Network Modeling Analysis in Health Informatics and Bioinformatics (2022) 11:10
https://doi.org/10.1007/s13721-021-00349-9

ORIGINAL ARTICLE

Identifying the candidate genes using co-expression, GO, and machine learning techniques for Alzheimer's disease

Shailendra Sahu[1] · **Pankaj Singh Dholaniya**[2] · **T. Sobha Rani**[1]

Received: 3 August 2021 / Revised: 20 November 2021 / Accepted: 22 November 2021 / Published online: 4 February 2022

Abstract

Alzheimer's disease is a neurological disorder that affects an individual's memory, motor functions, behaviour, and thought process. It has been observed that the hippocampus is the first region that gets affected by Alzheimer's. Hence, a study of the hippocampus region can identify genes responsible for the occurrence of the early stage of the disease. Most often, t-test and correlation are used to identify significant genes at the initial level. As the genes are differentially expressed, their classification power is generally high. These genes might appear significant, but their degree of specificity towards the disease might be low, leading to misleading interpretations. Similarly, there may be many false correlations between the genes that can affect the identification of relevant genes. This paper introduces a new framework to reduce the false correlations and find the potential biomarkers for the disease. The framework concerned uses the t-test, correlation, Gene Ontology (GO) categories, and machine learning techniques to find potential genes. The proposed framework detects Alzheimer-related genes and achieves more than 95% classification accuracy in every dataset considered. Some of the identified genes which are directly involved in Alzheimer are APP, GRIN2B, and APLP2. The proposed framework also identifies genes like ZNF621, RTF1, DCH1, and ERBB4, which may play an important role in Alzheimer's. Gene set enrichment analysis (GSEA) is also carried out to determine the major GO categories: down-regulated and up-regulated.

Keywords Microarray data · Gene co-expression network · Gene ontology similarity · Feature selection · Classification.

1 Introduction

Alzheimer's disease is a prevalent form of dementia. It is an irreversible disease with a progressive loss of memory and worsening cognitive function. The leading cause of AD is said to be the abnormal deposits of protein forms amyloid plaques and tau tangles throughout the brain (Alzheimer's 2015). Hippocampus is the brain region associated with all stages of semantic memory and is said to be affected first in AD (Duff and Covington 2020; Anand and Dhikav 2012). APOE is said to be the most common gene associated with AD (Alzheimer's 2015). Apart from APOE, APP, PSEN1, and PSEN2 are also observed as the cause of AD (Lanoiselée 2017). Various studies have been carried out to identify the genes which are differentially expressed in the AD affected brains (Lanoiselée 2017; Ray 2017). *T*-test and gene correlation networks are the most common statistical techniques used to identify the significant genes. The t-test is used to test the significant difference in gene expression levels (Zhou 2008). For example, Zhu and Yang (2016) used the rejection region of the *t*-test to identify the candidate gene for AD. However, the *t*-test only gives the significant difference in the mean expression values of genes between control and disease sets, which is not enough to determine the significant influence of genes on the disease. There could be many other reasons apart from the disease, which can result in a change in the expression value of a particular gene. Ray (2017) analyzed the preservation patterns of gene co-expression networks during Alzheimer's disease progression. Like the *t*-test, the correlation between two genes is not enough to tell that two correlated genes interact with

✉ Shailendra Sahu
shailendrasahu668@gmail.com

✉ Pankaj Singh Dholaniya
pankaz@uohyd.ac.in

✉ T. Sobha Rani
sobharani@uohyd.ac.in

[1] School of Computer and Information Sciences, University of Hyderabad, Hyderabad, India

[2] Department of Biotechnology and Bioinformatics, University of Hyderabad, Hyderabad, India

each other. Hu and Yu (2020) constructed a co-expression network using WGNCA and analyzed their clinical features. As a result, they identified four genes(ENO2, ELAVL4, SNAP91, and NEFM) said to be associated with AD. Xia et al. (2014) constructed the co-expression network using the method proposed by Ruan and Zhang (2006). Then, they ranked the genes based on a new topological overlap formula, a modified version of the formula described in Ray and Zhang (2010), Ray et al. (2012). The main concern with constructing a co-expression network using this method is that it depends on the user-defined value α. Different values of α result in a different number of edges. This means that every gene in the co-expression network is connected to its top α co-expressed genes. It may impact the removal of positive edges.

As the gene expression datasets are vast, various machine learning techniques are used along with the other statistical methods. Takahiro et al. (2016); Nishiwaki et al. (2016) used the random forest to identify the AD-related genes. In AL-Dlaeen and Alashqur (2014), AL-Dlaeen et al. used a decision tree classifier to predict the AD. There are many other algorithms, such as the K-means clustering algorithm, Principal component analysis(PCA), ant colony algorithm (ACO), independent component analysis algorithm (ICA), the angle cosine distance algorithm, and Chebyshev inequality algorithm (ACD), which produce less efficient and unstable results (Zhu and Yang 2016). Sharma and Dey (2021) combined two feature selection techniques, LASSO and Random forest, for gene selection and achieved a high classification accuracy. In Ramaswamy (2021), Ramya et al. used the *t*-test, signal-to-noise ratio, and f-test for the initial selection of genes and then selected genes were used in a modified particle swarm optimization algorithm to obtain further refined genes. Cheng and Liu (2021) observed that the machine learning model's average classification accuracy is higher than that of conventional methods. Apart from this, the authors also observed that machine learning approaches could also recognize oxidative phosphorylation genes in the Alzheimer's pathway. Saputra (2020) compared different decision trees with particle swarm optimization as feature selection methods and observed that the random forest gives the best accuracy. Kuang et al. (2021) compared the performance of three machine learning algorithms, artificial neural network (ANN), and decision tree and logistic regression models, to predict the AD. They found that ANN worked better than the other two models, and observed that the age, daily routine, urine neuronal thread protein associated with AD, smoking, alcohol intake, and sex are the crucial factors.

Almost every feature selection technique is applied on differentially expressed genes, i.e., genes obtained after the t-test. As the genes are differentially expressed, their classification power is generally high. These genes might appear significant, but their degree of specificity towards the disease might be low, leading to misleading interpretations. Some genes are expressed in basic cellular pathways and possess a higher probability of being differentially expressed across several biological conditions (Crow and Lim 2019). Nevertheless, as AD's causes probably include genetic, environmental, and lifestyle factors, different genes are identified as important in different AD datasets. Due to these various factors involved in AD, statistical methods and machine learning techniques alone are inadequate.

Table 1 Dataset description

Datasets	Control	AD
GSE48350 (dataset 1)	25	19
GSE5281 (dataset 2)	13	10
GSE28146 (dataset 3)	8	22

2 Dataset

The gene expression datasets GSE48350, GSE5281, and GSE28146, are downloaded from Gene Expression Omnibus (GEO), NCBI. The datasets GSE48350[1] (dataset 1) and GSE5281[2](dataset 2) contain gene expression data of control and Alzheimer's disease patients. The dataset GSE28146[3](dataset 3) contains microarray data of the hippocampal gray matter. The GSE48350 and GSE5281 datasets contain samples from different brain regions. We took only Hippocampus data for analysis as it is said to be affected first in Alzheimer's disease Anand and Dhikav (2012). Table 1 describes the data.

3 Proposed framework

This paper introduces a new framework, including t-test, correlation network, GO similarity matrix, and feature selection for filtering genes of less interest. Figure 1 shows the proposed framework.

Initially, differentially expressed genes are identified using the t-test. Then, the identified genes are used to create two separate correlation networks for AD and control sets using Pearson's correlation. There may be many false correlations, so a GO similarity matrix is introduced to reduce the false correlations. GO matrix consists of the number of similar GO terms between every pair of genes. Then, the GO similarity matrix is used to eliminate edges in the correlation

[1] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48350.
[2] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5281.
[3] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28146.

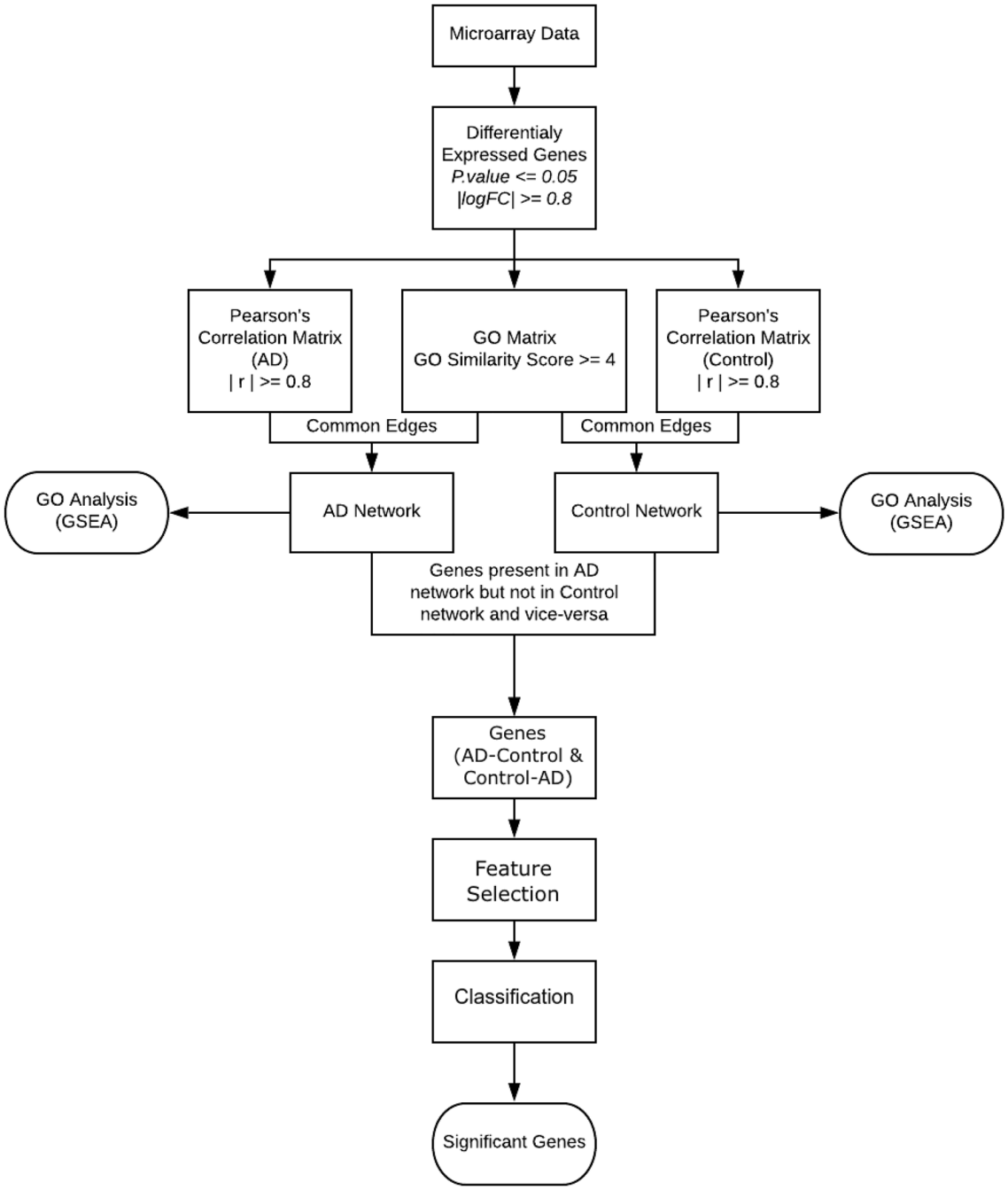

Fig. 1 Framework used to identify the potential biomarkers in Alzheimer's disease

networks that do not fall under the pre-defined criteria. The resultant correlation networks are then used for further analysis. Genes present in the control correlation network but not in the AD correlation network and vice versa are selected as the genes of interest. A separate Gene Set Enrichment Analysis (GSEA) has been carried out for selected genes to identify the affected GO categories. The feature selection algorithm is now applied to the selected genes to determine the most important genes from the important ones. In the final stage, the classification accuracy of the final set of genes is checked using a classification algorithm. All the components of the proposed framework are explained in detail in the following sections.

3.1 *T*-test

A *t*-test was performed on all the datasets, i.e., GSE48350, GSE5281, and GSE28146, to find the significant difference

in the expression values of genes in control and AD patients using GEO2R [NCBI]. p value ≤ 0.05 and fold count, $|logFC| \geq 0.8$ are used as the threshold values. These are standard values used in the literature. As many genes have different probe ids, we took the average expression and fold count values. 696, 7222, and 1893 Differentially expressed genes (DEGs) are obtained from dataset 1, dataset 2, and dataset 3, respectively.

3.2 Gene co-expression network

Pearson's correlation is used to calculate the correlation between each pair of genes after performing the t-test. ± 0.8 is taken as the threshold value as it is interpreted as strong/high correlation (Akoglu 2018; Mukaka 2012). They have pointed out that a correlation value of 0.7–0.9 indicates a high positive correlation and 0.9 as a very high positive correlation. Hence, a value of 0.8 is chosen as the threshold. All the correlation values which are greater than or equal to |0.8| are considered as 1, and the rest of the values are considered as 0. The resultant adjacency matrix is used to create the gene co-expression matrix. Two separate networks for control and AD are constructed using the binarised Pearson correlation values as edges.

3.3 GO similarity matrix

Gene ontology (GO) (Ashburner 2000) has become an accepted norm to evaluate the practical connections among gene products. GO is a scientific classification of biological terms identified with the properties of genes or their products. There are three GO categories: biological process, cellular component, and molecular function. Two proteins engaged with the same biological process are bound to interact than proteins engaged with various biological processes (Zhao and Wang 2018). Besides, two proteins need to come into close contact (essentially momentarily) to communicate; subsequently, co-localization can likewise be utilized to anticipate protein–protein interactions. Hence, the proposed framework uses GO categories for measuring the strength of the connection between genes in the correlation network.

GO similarity matrix consists of the GO similarity score between a pair of genes. Go similarity score is calculated as the number of common GO terms between two genes. For example, if Gene1 has 5 GO terms GO1, GO2, GO3, GO4, and GO5, and Gene2 has 4 GO terms GO1, GO3, GO5, and GO6. There are three common GO terms between the genes Gene1 and Gene2, which are GO1, GO2, and GO5. Hence, the GO similarity score ($GO_{(Gene1,Gene2)}$) between Gene1 and Gene2 is 3. GO categories of the differentially expressed genes (DEGs) identified by the t-test are used to construct the GO similarity matrix. The GO categories of all the DEGs are downloaded from DAVID (The Database for Annotation, Visualization, and Integrated Discovery) (Huang 2007). In the first dataset (GSE48350), out of 696 DEGs, 646 DEGs have known GO terms, and in the second dataset (GSE5281), out of 7222 DEGs, 6377 DEGs have known GO terms. In dataset 3 (GSE28146), out of 1893 DEGs, 1210 DEGs have known GO terms. All the three GO categories, i.e., Biological Process (BP), Molecular Function (MF), and Cellular Component (CC), are considered for the construction of the GO similarity matrix. Gene similarity matrix consists of the GO similarity score between all pairs of genes, as shown in Fig. 2.

$$\begin{Bmatrix} & Gene1 & Gene2 & ..... & GeneN \\ Gene1 & 0 & GO_{(1,2)} & ..... & GO_{(1,N)} \\ Gene2 & GO_{(2,1)} & 0 & ..... & GO_{(2,N)} \\ . & . & . & ..... & . \\ . & . & . & ..... & . \\ . & . & . & ..... & . \\ GeneN & GO_{(N,1)} & GO_{(N,2)} & ..... & 0 \end{Bmatrix}$$

Fig. 2 GO similarity matrix

This GO similarity matrix is used to create the GO network. To determine the cut-off score for the GO similarity score, 4000 genes (except the genes considered in the experiment) having nearly 11000 edges that are experimentally proven are taken [DAVID]. The GO similarities between the genes having experimentally proven interactions are analyzed. The average number of similar GO terms between two genes [having experimentally proven edges (interactions)] is 3.14. Hence, the ceiling value 4 is taken as the threshold value. All the edges whose weight (GO similarity score) is less than four are deleted. An edge between two genes is to be considered if they have at least four common GO terms.

3.4 Common genes and edges between GO and correlation networks

A combined network is constructed to take care of the false correlations by mapping gene correlation networks (Control and AD) to the GO network. As genes sharing more GO terms will tend to have a high biological association, combining the correlation and GO network helps to eliminate the edges with less biological significance (Martin et al. 2004; Zhao and Wang 2018). A combined AD network is constructed using the common edges between the AD correlation network and the GO network. A similar combined network is constructed for the control network using the control correlation network and GO network. Table 2 shows the count of edges in the correlation network and GO network.

The common control network consists of 240 genes and 673 edges in dataset 1, 2487 genes and 20486 edges in dataset 2, and 595 genes and 989 edges in dataset 3. The common AD network consists of 219 genes and 774 edges in

Table 2 Edge description of correlation and GO networks

Dataset	Edges		
	Correlation N/W		GO N/W
	Control	AD	
GSE48350	16441	22814	6740
GSE5281	403894	364752	767620
GSE28146	52128	508	43803

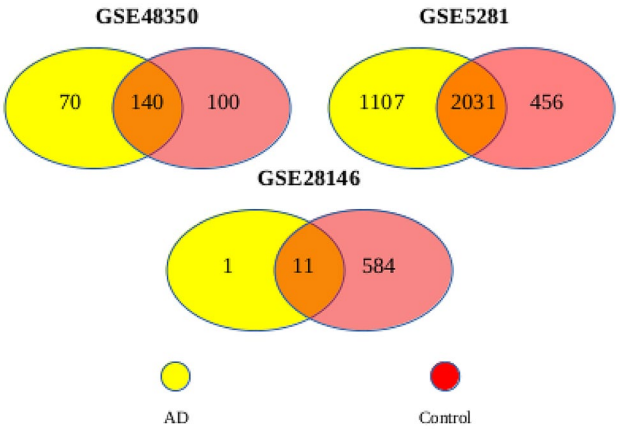

Fig. 3 Number of genes in (AD-control) and (control-AD) N/W

dataset 1, 3138 genes and 15499 edges in dataset 2, and 12 genes and 7 edges in dataset 3.

3.5 Analysis of networks

Generally, a gene of interest behaves differently in normal and affected persons. Hence, both AD and control common networks are analyzed and culled the genes present in the AD network but not in the control network (AD-CTRL). As a result, 79 such genes are identified in dataset 1, 1107 genes in dataset 2, and 1 gene in dataset 3. Similarly, 100, 456, and 584 genes in dataset 1, dataset 2 and dataset 3 are identified, which are present in the control network but not in the AD network. Figure 3 shows the Venn diagram of genes.

3.6 Gene set enrichment analysis

The gene set enrichment analysis of AD and control networks is performed using GSEA 4.0 application, which can be downloaded from http://software.broadinstitute.org/gsea (Subramanian and Tamayo 2005). The all_GENE_ONTOLOGY database is used for this analysis. In dataset 1, we found that 44 and 13 GO terms are down-regulated and up-regulated, respectively, in the AD network. In the control network of dataset 1, 99, and 20 GO terms are down-regulated and up-regulated, respectively. Similarly, in dataset 2, 298 and 148 GO terms are down-regulated and up-regulated, respectively, in the AD network. In contrast, in the control network, 307 and 321 GO terms are down-regulated and up-regulated. We found a total of 11 and 21 GO terms, which got down-regulated in both the AD networks of dataset 1 and dataset 2 and control networks of dataset 1 and dataset 2, respectively. Similarly, 8 and 16 common GO terms got up-regulated in AD and control networks of dataset 1 and dataset 2. Tables 3 and 4 list the GO terms which got up-regulated/down-regulated in AD network but not in control network and vice versa. Tables 5 and 6 list the GO terms which got down-regulated and up-regulated in the control and AD networks, respectively. All GO terms related to dataset 3 are provided in supplementary data.

Table 3 GO terms UP-regulated in AD but not in control N/W and vice versa

GO terms up-regulated in AD but not in control network	GO terms up-regulated in control but not in AD Network
	GO BIOLOGICAL ADHESION
GO ENZYME LINKED RECEPTOR PROTEIN SIGNALING PATHWAY	GO DNA BINDING TRANSCRIPTION FACTOR ACTIVITY
	GO DOUBLE STRANDED DNA BINDING
GO LOCOMOTION	GO NEGATIVE REGULATION OF TRANSCRIPTION BY RNA POLYMERASE II
	GO POSITIVE REGULATION OF RNA BIOSYNTHETIC PROCESS
GO NEGATIVE REGULATION OF RNA BIOSYNTHETIC PROCESS	GO REGULATORY REGION NUCLEIC ACID BINDING
	GO RESPONSE TO WOUNDING
	GO SEQUENCE SPECIFIC DNA BINDING
GO POSITIVE REGULATION OF LOCOMOTION	GO SEQUENCE SPECIFIC DOUBLE STRANDED DNA BINDING
	GO SKELETAL SYSTEM DEVELOPMENT
	GO TRANSCRIPTIONAL FACTOR BINDING
GO REGULATION OF CELL POPULATION PROLIFERATION	GO TRANSITION METAL ION BINDING
	GO ZINC ION BINDING

Table 4 GO terms down-regulated in AD but not in control N/W and vice versa

GO terms down-regulated in AD But not in control network	GO terms down-regulated in control But not in AD network
	GO AXON
	GO AXON PART
	GO INTRACELLULAR TRANSPORT
	GO MEMBRANE PROTEIN COMPLEX
	GO ORGANELLE LOCALIZATION
	GO POSTSYNAPSE
GO SYNAPTIC VESICLE MEMBRANE	GO SYNAPSE
	GO TRANSMEMBRANE TRANSPORT
	GO TRANSPORT VESICLE MEMBRANE
	GO VESICLE LOCALIZATION
	GO VESICLE MEDIATED TRANSPORT IN SYNAPSE

3.7 Classification

As all the genes in the combined network may be important concerning Alzheimer's disease, only the top genes are picked up using feature selection are chosen for the discussion. Although any classification method can be used, the main purpose of this stage is to see the proposed approach's effectiveness for selecting the potential candidate genes in discriminating between control and genes. To analyze whether the identified genes are able to classify the disease or not, the decision tree and random forest are used for the classification. As an input to the decision tree and random forest, the expression value of genes present in the AD network but not in the control network and genes present in the control network but not in the AD network are used. J48 decision tree and random forest are used with tenfold cross-validation. A total of 2 decision trees are constructed, one for dataset 1 and another for dataset 2. Feature selection is also performed using correlation-based feature subset selection for machine learning algorithms (Hall 2000). After performing feature selection, we got 13 genes out of 179 genes (79 + 100, Fig. 3), in dataset 1 (GSE48350), 101 genes out of 1563 (1107 + 456, Fig. 3) genes, and 54 genes out of 585 (1 + 584, Fig. 3) genes, in dataset 2 (GSE5281). Table 7 shows the accuracy obtained for each dataset.

4 Comparison

For the comparison purpose, we have considered two recently published frameworks: the first is based on Lasso and random forest (LASSO & RF) (Sharma and Dey 2021), and the second is based on *t*-test, genetic algorithm, and a modified particle swarm optimization algorithm (MPSO) (Ramaswamy 2021). For a fair comparison, if the number of genes obtained by the frameworks is more than 20, we chose only the top 20 genes for the comparison. Table 8 shows the top genes obtained from the different frameworks for different datasets. Tables 9 and 10 list all the identified genes in dataset 1 and dataset 2, respectively. Genes selected for dataset 3 are provided in supplementary data.

As observed from Table 8, though the significant genes obtained after feature selection are almost different for all the datasets, yet the accuracy of the genes acquired is nearly the same in all datasets (Table 7). Hence, this does not provide us with any inference. Therefore, we compared the degree of specificity of genes obtained by the proposed framework, LAASO & RF and MPSO, towards Alzheimer's disease. We checked the direct interactions of the genes obtained with the AD pathway genes using the STRING database. We did not find any common pattern in the number of interactions, making it difficult to draw any conclusion. We further used DAVID[4] to obtain the diseases of the genes obtained after feature selection which did not yield significant results as the number of genes is less, and some are not characterized. It is well known that interacting proteins regulate the function of a protein (Swamy 2021). Therefore, retrieving the interacting partners and the associated diseases can give us a deeper insight into the genes obtained from our framework. HIPPIE[5] is used to fetch the high confidence primary interacting proteins of the genes obtained from our analysis. The primary interacting genes are then subjected to DAVID analysis to obtain the corresponding diseases.

It is observed that in dataset 1 and dataset 2, primary interactions of the genes obtained by the proposed framework are directly associated with Alzheimer's disease with high significance. In contrast, the interacting partners of genes obtained from other algorithms are not at all related

[4] https://david.ncifcrf.gov/.

[5] http://cbdm-01.zdv.uni-mainz.de/mschaefer/hippie/.

Table 5 Down-regulated and up-regulated GO terms in control network

Down-regulated			Up-regulated		
GO term	No. of genes in dataset 1	No. of genes in dataset 2	GO term	No. of genes in dataset 1	No. of genes in dataset 2
GO AXON	48	150	GO BIOLOGICAL ADHESION	32	260
GO AXON PART	31	93	GO CELL MOTILITY	40	307
GO CYTOPLASMIC VESICLE PART	39	306	GO DNA BINDING TRANSCRIPTION FACTOR ACTIVITY	32	392
GO DISTAL AXON	25	67	GO DOUBLE STRANDED DNA BINDING	19	250
GO EXOCYTIC VESICLE	23	52	GO NEGATIVE REGULATION OF TRANSCRIPTION BY RNA POLYMERASE II	18	256
GO EXOCYTOSIS	26	190	GO POSITIVE REGULATION OF RNA BIOSYNTHETIC PROCESS	41	437
GO INTRACELLULAR TRANSPORT	49	384	GO POSITIVE REGULATION OF TRANSCRIPTION BY RNA POLYMERASE II	26	335
GO MEMBRANE PROTEIN COMPLEX	33	217	GO REGULATORY REGION NUCLEIC ACID BINDING	22	275
GO NEURON PROJECTION TERMINUS	15	35	GO RESPONSE TO WOUNDING	17	136
GO ORGANELLE LOCALIZATION	29	161	GO SEQUENCE SPECIFIC DNA BINDING	21	296
GO POSTSYNAPSE	38	164	GO SEQUENCE SPECIFIC DOUBLE STRANDED DNA BINDING	18	236
GO PRESYNAPSE	42	121	GO SKELETAL SYSTEM DEVELOPMENT	16	105
GO SECRETORY VESICLE	31	183	GO TRANSCRIPTION FACTOR BINDING	17	225
GO SYNAPSE	72	290			
GO SYNAPSE PART	67	235	GO TRANSCRIPTION FACTOR BINDING	17	225
GO TRANSMEMBRANE TRANSPORT	49	246			
GO TRANSPORT VESICLE	26	27	GO TRANSITION METAL ION BINDING	15	171
GO TRANSPORT VESICLE MEMBRANE	21	46			
GO VESICLE LOCALIZATION	17	83	GO TUBE DEVELOPMENT	24	239
GO VESICLE MEDIATED TRANSPORT IN SYNAPSE	16	54			
GO WHOLE MEMBRANE	49	28	GO ZINC ION BINDING	15	144

GO terms which got down-regulated and up-regulated in control network with p value ≤ 0.05

to any neurological disorders. Although in dataset 3, genes obtained from the proposed framework, LASSO & RF, and MPSO framework have interacting partners implicated in Alzheimer's disease. However, it is interesting to note that the significance and count of genes associated with AD in the proposed framework are quite high compared to the LASSO & RF and MPSO framework. The supplementary data provide the table of all the diseases related to the genes, the gene count, and their corresponding p values.

5 Results

Using the introduced framework, we are able to identify genes in all datasets that are directly or indirectly related to AD with a high classification power. More than 95% accuracy is achieved for classifying the disease and control using the identified genes. Tables 9 and 10 list the genes identified. The link between the identified genes and the AD pathway genes is analyzed to find out the importance of the identified

Table 6 Down-regulated and up-regulated GO terms in AD network

Down-regulated			Up-regulated		
GO term	No. of genes in dataset 1	No. of genes in dataset 2	GO term	No. of genes in dataset 1	No. of genes in dataset 2
GO CYTOPLASMIC VESICLE PART	46	385	GO CELL MOTILITY	33	401
GO DISTAL AXON	29	91	GO ENZYME LINKED RECEPTOR PROTEIN SIGNALING PATHWAY	21	305
GO EXOCYTIC VESICLE	30	64			
GO EXOCYTOSIS	25	234	GO LOCOMOTION	38	461
GO NEURON PROJECTION TERMINUS	17	46	GO NEGATIVE REGULATION OF RNA BIOSYNTHETIC PROCESS	16	416
GO PRESYNAPSE	51	149			
GO SECRETORY VESICLE	3 5	233	GO POSITIVE REGULATION OF LOCOMOTION	16	152
GO SYNAPSE PART	83	289			
GO SYNAPTIC VESICLE MEMBRANE	22	33	GO POSITIVE REGULATION OF TRANSCRIPTION BY RNA POLYMERASE II	19	413
GO TRANSPORT VESICLE	33	112	GO REGULATION OF CELL POPULATION PROLIFERATION	24	427
GO WHOLE MEMBRANE	49	430	GO TUBE DEVELOPMENT	20	277

GO terms which got down-regulated and up-regulated in AD network with p value ≤ 0.05

Table 7 Accuracy obtained using different decision trees

Framework	Dataset 1		Dataset 2		Dataset 3	
	Decision tree	Random forest	Decision tree	Random forest	Decision tree	Random forest
Proposed framework	97.73%	100%	95.65%	100%	63.33%	96.67%
LASSO & RF	97.73%	100%	95.65%	100%	83.33%	96.67%
MPSO	97.73%	100%	95.65%	100%	70%	86.67%

Table 8 Different genes selected by proposed algorithm

Framework	Dataset1	Dataset2	Dataset3
Proposed	ATP2B3, FGF12, MDFIC, NSG1, TAC1, ZNF621, BTK, CD44, CD5, DACH1, ERBB4, RTF1, TAB3	ABI2, ELAVL3, AP2A2, CEP97, ADGRB3, SRRM2, AGFG1, SEC22C, EAPP, AKAP13, TNRC6B, ARHGAP21, CHMP2A, BICD1, FAM120A, COPG1, YTHDC1, INTS3, ERC1, BRD9	ARL8B, PMAIP1, THRB, BHLHE40, ZNF711, BNIP3, DIS3, ZMYM2, HNRNPA0, MAPKAPK2, KPNA6, KBTBD7, MAP2K6, AHNAK, CD44, IL1R1, LRP8, NCOA2, CDH5, ZBTB17
Lasso and Random Forest	ANKIB1, FBRSL1, LOC101927151, RAE1, RTF1, SLC25A46, ZNF621	ARHGAP5, CDK5RAP2, CKMT1A, CKMT1B, DUSP8, FAM120A, FAM168A, FAM63A, KTN1, LOC101927562, OSBPL1A, PEBP1, RHOB, TESK1, ZNF532	BNIP3, CD44, HPS3, MCCC1, NSUN6, ST6GALNAC5
MPSO	ZNF621, LOC101927151, SLC25A46, ANKIB1, RAE1, RTF1	ANKRD12, ELAVL3, ERC1, GPR155, KTN1, NAV1	IL13RA1, DEFB125, RFX4, CXorf38, JAM3, ZFP41, TGFB1I1, TTL

genes in this work. As a result, it is found that most genes have either direct or one-hop interaction with the AD pathway genes. Table 11 shows some direct interactions between top genes of dataset 1 and AD pathway genes. As the top genes in both datasets are different, we tried to determine the relationship between both datasets' top genes. STRING

Table 9 Genes identified in dataset 1 (GSE48350)

Gene symbol	*p* value	logFC
ATP2B3	0.000131	−0.9364480
BTK*	1.05e-05	−0.8254824
CD44*	0.000118	1.04783081
CD5	0.00123	0.87985092
DACH1	2.8e-05	0.94417236
ERBB4*	0.00074	0.82142565
FGF12	0.00134	−1.1104277
MDFIC	0.000417	0.8739675
NSG1*	0.001075	−1.05723075
TAB3	0.00173	1.00097339
RTF1	1.46e-17	−1.7616841
TAC1*	0.0139	1.32196006
ZNF621	2.32e-24	2.96528119

*Identified in literature

database[6] is used to find interactions between the genes, only interactions that are experimentally proven or are from the curated database with at least medium confidence value 0.4 (as mentioned in STRING database) are considered. All the top genes of dataset 1 have either direct or one-hop connections with at least one top gene of dataset 2 (a few of the interactions are shown in Table 12). We also checked the GO similarity between the top genes of both datasets and the GO similarity of top genes with the AD pathway genes to find the similarity between them. Also checked the primary interactions of the identified genes and found them related to AD with high significance compared to the genes identified by other considered frameworks. All the interactions, GO similarity, disease-associated, and primary interactions files can be downloaded from *"Supplementary Data"*.

In dataset 1, out of 13 identified genes, 5 (BTK, CD44, ERBB4, NSG1, and TAC1) are found to be related to AD in the recent literature. Similarly, many genes (ADAM22, AGFG1, GRIN2B, MPRIP, ZNF532, etc.), identified in dataset 2 are listed in the literature. Gene ATP2B3[7] has human phenotype ontology of ataxia, cerebellar atrophy, cerebellar hypoplasia, and clumsiness. Gene FGF12 has a human phenotype ontology of abnormal myelination, abnormality of vision, and absence of speech. In Keaney (2019) observed that the activation of phospholipase gamma 2, a genetic risk factor in AD, is decreased due to the blockade of **BTK**. Pinner (2017) investigated the expression values of **CD44** splice variants in the hippocampus region of AD patients and compared it with the control patients and observed that the expression values of splice variants of CD44 are significantly higher in AD patients when compared to the normal person. The research suggested that some splice variants of CD44 contribute to AD pathology. Woo (2011) found that up-regulation of the immunoreactivity of **ERBB4** may involve in Alzheimer's disease progression. Abhik Ray and Gerecke (2003) observed that Neuregulin-1 and **ERBB4** immunoreactivity is associated with plaques formation in the AD brain. Norstrom (2010) Norstrom et al. reported that NEEP21 protein (gene name: **NSG1**) affects the processing of APP and Aβ production. Magistri (2015) analyzed that in the hippocampus region of the brain in AD patients, TAC1 is down-regulated compared to controls hippocampus.

The GSEA analysis shows that out of 22, 12 GO terms that got down-regulated in the control network are not being regulated in the AD network and vice versa, which indicates that there may be a disturbance in the regulation of those 12 GO terms. Similarly, out of 21, 18 GO terms that got up-regulated in the control network is not being regulated in the AD network vice versa. Tables 3 and 4 list the GO categories which may got disturbed. In the identified GO terms, we find that some are found to be disturbed in the Alzheimer's disease, like, GO SYNAPTIC VESICLE MEMBRANE, GO AXON, GO TRANSPORT VESICLE MEMBRANE, GO VESICLE MEDIATED TRANSPORT IN SYNAPSE, GO NEGATIVE REGULATION OF RNA BIOSYNTHETIC PROCESS, GO REGULATION OF CELL POPULATION PROLIFERATION, GO NEGATIVE REGULATION OF TRANSCRIPTION BY RNA POLYMERASE II, GO RESPONSE TO WOUNDING, GO SKELETAL SYSTEM DEVELOPMENT, GO POSITIVE REGULATION OF RNA BIOSYNTHETIC PROCESS, and GO ZINC ION BINDING. Blennow and Bogdanovic (1996) found that the level of **synaptic vesicle membrane** protein rab3a was reduced in Alzheimer's disease in the hippocampus. In the studies, it is found that in Alzheimer's disease, the amyloid-beta disturbed the **vesicle transport in synapse** in the hippocampus (Seifert and Eckenstaler 2016; Kelly and Ferreira 2007). Wu and Zhang (2016) observed that the **cell proliferation** gets slowdown when the APP is overexpressed. Watt (2010) discussed the role of **Zinc** in Alzheimer's disease. Zinc binds to amyloid-beta, advancing its conglomeration into neurotoxic species, and disturbance of zinc homeostasis in the brain results in synaptic and memory deficiencies. Kiecolt-Glaser and Marucha (1995) observed that **wound healing** took a long time significantly in AD patients than in controls. Chen and Lo (2017) conclude that AD increase the risk of osteoporosis (**Skelton disorder**). The overexpression of amyloid-beta might happen in both cerebrum and bone, meddling with the RANKL signalling cascade, improving osteoclast activities, and prompting osteoporosis.

[6] https://string-db.org/.

[7] https://www.genecards.org.

Table 10 Genes identified in Dataset 2 (GSE5281)

Gene symbol	*P* value	logFC	Gene symbol	*p* value	logFC
ABI2	0.0067832	0.961026636667	INTS3	2.79E-06	−1.37560425
ACBD5	1.25E-06	2.3668954	IPO7	0.000274	0.85463051
ACO1	0.000295	−0.98042327	JPH3	0.000359	2.11788507
AGFG1	9.36E-06	2.19387208	KDM5A	0.00524225	1.9725445
ACTR1B	0.000606	−0.97274028	KIF1A	2.31E-05	1.27250433
ACVR1B	0.000422	1.39831289	KTN1	4E-10	2.34921503
ADAM22	0.0216053333333	0.53768401	L1CAM	4.46E-07	1.83246955
ADAM23	4.08E-06	−1.73163989	LAMP1	0.0038554085	−1.73519342
ADCY2	2.95E-05	1.31786542	MAGI2	3.35E-08	3.020489
AKAP13	0.00647204185714	1.04407532714	MAP6	6.54E-07	1.89564215
AKAP8L	0.000989	1.87231609	MARK3	0.0001442	−0.04287374
ANK3	0.000124	2.44887394	MIB1	3.95E-09	2.14433497
AP2A2	0.000291433333333	-1.11659461	MORF4L2	0.001740079	2.404579215
AP3D1	0.00770605	1.39260274	MRPS5	0.000319606666667	2.40202342333
ADGRB3	6.93E-07	2.46317656	NAP1L4	2.12E-05	−1.09098646
ANKRD11	0.00421	1.24675	NEUROD1	4.28E-05	1.56594988
ARHGAP21	4.55E-08	2.64329836	CBX3	0.00117	2.43672
BBX	0.003446645924	1.888399814	NR2F2///NR2F1	5.13E-05	1.37623146
BICD1	7.65E-07	1.88214093	NSL1	0.014850297	1.450132955
BNIP3L	0.000345	0.94418923	NUCKS1	2.05365E-06	0.11134135
BRD9	4.94E-06	−1.80179381	PNISR	0.000345876666667	1.60721201
C12orf10	0.00169	−0.82914204	PRKAB2	0.0002555	1.431380315
			PTBP3	4.61E-05	1.91441374
CAMSAP2	0.00028575	0.050696345	PTP4A1	0.005201145	0.24193515
CAPRIN2	4.8E-06	−1.48271188	PTPRJ	4.73E-06	−1.34847657
CBL	1.29E-06	1.32748593	REV3L	1.08E-06	1.33991121
CEP97	0.00550010933333	1.99005593	RFK	4.39E-05	1.1322082
CHMP2A	5.25E-05	−1.91874975	KDELR2	3.08E-05	−1.07980691
CLN8	6.19E-06	1.24748557	SEC22C	9.851395E-05	1.871307925
COPG1	3.09E-08	−2.10125131	SLC25A36	0.010800795	1.20768185
CORO1C	1.01E-05	−1.35341065	SLC8A1	8.59E-08	1.66146541
CTSC	0.005502715	1.421647035	SRRM2	0.0028989782	2.03190842
DGKG	1.35E-05	2.67416378	STOML2	6.49E-08	−2.13214491
EAPP	1.16E-08	−1.71677758	SUZ12P1///SUZ12	0.000103	0.8732712
EIF5B	0.00467275	1.32003578667	TBL1XR1	1.155237E-07	−0.45937598
ELAVL3	1.33E-10	2.89606404	TNPO2	9.1E-07	1.7570126
ELMO1	3.08E-06	−2.49796439	TNRC6B	2.17015266667E-05	1.56078730333
ERC1	4.31E-10	2.47720499	TRIM23	1.72E-05	1.39388483
ERCC3	2.21E-06	−1.46854842	MNT	5.07E-05	1.43479172
MICAL1	0.0234	1.22246841	PABPC3	3.48E-06	1.14393254
ESF1	0.009050321	1.38541814	UNKL	3.21E-05	1.08877527
FAM120A	0.000426554166667	1.51672673333	USP10	0.000656245	1.216217585
UHMK1	1.07E-06	1.8235725	WDR82	6.58E-06	3.907264
ZNF532	1.66E-09	−2.46526681	YTHDC1	0.00066	1.7781
GALNT1	0.00010224	1.28699208	ZBTB1	0.00018795	2.059890885
GLG1	9.93E-09	1.55542645	ZMAT3	9.7E-06	1.33505671
GOLGA2	0.00525002385	1.82953746	ZNF148	0.000365265	1.22672275
GRIN2B	0.0004368415	1.678765335	ZNF264	4.16E-06	1.46456226
GRK3	1.63E-09	2.31565161	ZNF652	0.0006725	1.62317446
HSPH1	5.07E-06	1.55584984	ZNF770	7.36E-05	1.06839715

Table 10 (continued)

Gene symbol	*P* value	logFC	Gene symbol	*p* value	logFC
INO80D	0.0056500398	0.291794275	MPRIP	2.1775E-05	−2.412150

Table 11 STRING interactions between top genes of dataset 1 and AD pathway genes

Top genes of dataset 1	AD pathway genes	STRING interaction score
ATP2B3	CALM1	0.69
BTK	FAS	0.935
ERBB4	PSEN1	0.9
FGF12	CALM1	0.96
TAB3	TNF	0.902
TAC1	APP	0.9

Table 12 STRING interactions between top genes of dataset 1 and data set 2

Top genes of dataset 1	Top genes of dataset1	STRING interaction score
BTK	CBL	0.95
CD44	ANK3	0.8
ERBB4	GRIN2B	0.9
ATP2B3	CALM1	0.69
CALM1	ADCY2	0.64
CD5	CD4	0.861
CD4	AGFG1	0.9
DACH1	NCOR1	0.426
NCOR1	TBL1XR1	0.98
FGF12	CALM1	0.96
CALM1	ADCY2	0.64
MDFIC	CTNNB1	0.9
CTNNB1	TBL1XR1	0.935
FGF12	SRPK2	0.442
SRPK2	SRRM2	0.442
TAB3	MAP3K7	0.986
MAP3K7	PRKAB2	0.817
RTF1	SUPT16H	0.995
SUPT16H	ERCC3	0.9
TAC1	TACR1	0.965
TACR1	AGFG1	0.9
ZNF621	TRIM28	0.922
TRIM28	ZNF770	0.902

6 Conclusions

In summary, in this paper, a framework that includes t-test, correlation, GO categories, and machine learning techniques is developed to identify the potential biomarkers for Alzheimer's disease. The GO categories are analyzed and used to create a more biologically significant network, which helps in eliminating false correlations. Feature selection is used to list out the top genes. Then, using the J48 decision tree and random forest, their classification power is estimated and obtained more than 95% accuracy for all the datasets. Biological interactions between the top genes of all datasets are studied in which the top genes either have direct or one-hop experimentally proven interactions with one another. Biological interactions between top genes and AD pathway genes are also studied. As a result, many of the genes were found to have direct experimentally proven interactions with the AD pathway genes. Primary interactions of selected genes show that the genes selected by the proposed framework are associated with Alzheimer's disease. Gene set enrichment analysis of AD and control networks is also carried out and found that GO terms which got up-regulated/down-regulated in AD network but not in control network and vice versa, may get disturbed in Alzheimer's disease. The literature shows that the genes identified by the decision tree classifier whose logFC values indicate that these genes that need to be up-regulated are down-regulated and vice versa. The results consist of the genes and GO terms that are related to Alzheimer's disease in the literature, which adds more credibility to the results. The results show that even though the classification power of genes identified by other frameworks are high or the same, the genes identified by the proposed framework have a high degree of association with AD in comparison to the genes identified by the other frameworks considered. In future, the proposed framework can be applied to other diseases too, and an automated tool based on the proposed framework can be developed.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13721-021-00349-9.

References

Akoglu H (2018) User's guide to correlation coefficients. Turk J Emerg Med 18(3):91–93

AL-Dlaeen D, Alashqur A(2014) Using decision tree classification to assist in the prediction of Alzheimer's disease. 6th International conference on computer science and information technology (CSIT), Amman, pp. 122-126

Alzheimer's Disease Fact Sheet, National Institute of Aging, U.S. Department of Health and Human Services. 2015. https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet

Anand KS, Dhikav V (2012) Hippocampus in health and disease: an overview. Ann Indian Acad Neurol 15(4):239–46
Ashburner M et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25(1):25–9
Blennow K, Bogdanovic N et al (1996) Synaptic pathology in Alzheimer's disease: relation to severity of dementia, but not to senile plaques, neurofibrillary tangles, or the ApoE4 allele. J Neural Transm 103(5):603–618
Chaudhury AR, Gerecke KM et al (2003) Neuregulin-1 and ErbB4 immunoreactivity is associated with neuritic plaques in Alzheimer disease brain and in a transgenic model of Alzheimer disease. J Neuropathol Exp Neurol 62(1):42–54
Chen Y-H, Lo RY (2017) Alzheimer's disease and osteoporosis. Ci Ji Yi Xue za Zhi (Tzu-chi Med J) 29(3):138–142
Cheng J, Liu HP et al (2021) Machine learning compensates fold-change method and highlights oxidative phosphorylation in the brain transcriptome of Alzheimer's disease. Sci Rep 11:13704
Crow M, Lim N et al (2019) Predictability of human differential gene expression. Proc Natl Acad Sci 116(13):6491–6500
Duff MC, Covington NV et al (2020) Semantic memory and the hippocampus: revisiting, reaffirming, and extending the reach of their critical relationship. Front Hum Neurosci 13:471
Hall M (2000) Correlation-based feature selection for machine learning. Dep Comput Sci 19
Hu R-T, Yu Q et al (2020) Co-expression network analysis reveals novel genes underlying Alzheimer's disease pathogenesis. Front Aging Neurosci 12:432
Huang DW et al (2007) The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 8(9):R183
Keaney J et al (2019) Inhibition of Bruton's tyrosine kinase modulates microglial phagocytosis: therapeutic implications for Alzheimer's disease. J Neuroimmune Pharmacol Off J Soc NeuroImmune Pharmacol 14(3):448–461
Kelly BL, Ferreira A (2007) $A\beta $ disrupted synaptic vesicle endocytosis in cultured hippocampal neurons. Neuroscience 147(1):60–70
Kiecolt-Glaser JK, Marucha PT et al (1995) Slowing of wound healing by psychological stress. Lancet 346(8984):1194–1196
Kuang J, Zhang P, Cai T et al (2021) Prediction of transition from mild cognitive impairment to Alzheimer's disease based on a logistic regression-artificial neural network-decision tree model. Geriatr Gerontol Int 21(1):43–47
Lanoiselée H-M et al (2017) APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: a genetic screening study of familial and sporadic cases. PLoS Med 14(3):e1002270
Magistri M et al (2015) Transcriptomics profiling of Alzheimer's disease reveal neurovascular defects, altered amyloid-$\beta $ homeostasis, and deregulated expression of long noncoding RNAs. J Alzheimer's Dis JAD 48(3):647–665
Martin D, Brun C, Remy E et al (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biol 5:R101
Mukaka MM (2012) Statistics corner: a guide to appropriate use of correlation coefficient in medical research. Malawi Med J J Med Assoc Malawi 24(3):69–71
Nishiwaki K, Kanamori K, Ohwada H (2016) Finding a disease-related gene from microarray data using random forest. IEEE 15th international conference on cognitive informatics and cognitive computing (ICCICC), Palo Alto, pp. 542-546
Norstrom EM et al (2010) Identification of NEEP21 as a $\beta $-amyloid precursor protein-interacting protein in vivo that modulates amyloidogenic processing in vitro. J Neurosci Off J Soc Neurosci 30(46):15677–15685
Pinner E et al (2017) CD44 splice variants as potential players in Alzheimer's disease pathology. J Alzheimer's Dis JAD 58(4):1137–1149
Ramaswamy R et al (2021) Feature selection for Alzheimer's gene expression data using modified binary particle swarm optimization. IETE J Res 2021:1–12
Ray M, Zhang W (2010) Analysis of Alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks. BMC Syst Biol 4:136
Ray M, Yunis R, Chen X, Rocke DM (2012) Comparison of low and high dose ionizing radiation using topological analysis of gene co-expression networks. BMC Genomics 13(1):190
Ray S et al (2017) A comprehensive analysis on preservation patterns of gene co-expression networks during Alzheimer's disease progression. BMC Bioinform 18(1):579
Ruan J, Zhang W (2006) Identification and evaluation of functional modules in gene co-expression networks. Syst Biol Comput Proteomics Lecture Notes Comput Sci 4532(1):57–76
Saputra RA et al (2020) Detecting Alzheimer's disease by the decision tree methods based on particle swarm optimization. J Phys Conf Ser 1641:012025
Seifert B, Eckenstaler R et al (2016) Amyloid-beta induced changes in vesicular transport of BDNF in hippocampal neurons. Neural Plast 2016:4145708
Sharma A, Dey P (2021) A machine learning approach to unmask novel gene signatures and prediction of Alzheimer's disease within different brain regions. Genomics 113(4):1778–1789
Subramanian A, Tamayo P et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci 102(43):15545–15550
Swamy K et al (2021) Protein complexes form a basis for complex hybrid incompatibility. Front Genet 12:144
Takahiro K, Kazutaka N, Hayato O (2016) Finding unknown disease-related genes by comparing random forest results to secondary data in medical science study. Proceedings of the 7th international conference on computational systems-biology and bioinformatics (CSBio '16), pp. 24-27
Watt NT et al (2010) The role of Zinc in Alzheimer's disease. Int J Alzheimer's Dis 2011:971021
Woo R-S et al (2011) Expression of ErbB4 in the neurons of Alzheimer's disease brain and APP/PS1 mice, a model of Alzheimer's disease. Anat Cell Biol 44(2):116–27
Wu Y, Zhang S et al (2016) Regulation of global gene expression and cell proliferation by APP. Sci Rep 6:22460
Xia J, Rocke DM, Perry G, Ray M (2014) Differential network analyses of Alzheimer's disease identify early events in Alzheimer's disease pathology. Int J Alzheimer's Dis. https://doi.org/10.1155/2014/721453
Zhao C, Wang Z (2018) GOGO: an improved algorithm to measure the semantic similarity between gene ontology terms. Sci Rep 8:15107
Zhou S (2008) Probability theory and mathematical statistics, 4th edn. Higher Education Press, Beijing
Zhu G, Yang P (2016) Identifying the candidate genes for Alzheimer's disease based on the rejection region of T-test. International conference on machine learning and cybernetics (ICMLC), Jeju, vol. 2, pp. 732–736