P. Quinn, “Computational identification of key genes that may regulate gene expression reprogramming in Alzheimer’s patients,” *PLoS One*, vol. 14, no. 9, Sep. 2019, doi: 10.1371/JOURNAL.PONE.0222921.
10. J. Pantel, “[Alzheimer’s disease from Auguste Deter to the present: Progress, disappointments and open questions].,” Z. Gerontol. Geriatr., vol. 50, no. 7, pp. 576–587, Sep. 2017, doi: 10.1007/S00391-017-1307-2.
11. “Gene expression signatures of Alzheimer’s disease | National Institute on Aging.” https://www.nia.nih.gov/news/gene-expression-signatures-alzheimers-disease (accessed Dec. 27, 2021).
12. J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, “Multimodal deep learning models for early detection of Alzheimer’s disease stage,” Sci. Reports 2021 *111*, vol. 11, no. 1, pp. 1–13, Feb. 2021, doi: 10.1038/s41598-020-74399-w.
13. F. Ciccocioppo *et al.*, “The Characterization of Regulatory T-Cell Profiles in Alzheimer’s Disease and Multiple Sclerosis,” *Sci. Reports* 2019 *91*, vol. 9, no. 1, pp. 1–9, Jun. 2019, doi: 10.1038/s41598-019-45433-3.
14. M. Goedert and M. G. Spillantini, “A Century of Alzheimer’s Disease,” *Science (80-.).*, vol. 314, no. 5800, pp. 777–781, Nov. 2006, doi: 10.1126/SCIENCE.1132814.
15. Ó. Álvarez-Machancoses, E. J. D. Galiana, A. Cernea, J. F. de la Viña, and J. L. Fernández-Martínez, “On the Role of Artificial Intelligence in Genomics to Enhance Precision Medicine,” Pharmgenomics. Pers. Med., vol. 13, p. 105, 2020, doi: 10.2147/PGPM.S205082.
16. W. B. Grant, A. Campbell, R. F. Itzhaki, and J. Savory, “The significance of environmental factors in the etiology of Alzheimer’s disease,” *J. Alzheimer’s Dis.*, vol. 4, no. 3, pp. 179–189, Jan. 2002, doi: 10.3233/JAD-2002-4308.
17. A. Becker, “Artificial intelligence in medicine: What is it doing for us today?,” *Heal. Policy Technol.*, vol. 8, no. 2, pp. 198–205, Jun. 2019, doi: 10.1016/J.HLPT.2019.03.004.
18. M. Narayanan *et al.*, “Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases,” Mol. Syst. Biol., vol. 10, no. 7, p. 743, Jul. 2014, doi: 10.15252/MSB.20145304.
19. B. Zhang *et al.*, “Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer’s disease,” Cell, vol. 153, no. 3, pp. 707–720, Apr. 2013, doi: 10.1016/J.CELL.2013.03.030.
20. M. J. Heller, “DNA Microarray Technology: Devices, Systems, and Applications,” *http://dx.doi.org/*10.1146/annurev.bioeng.4.020702.*153438*, vol. 4, pp. 129–153, Nov. 2003, doi:

10.1146/ANNUREV.BIOENG.4.020702.153438.

21. R. K. Singh and M. Sivabalakrishnan, “Feature Selection of Gene Expression Data for Cancer Classification: A Review,” Procedia Comput. Sci., vol. 50, pp. 52–57, Jan. 2015, doi: 10.1016/J.PROCS.2015.04.060.
22. C. S. Wilhelm-Benartzi *et al.*, “Review of processing and analysis methods for DNA methylation array data,” *Br. J. Cancer* 2013 *1096*, vol. 109, no. 6, pp. 1394–1402, Aug. 2013, doi: 10.1038/bjc.2013.496.
23. P. Meesad and K. Hengpraprohm, “Combination of KNN-based feature selection and KNN-based missing-value imputation of microarray data,” *3rd Int. Conf. Innov. Comput. Inf. Control. ICICIC’08*, 2008, doi: 10.1109/ICICIC.2008.635.
24. T. N and T. J, “Complete pipeline for Infinium(®) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation,” Epigenomics, vol. 4, no. 3, pp. 325–341, Jun. 2012, doi: 10.2217/EPI.12.21.
25. S. W. Jones *et al.*, “The identification of differentially expressed microRNA in osteoarthritic tissue that modulate the production of TNF-α and MMP13,” Osteoarthr. Cartil., vol. 17, no. 4, pp. 464–472, Apr. 2009, doi: 10.1016/J.JOCA.2008.09.012.
26. Z. Li, W. Xie, and T. Liu, “Efficient feature selection and classification for microarray data,” *PLoS One*, vol. 13, no. 8, p. e0202167, Aug. 2018, doi: 10.1371/JOURNAL.PONE.0202167.
27. C. Kang, Y. Huo, L. Xin, B. Tian, and B. Yu, “Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine,” J. Theor. Biol., vol. 463, pp. 77–91, Feb. 2019, doi: 10.1016/J.JTBI.2018.12.010.
28. B. Remeseiro and V. Bolon-Canedo, “A review of feature selection methods in medical applications,” Comput. Biol. Med., vol. 112, p. 103375, Sep. 2019, doi: 10.1016/J.COMPBIOMED.2019.103375.
29. S. Maldonado and J. López, “Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification,” Appl. Soft Comput. J., vol. 67, pp. 94–105, 2018, doi: 10.1016/j.asoc.2018.02.051.
30. V. Bolón-Canedo and A. Alonso-Betanzos, “Ensembles for feature selection: A review and future trends,” Inf. Fusion, vol. 52, pp. 1–12, Dec. 2019, doi: 10.1016/J.INFFUS.2018.11.008.
31. B. Pes, “Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains,” Neural Comput. Appl., vol. 32, no. 10, pp. 5951–5973, 2020, doi: 10.1007/s00521-019-04082-3.
32. B. Seijo-Pardo, V. Bolón-Canedo, and A. Alonso-Betanzos, “Using a feature selection ensemble on DNA microarray datasets,” *ESANN 2016–24th Eur. Symp. Artif. Neural Networks*, no. January 2017, pp. 277–282, 2016.
33. Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nat.* 2015 *5217553*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
34. J. M. Johnson and T. M. Khoshgoftaar, “Survey on deep learning with class imbalance,” J. Big Data, vol. 6, no. 1, Dec. 2019, doi: 10.1186/S40537-019-0192-5.
35. A. Reyes-Nava, H. Cruz-Reyes, R. Alejo, E. Rendón-Lara, A. A. Flores-Fuentes, and E. E. Granda-Gutiérrez, “Using deep learning to classify class imbalanced gene-expression microarrays datasets,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11401 LNCS, pp. 46–54, 2019, doi: 10.1007/978-3-030-13469-3_6.
36. A. M. Fred Agarap, “Deep Learning using Rectified Linear Units (ReLU),” Mar. 2018, Accessed: Dec. 27, 2021. [Online]. Available: https://arxiv.org/abs/1803.08375v2.

Figures

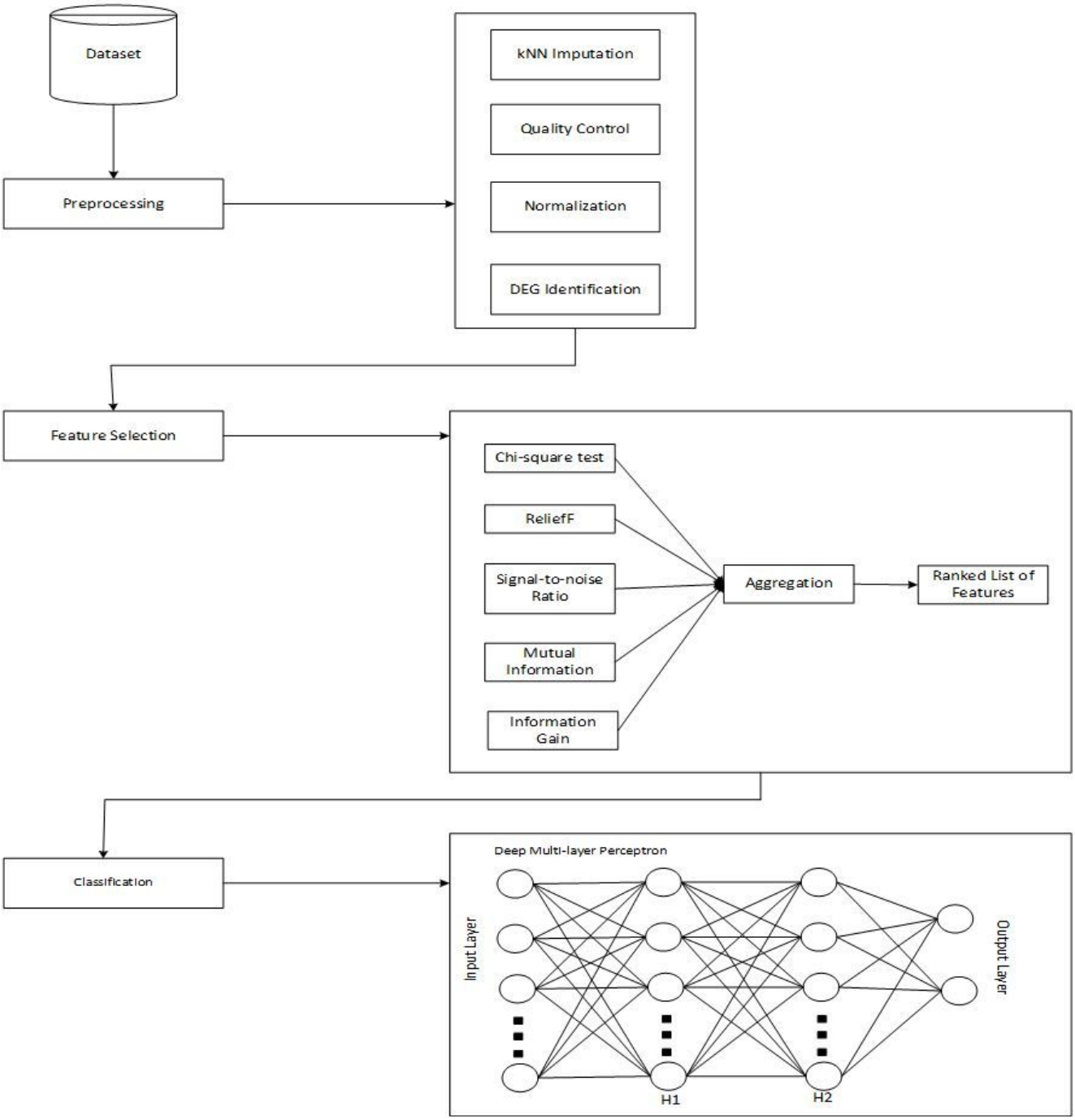

Figure 1

Process Flow

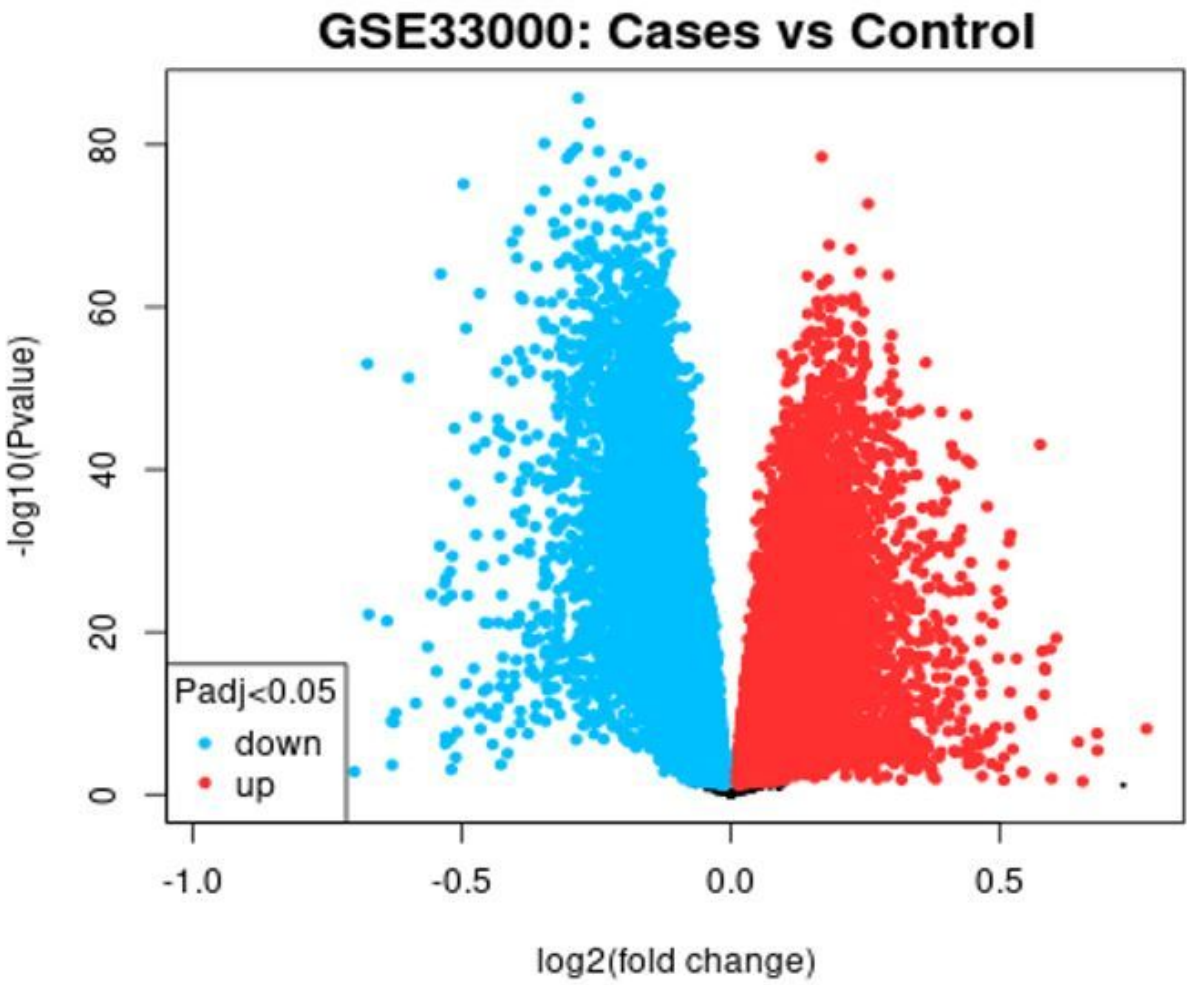

Figure 2

Volcano Plot (GSE33000)

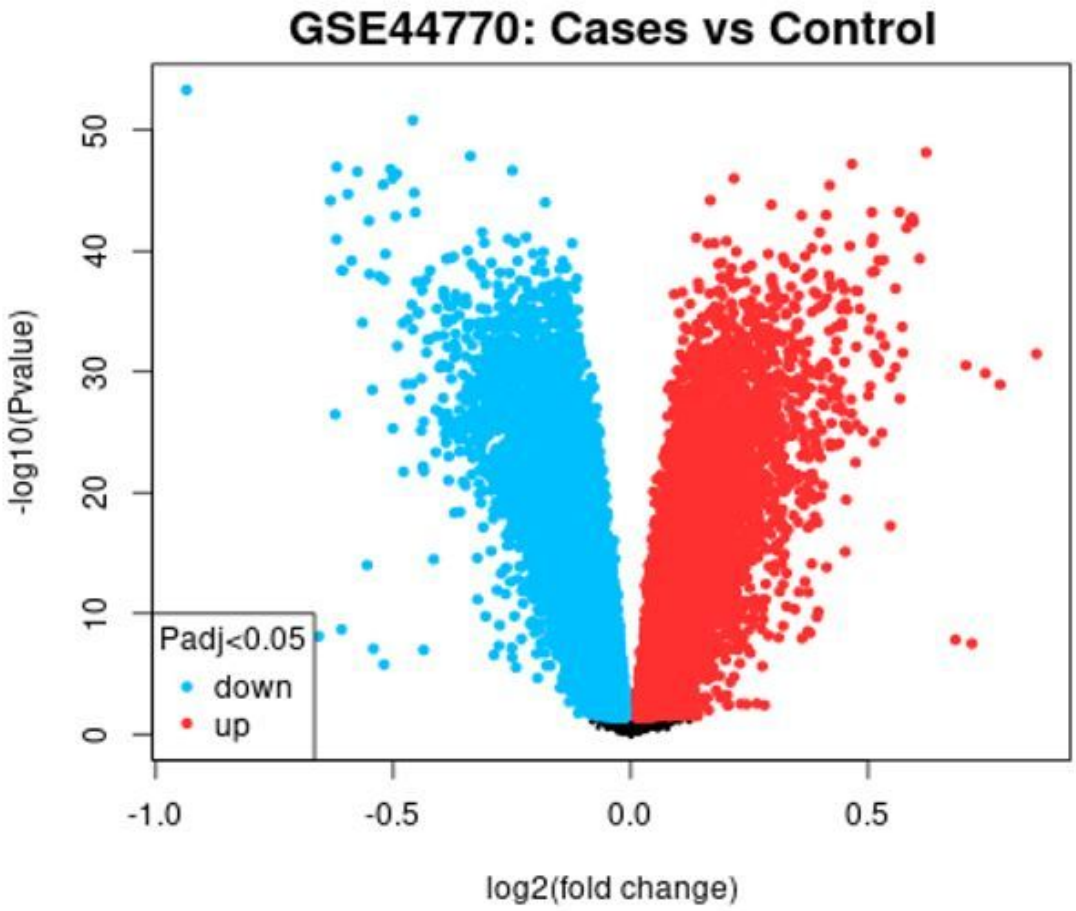

Figure 3

Volcano Plot (GSE44770)

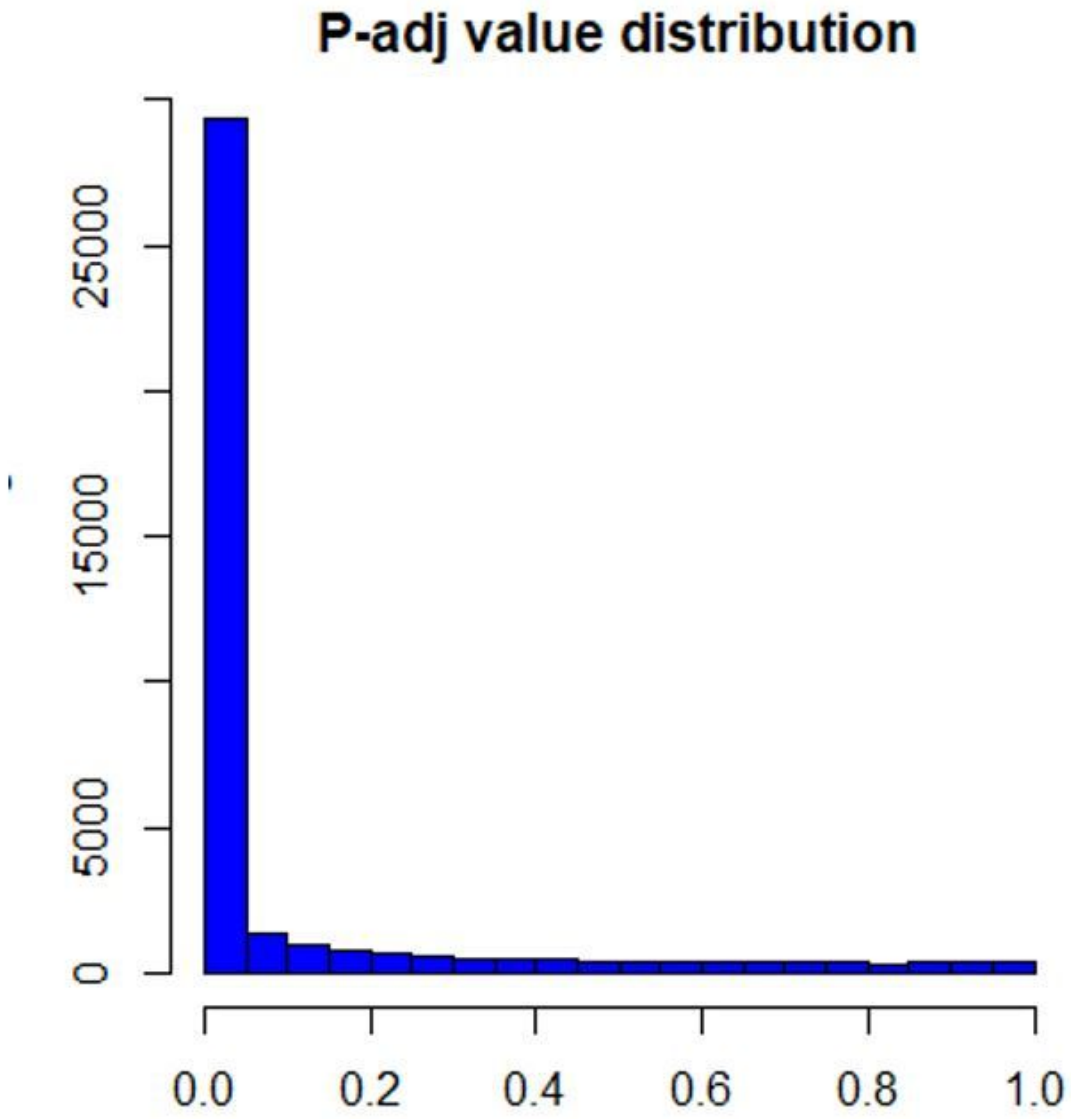

Figure 4

Adj P-value (GSE33000)

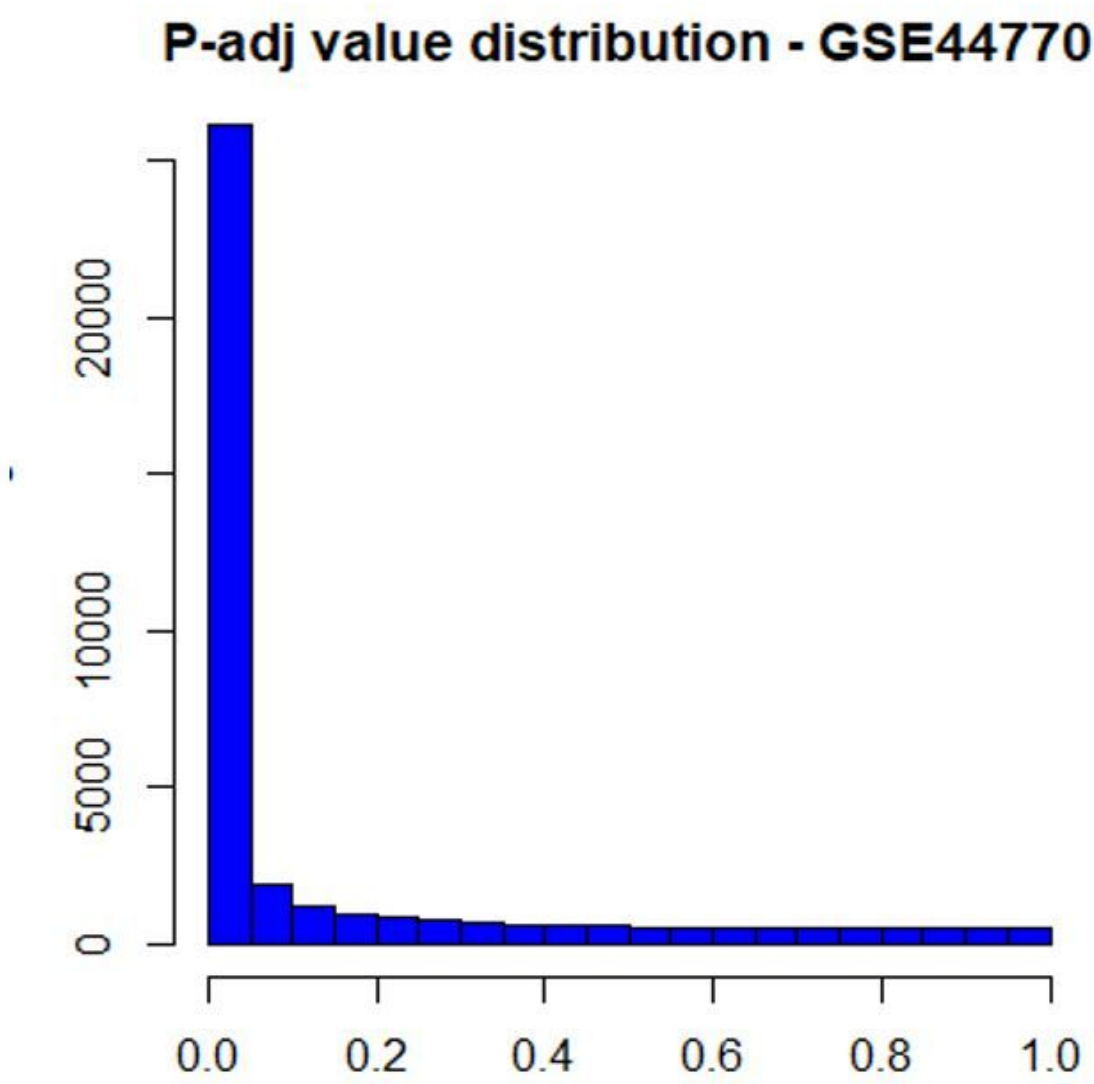

Figure 5

Adj p-value (GSE44770)

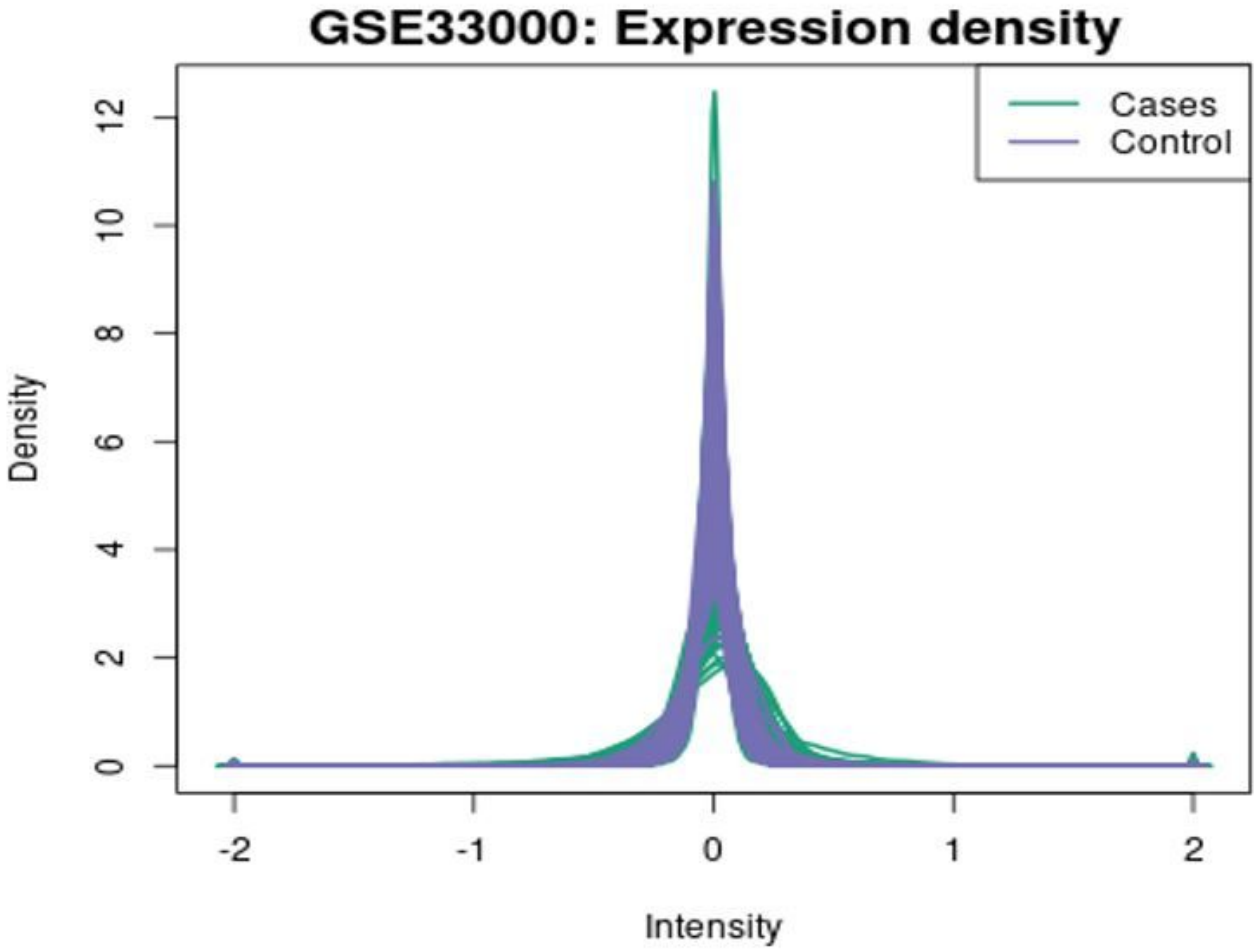

Figure 6

Density Plot (GSE33000)

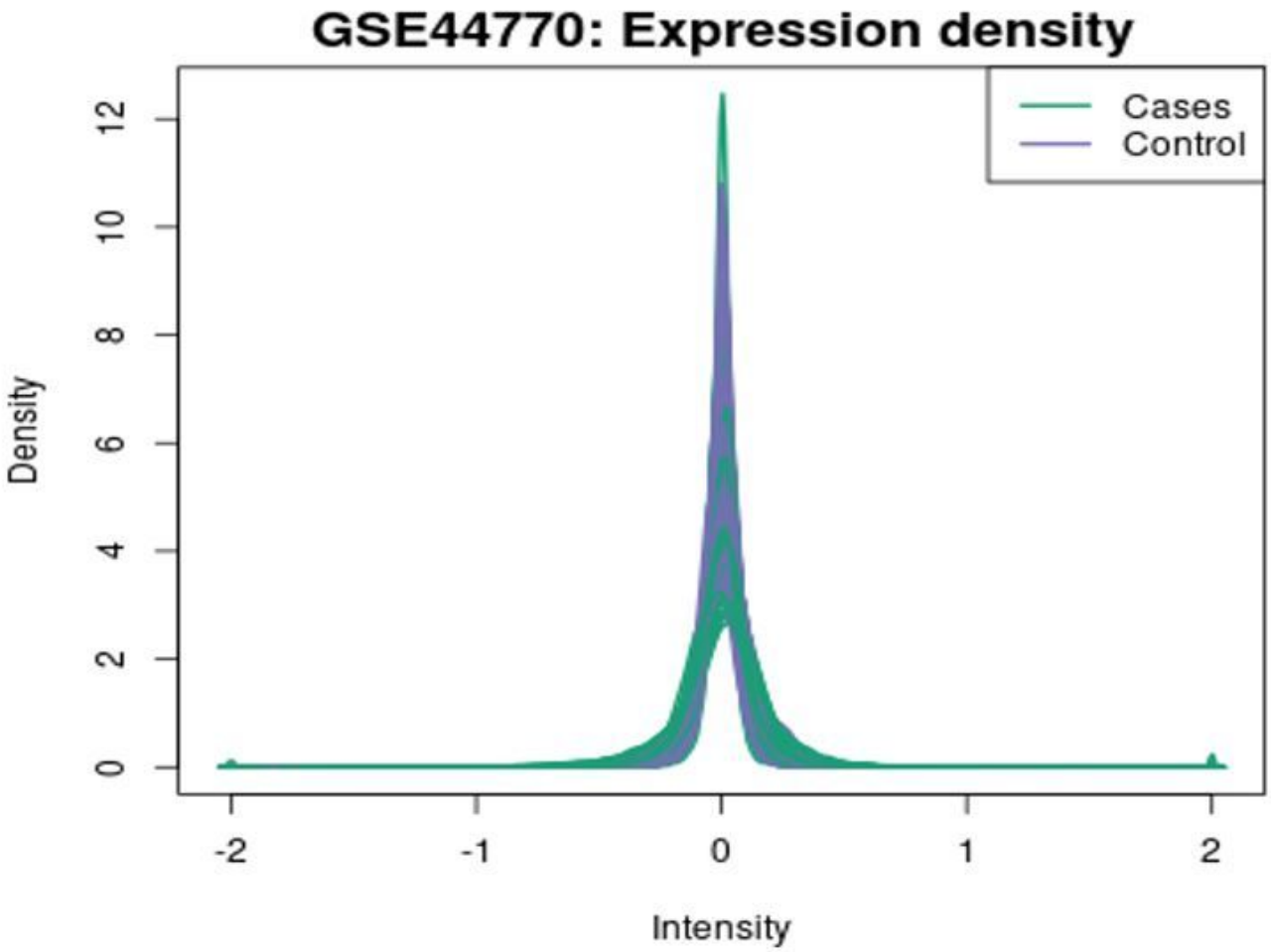

Figure 7

Density Plot (GSE44770)

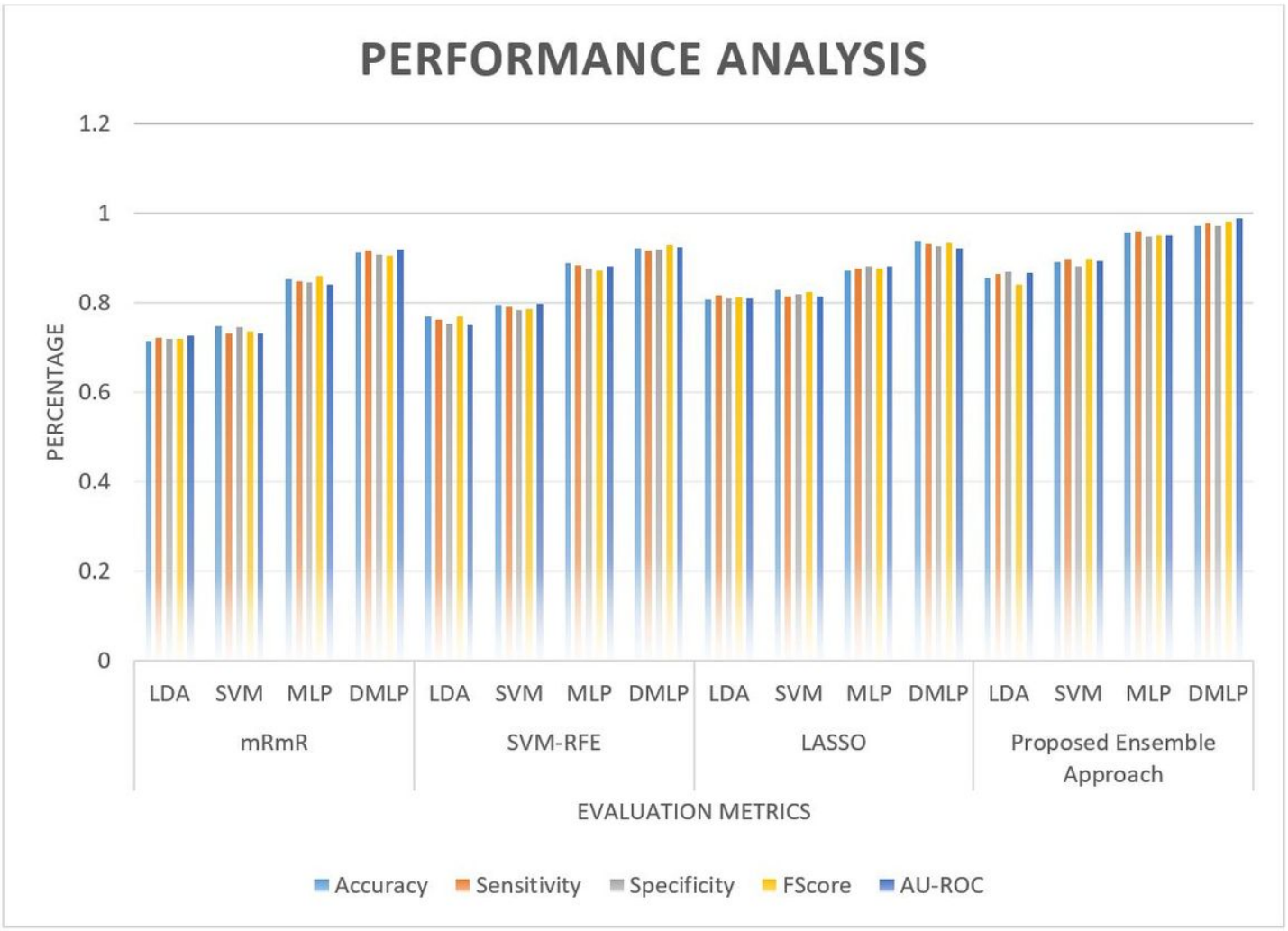

Figure 8

Performance Analysis