Detection of Novel Biomarker Genes of Alzheimer's Disease Using Gene Expression Data

Shehan Perera
Department of Computer Engineering
University of Peradeniya
Peradeniya, Sri Lanka
mario.perera@eng.pdn.ac.lk

Kaveesha Hewage
Department of Computer Engineering
University of Peradeniya
Peradeniya, Sri Lanka
kaveesha.dilshani@eng.pdn.ac.lk

Chamara Gunarathne
Department of Computer Engineering
University of Peradeniya
Peradeniya, Sri Lanka
chamarag@eng.pdn.ac.lk

Rajitha Navarathna
99X Technology
Colombo, Sri Lanka
rajithae03@gmail.com

Damayanthi Herath
Department of Computer Engineering
University of Peradeniya
Peradeniya, Sri Lanka
damayanthiherath@eng.pdn.ac.lk

Roshan G. Ragel
Department of Computer Engineering
University of Peradeniya
Peradeniya, Sri Lanka
roshanr@eng.pdn.ac.lk

***Abstract*—It is well recognized, that most common form of dementia is Alzheimer's disease and a successful cure or medication is not discovered. A plethora of research has been conducted to understand the underlying mechanism and the pathogenesis of the Alzheimer's disease. To explore the underlying genetic structure of the disease, gene expression data is being used by many researches and computational and statistical approaches were used to identify possible genes that are risk. In this paper, we propose a machine learning framework that can be used to identify possible bio-marker genes. Our experiments discover possible set of 14 genes, which some of them are validated by biological sources. We also present a critical analysis of the propose machine learning framework using GSE5281 gene dataset.**

***Index Terms*—machine learning, alzheimer's disease, feature engineering, gene expression**

I. Introduction

Alzheimer's disease (AD) has been known for more than 100 years and it accounts for 60 – 80 % of dementia cases [1]. Currently, there are around 50 million people suffering from dementia worldwide. It is an irreversible, neurodegenerative brain disease featuring clinical symptoms usually starting at an age of over 65 years. The symptoms include short-term memory loss, challenges in completing daily activities, bafflement, problems in speaking and writing, changes in behavior and mood swings [2]. Although in many cases, the symptoms become significantly visible only after the age of 65, the disease may have initiated 20 years before the symptoms become clearly visible [3].

To understand the exact reasons for the cause of the disease and to treat it before the symptoms begin to appear, a clear understanding of the underlying structure of the disease is needed. Gene expression is one such form of data that can be applied to deepen our knowledge of the disease. However, the analysis of gene expression data is computationally and statistically expensive. To overcome this method, various machine learning techniques are used.

II. Related Work

There are many pieces of research carried out to find the best biomarker genes using machine learning techniques. A study by Pang et al. [4] has carried out a study in which they explore the relevance of a Special Local Clustering Algorithm to group the similar set of genes. Through this approach, they were able to identify the significantly varied genes as isolated points. Kong et al. [5] in 2009 had applied two unsupervised analysis methods for microarray data analysis, principal component analysis (PCA) and independent component analysis (ICA). The results obtained through this study is further enhanced by Kong et al. [6] where they have presented an integration of the ICA method and nonnegative matrix factorization (NMF) to identify the significant genes related to AD. Many studies including the work by Pang et al. [4] and Walker et al. [7] have only considered identifying the most informative individual genes among the gene set. Also, interpretation was a challenging task due to large number of genes, gene complexity and the high collinearity among the gene expression profiles for biologists. Therefore, computational and mathematical models were used to extract underlying features from the multivariable datasets. Scheubert et al. [8] propose to use three feature selection methods namely (i) information gain, (ii) random forest and (iii) a wrapper approach of genetic algorithm & support vector machine (SVM) [9] to identify the best method for the identification using a small biomarker set.

Although the top-ranking genes provide valuable insights and information about the genes, the molecular mechanism of the disease is hard to explain through a function of a single cell. Therefore, it is important to focus on the small and less redundant sets of genes and gene pairs which contribute more to the disease. In our work, our main aim is to find the best biomarker genes that could be risk factors for Alzheimer's disease. We also analyze and compare the best feature selection techniques along with the classifiers. In our solution, we analyze the contribution of PCA, feature importance from the random forest and extra tree classifier as feature selection

978-1-7281-9975-7/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on November 19,2022 at 14:03:30 UTC from IEEE Xplore. Restrictions apply.

techniques. We obtain the best possible biomarker genes and validate them with the already identified risk genes to interpret novel results.

III. Method

The proposed framework consists of data pre-processing, feature reduction, feature extraction, modelling, classification and biological validation.

A. Data Pre-processing

We use GSE5281 [10] brain dataset and map the probe set IDs to gene official symbols according to the GPL570 annotation table. The dataset contains 161 samples out of which 87 are diagnosed with Alzheimer's disease and 74 samples are from the healthy control group. Totally, there are 24,438 unique gene symbols. We split the analysis of the dataset into two parts. They are:

- Analyzing the performance of classification algorithms
- Discovering the best biomarkers

B. Differentially Expressed Genes

As the first step of our analysis, we need to filter out the differentially expressed genes. These genes have significantly different expression values between the affected samples by Alzheimer's disease and control samples.

Through the t-test, pValue for each gene is calculated. In our research, the null hypothesis would be that the mean of the two classes of samples is equal. If pValue < 0.05, then null hypothesis has been rejected accepting the alternative hypothesis where there is a difference between the mean values of AD and non-AD samples. By this, filtering out the features with pValue < 0.05, we ensure that we get the most significantly expressed genes. We evaluate the pValue of each gene across the control and the disease samples via Welch's two-sample t-test [11].

We also calculate the corresponding fold value. Fold change was determined as:

$$\log FC_i = X_i – Y_i, \tag{1}$$

where X_i is the mean of the gene expression values of gene i for AD samples, and Y_i is the mean of the gene expression values of gene i for non-AD samples. Both values are in log-scale.

C. Feature Selection

PCA: This technique uses orthogonal transformation to convert a set of observations from correlated features into a set of linearly uncorrelated variables called principal components. In this research, we have selected a total of 50 unique genes by selecting the highest scored feature in each component. PCA is shown to be effective in reducing the number of features of genes [12].

Random forest: This consists of a number of decision trees. Every node in the decision tree is a condition on a single feature, designed to split the dataset into two so that similar response values end up in the same set [13]. The optimal condition is based on a measure called impurity. For a forest, the impurity decrease from each feature can be averaged and the features are ranked according to this measure. We selected the 50 features with the highest scores.

Extra Tree Classifier: Feature importance by extra tree classifier implements a meta estimator that fits a number of randomized extra trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The features with the highest 50 scores are sorted out by this method as well.

D. Classification

After performing statistical sets, we evaluated the performance for a dataset with 2000 genes. To ensure that results are not overly optimistic, the training was performed on two-thirds of data set while the other one third was used for testing.

For our experiments, we use six classification methods namely, (a) Naive Bayes, (b) Decision trees, (c) Random forest, (d) Nearest neighbor, (e) SVM with linear and (f) Gaussian kernel. We report the results using accuracy metric. Simple classifiers and ensemble learning is used for two class classification.

IV. Results and Discussion

We used univariate selection which is a statistical test that could be used to select those features that have the strongest relationship with the output variable (AD or non AD), to reduce thousands of differential genes upto small subsets as it works better with larger datasets. We selected 200 features with the highest scores for further analysis.

A. Classification

As the first step of our analysis, we look at the classification performance of five different classification methods, implemented as described in the Methods section. Table I shows the cross-validated classification accuracy for all five methods on the dataset.

TABLE I
Testing and Training Accuracy for 2000 Genes

Model	Testing	Training
Naive Bayes	81.63%	83.93%
C4.5 decision tree	87.76%	78.57%
Nearest neighbor	85.71%	86.58%
Random Forest	83.67%	87.48%
SVM + Gaussian kernel	83.67%	91.06%
SVM + linear kernel	89.80%	93.76%

For both training and testing, SVM with linear kernel shows the highest classification accuracy with 93.8% on training and 89.8% on testing. The lowest accuracies are obtained by the C4.5 decision tree classifier [14] and the Naive Bayes classifier compared to other classifiers. As SVM with linear kernel shows the best results, we use this classifier for evaluating the quality of the genes selected with different feature selection methods.

We use feature selection by PCA, random forest and the extra tree classifier to find biomarkers with potentially high importance in Alzheimer's disease.

Experimental results with different classifiers with the selected features is given in Fig. 1. The features extracted are proven to have an effect on AD according to Fig. 1. As shown

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on November 19,2022 at 14:03:30 UTC from IEEE Xplore. Restrictions apply.

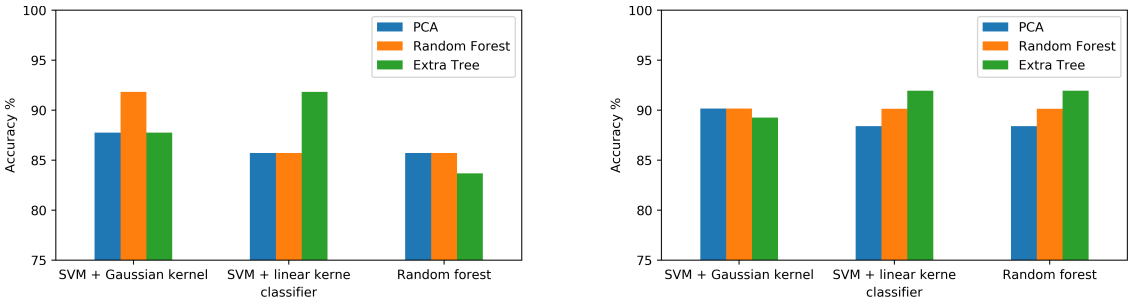

Fig. 1. Testing (left) and training (right) accuracies of three classifiers based on features extracted from three feature selection methods.

in Fig. 1, PCA has identified more significantly different genes compared with other techniques.

B. Performance of Small Biomarker Sets

Next, we calculated the correlation values of each gene that are only identified by one feature selection method. Fig. 2 shows the correlation matrix plot for 34 genes selected by PCA, 18 genes by extra tree classifier and 18 genes from random forest. We also plot the correlated feature percentages with different correlation values as shown in Fig. 3. When correlation value is 0.86, it has 40% correlated features in extra tree and 60% correlated features in random forest whereas the correlation value for PCA is just 0%. This proves that PCA is better in finding new biomarker genes that provide new information. Therefore, the genes identified by PCA may provide a novel insight into molecular processes related to AD, than the other two approaches.

The resulting 50 genes of each fold are sorted by their importance based on applying PCA, random forest and the extra tree classifier. Then we select the best 50, 40, 30, 20, 15, 10, 5, 3, 2 and 1 genes for each algorithm and utilize these to compute the classification performance using an SVM with gaussian kernel, linear kernel as well as random forest (see Fig. 4). As shown in Fig. 4, the highest accuracy is given when the number of features is in the range of 10-15. This means that the models work better with small sets of genes than with a large number.

All three feature selection methods give 8 common genes. We call these 8 genes as the overlapped set of genes. According to Fig. 5, these overlapped genes give a higher accuracy than any other gene set obtained through single feature selection method. But according to Fig. 4, the classifiers give the best accuracy when the number of selected features is in the range of 10-15. To get the optimal number of genes selected, we consider correlation coefficients. By considering the correlation coefficient values, we select another new 9 genes. Genes SLC39A12, CTD-3092A11.2 and RP11-271C24.3 were found in both the overlapped set and the newly found 9 genes. Therefore, we removed these three genes from the overlapped set of genes. The overlapped genes together with the new 9 genes shows 93.9% accuracy by SVM and 91.4% accuracy with random forest classifier. Table II shows the accuracy, sensitivity and specificity of the model with the 9 genes found only by PCA.

When considering the correlation matrix of selected genes (set of overlapped genes (8) with the 9 genes found by the correlation coefficients of genes selected by PCA alone), it was visible that these genes were less co-related. We selected the less co-related genes to minimize the redundancy of the selected genes. The resulting genes are less correlated and might provide new information on the biomarker genes. Removing highly correlated genes and minimizing redundancy in genes can be useful in two main scenarios [13]. They are:

- When considering the design of diagnostic tools, where having a small set of probes is often desirable.
- To help understand the results from other gene selection approaches that return many corelated genes. This helps to understand which ones of those genes have the largest signal to noise ratio. This approach will give information on which genes could be used as substitutes for complex processes involving many correlated genes.

Next, we combine overlapped genes with these genes and compare the accuracy, sensitivity and specificity. Table III shows the accuracy, sensitivity and specificity of the model with the newly found genes from PCA together with the overlapped set. According to Table III, SVM + gaussian kernel gives highest testing accuracy. The mentioned set of genes gives the highest accuracy that have been achieved so far in our study.

TABLE II
PERFORMANCE EVALUATION OF THE 9 GENES FOUND ONLY BY PCA

Model	Accuracy	Sensitivity	Specificity
SVM + Gaussian kernel	93.9%	98.6%	96.6%
SVM + linear kernel	93.9%	98.6%	97.7%
Random Forest	91.84%	95.9%	96.6%

TABLE III
PERFORMANCE EVALUATION OF THE NEWLY FOUND GENES BY PCA AND THE OVERLAPPED SET

Model	Accuracy	Sensitivity	Specificity
SVM + Gaussian kernel	93.9%	97.3%	96.6%
SVM + linear kernel	93.9%	98.6%	97.7%
Random Forest	91.84%	95.8%	94.4%

Our method selected 14 genes, which may be biomarker for Alzheimer's disease. The selected gene symbols

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on November 19,2022 at 14:03:30 UTC from IEEE Xplore. Restrictions apply.

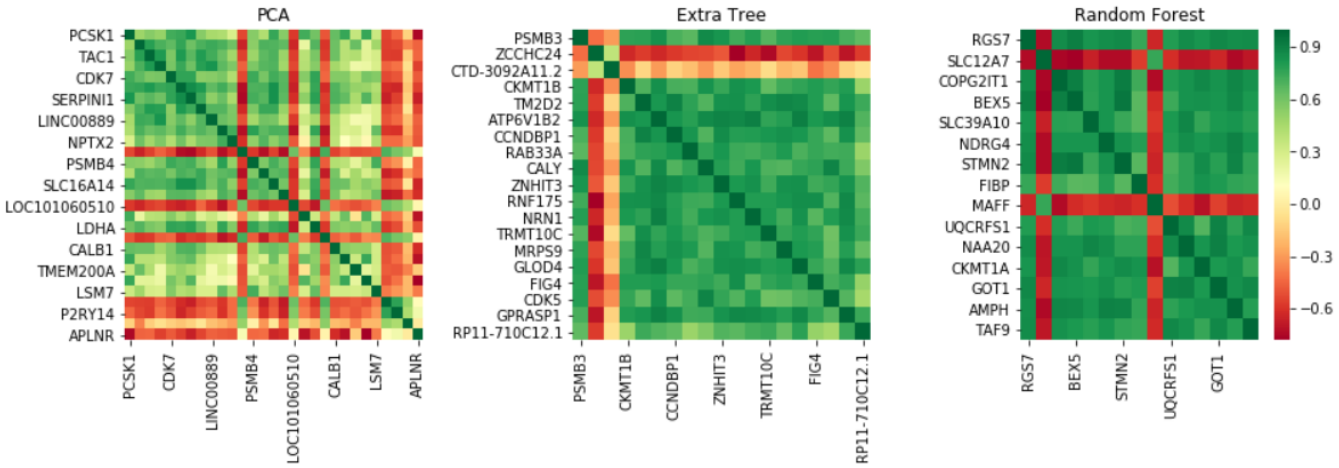

Fig. 2. Correlation matrix plots. Correlation plots considering the features found only by that particular feature selection method. (left to right) PCA, Feature importance by Extra tree classifier and Random forest.

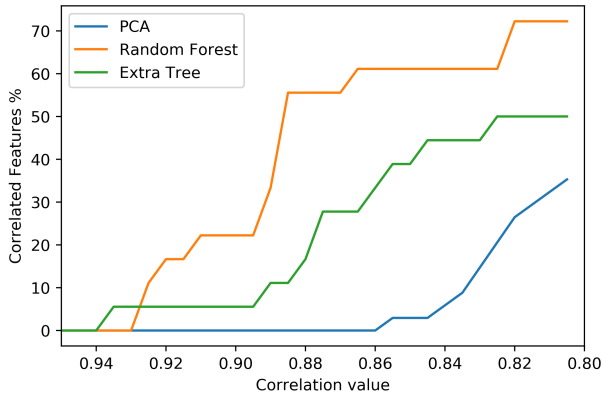

Fig. 3. An interpretation of the number of correlation features with the variation of correlation values. Note that PCA gives a far less indication of correlated genes. In PCA,there is no correlation found untill the correlation value is 0.86 whereas the percentages of correlated genes at the same point by other two methods is very large.

are AC004951.6, MAFF, SLC39A12, PCYOX1L, CTD-3092A11.2, RP11-271C24.3, PRO1804, PRR34-AS1, SST, CHGB, MT1M, JPX, APLNR and PPEF1.

C. Biological Validation

We find a significant enrichment for nearly all of the tested gene sets compared with in the list of genes associated with Alzheimer's disease by GeneCards [15]. There are 14 already identified genes within the 50 genes we selected from PCA. Our method concludes 14 genes can be novel genes for Alzheimer's disease. According to GeneCards, out of those predicted 14 genes, 4 genes have been discovered as AD related. SST, CHGB, SLC39A12 and MT1M are the 4 genes that have already been identified in the GeneCards. SST was identified by Squillario and Barla [16] as part of a 39 gene signature implicated in Alzheimer's disease. SST (somatostatin) expression has been shown to be less in cortex and hippocampus of Alzheimer affected brains [17]. Somatostatin also affects rates of neurotransmission in the central nervous system and is identified as co-related with memory impairment and cognitive function [17]. CHGB, which was involved in the signal-mediated sorting process and interfered by protein misfolding, was possibly associated with the increased vulnerability of motor neurons [18]. This gene encodes a tyrosine-sulfated secretory protein abundant in peptidergic endocrine cells and neurons. This protein may serve as a precursor for regulatory peptides. MT1M is also identified by GeneCards as a gene that has an impact on the disease. This gene encodes a member of the metallothionein superfamily, type 1 family. Metallothioneins have a high content of cysteine residues that bind various heavy metals. Some research have pointed out that dysregulated metal homeostasis is associated with Alzheimer's Disease. AD patients have decreased cortex and elevated serum copper levels along with extracellular amyloid-beta plaques containing copper, iron, and zinc. Myhre et al. [19] has pointed out MT1M as the most prominent gene involved in brain metal homeostasis pathways for which expression is significantly changed in AD. Previous research have found that over-expressed small MAFs can form homodimers and act as transcriptional repressors. Therefore, MAFF might play an important role in dysfunction of NRF2 regulatory network in AD [20]. Schubert et al. [8] have also found PCYOX1L as their highest-ranked gene that could have a potential risk towards the disease. However, a clear relationship between PCYOX1L and AD is not yet defined. But PCYOX1L has been identified as a peripheral priority gene in Parkinson's disease [21]. Many researchers have argued that there could be a relationship between Parkinson's and Alzheimer's disease implying that PCYOX1L also might have an impact on Alzheimer's disease as well.

D. Correlation Between the Selected Genes

The correlation between the 14 genes selected could be interpreted using the Graphical Lasso method. Fig. 6 shows

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on November 19,2022 at 14:03:30 UTC from IEEE Xplore. Restrictions apply.

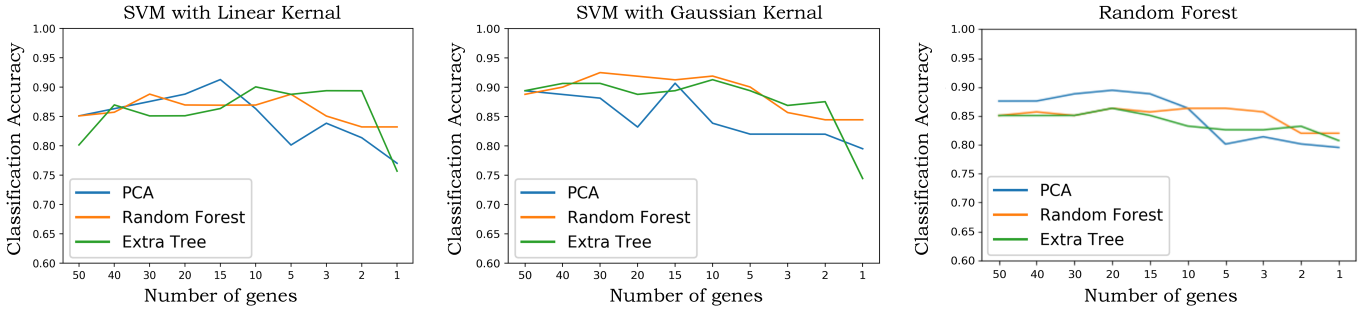

Fig. 4. Classification accuracy of selected genes. Classification accuracy of three classifiers using incrementally smaller sets of genes, identified by our three feature selection methods trained by three classifiers.

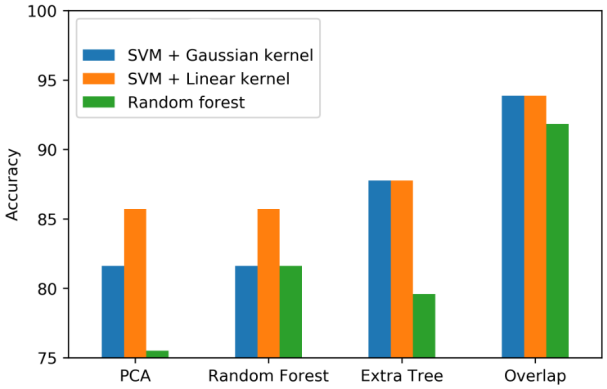

Fig. 5. Accuracy of the model with genes only found by each feature selection method when trained upon three classifiers; SVM with Gaussian and linear kernel and random forest.

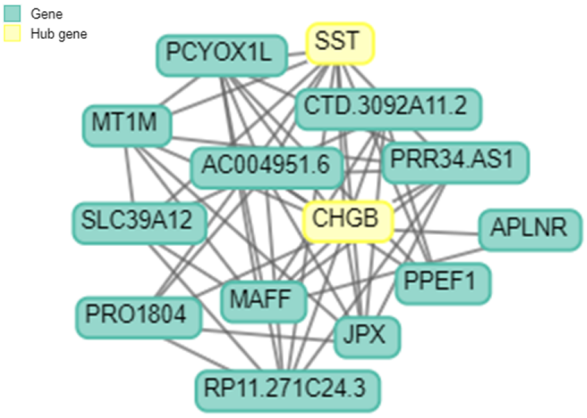

Fig. 6. Gene regulatory network constructed using graphical lasso method for the 14 genes selected.

a gene regulatory network for the 14 genes constructed using GeNeCk. It is a web server for gene network construction and visualization [22]. A gene regulatory network reveals how genes work together to carry out various biological processes. In gene networks, a gene that has many interactions with other genes is called a hub gene, which usually plays an essential role in gene regulation and biological processes [23]. Even in this network, it is clear that SST and CHGB act as hub genes in this network, which are two prominent biomarker genes already identified as prominent by the literature.

V. Conclusion

In this paper, we propose a machine learning framework that can be used to identify possible bio-marker genes. Through feature importance scores from random forest and extra tree classifier and co-relation matrix, we were able to identify 14 new candidate biomarker genes for Alzheimer's disease.

We also focused on reducing redundant genes and the importance of principle component analysis method in finding the less co-related genes. The usage of principle component analysis in a novel approach has led to the removal of highly co-related genes.

We expect that future biological experiments can test some of our computational predictions. In the future, we will test our method on many more datasets, representing wide variety of cellular phenotypes and diseases.

References

[1] A. Association *et al.*, "2018 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.

[2] K. G. Yiannopoulou and S. G. Papageorgiou, "Current and future treatments for Alzheimer's disease," *Therapeutic advances in neurological disorders*, vol. 6, no. 1, pp. 19–33, 2013.

[3] V. L. Villemagne, S. Burnham, P. Bourgeat, B. Brown, K. A. Ellis, O. Salvado, C. Szoeke, S. L. Macaulay, R. Martins, P. Maruff *et al.*, "Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study," *The Lancet Neurology*, vol. 12, no. 4, pp. 357–367, 2013.

[4] C.-Y. Pang, W. Hu, B.-Q. Hu, Y. Shi, C. R. Vanderburg, J. T. Rogers, and X. Huang, "A special local clustering algorithm for identifying the genes associated with Alzheimer's disease," *IEEE transactions on nanobioscience*, vol. 9, no. 1, pp. 44–50, 2010.

[5] W. Kong, X. Mou, Q. Liu, Z. Chen, C. R. Vanderburg, J. T. Rogers, and X. Huang, "Independent component analysis of Alzheimer's DNA microarray gene expression data," *Molecular neurodegeneration*, vol. 4, no. 1, p. 5, 2009.

[6] W. Kong, X. Mou, and X. Hu, "Exploring matrix factorization techniques for significant genes identification of Alzheimer's disease microarray gene expression data," in *BMC bioinformatics*, vol. 12, no. S5. Springer, 2011, p. S7.

[7] P. R. Walker, B. Smith, Q. Y. Liu, A. F. Famili, J. J. Valdés, Z. Liu, and B. Lach, "Data mining of gene expression changes in Alzheimer brain," *Artificial intelligence in medicine*, vol. 31, no. 2, pp. 137–154, 2004.

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on November 19,2022 at 14:03:30 UTC from IEEE Xplore. Restrictions apply.

[8] L. Scheubert, M. Luštrek, R. Schmidt, D. Repsilber, and G. Fuellen, "Tissue-based Alzheimer gene expression markers–comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets," *BMC bioinformatics*, vol. 13, no. 1, p. 266, 2012.
[9] E. B. Huerta, B. Duval, and J.-K. Hao, "A hybrid GA/SVM approach for gene selection and classification of microarray data," in *Workshops on Applications of Evolutionary Computation*. Springer, 2006, pp. 34–44.
[10] "GEO Accession viewer," [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5281. [Accessed: 11-Jul-2020].
[11] J. Algina, T. Oshima, and W.-Y. Lin, "Type I error rates for Welch's test and James's second-order test under nonnormality and inequality of variance when there are two groups," *Journal of Educational Statistics*, vol. 19, no. 3, pp. 275–291, 1994.
[12] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.
[13] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, p. 3, 2006.
[14] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree ID3 and C4. 5," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 2, pp. 13–19, 2014.
[15] "GeneCards®: The Human Gene Database," [Online]. Available: https://www.genecards.org/. [Accessed: 11-Jul-2020].
[16] M. Squillario and A. Barla, "A computational procedure for functional characterization of potential marker genes from molecular data: Alzheimer's as a case study," *BMC medical genomics*, vol. 4, no. 1, p. 55, 2011.
[17] J. Epelbaum, J.-L. Guillou, F. Gastambide, D. Hoyer, E. Duron, and C. Viollet, "Somatostatin, Alzheimer's disease and cognition: an old story coming of age?" *Progress in neurobiology*, vol. 89, no. 2, pp. 153–161, 2009.
[18] L.-M. Chi, X. Wang, and G.-X. Nan, "In silico analyses for molecular genetic mechanism and candidate genes in patients with Alzheimer's disease," *Acta Neurologica Belgica*, vol. 116, no. 4, pp. 543–547, 2016.
[19] O. Myhre, H. Utkilen, N. Duale, G. Brunborg, and T. Hofer, "Metal dyshomeostasis and inflammation in Alzheimer's and Parkinson's diseases: possible impact of environmental exposures," *Oxidative Medicine and Cellular Longevity*, vol. 2013, 2013.
[20] Q. Wang, W.-X. Li, S.-X. Dai, Y.-C. Guo, F.-F. Han, J.-J. Zheng, G.-H. Li, and J.-F. Huang, "Meta-analysis of Parkinson's disease and Alzheimer's disease revealed commonly impaired pathways and dysregulation of nrf2-dependent genes," *Journal of Alzheimer's Disease*, vol. 56, no. 4, pp. 1525–1539, 2017.
[21] L. B. Moran and M. B. Graeber, "Towards a pathway definition of Parkinson's disease: a complex disorder with links to cancer, diabetes and inflammation," *Neurogenetics*, vol. 9, no. 1, pp. 1–13, 2008.
[22] M. Zhang, Q. Li, D. Yu, B. Yao, W. Guo, Y. Xie, and G. Xiao, "GeNeCK: a web server for gene network construction and visualization," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–7, 2019.
[23] D. Yu, J. Lim, X. Wang, F. Liang, and G. Xiao, "Enhanced construction of gene regulatory networks using hub gene information," *BMC bioinformatics*, vol. 18, no. 1, p. 186, 2017.

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on November 19,2022 at 14:03:30 UTC from IEEE Xplore. Restrictions apply.