Establishment and Analysis of a Combined Diagnostic Model of Alzheimer's Disease With Random Forest and Artificial Neural Network

Sun, Dazhong; Peng, Haojun; Wu, Zhibing . Frontiers in Aging Neuroscience ; Lausanne (Jun 30, 2022).

ABSTRACT (ENGLISH)

Alzheimer's disease (AD) is a chronic neurodegenerative disease. Due to the limitations of existing diagnostic techniques for AD, it is necessary to develop novel diagnostic models to supplement existing methods. Few studies, however, have attempted to develop a diagnostic model based on gene biomarkers. To identify gene biomarkers and construct a diagnostic model, we used a computational method that combined two machine learning algorithms, including random forest (RF) and artificial neural network (ANN). We collected AD gene expression data from Gene Expression Omnibus (GEO) database; four datasets were utilized, including two training dataset for screening differentially expressed genes (DEGs) and two validation datasets. Firstly, based on RF, 6 key genes(NFKBIA, SST, KLF15, NDUFA7, MAFF, and UCHL1) in 177 key DEGs were identified to be vital for classification of AD and normal samples. The weights of these key genes were calculated and the diagnostic models were developed using ANN. Finally, two validation datasets were used to test and compare the performance by area under curve (AUC). Our model achieved an AUC of 0.822 in GSE109887, and 0.814 in GSE132903. To conclude, we uncovered gene biomarkers and successfully constructed a new diagnostic model of AD using an artificial neural network and verified its diagnostic efficacy in public datasets, which would be helpful for diagnosis.

FULL TEXT

Introduction

Alzheimer's disease (AD) is a type of chronic degenerative brain illness marked by central nervous system disorder that primarily affects people in their forties and fifties (Scheltens et al., 2021). The main clinical feature of AD is memory impairment, which may be accompanied by aphasia and personality behavior changes (Scheltens et al., 2016). Pathophysiological changes in AD may begin years before any clinical symptoms appear and may progress all the way to severe cognitive impairment (Aisen et al., 2017). As a result, AD cannot be identified just on the basis of clinical characteristics, and researchers have made exhaustive efforts to identify AD using clinical and biomarker data (Delaby et al., 2022). Understanding of AD has grown significantly over the past few decades while also highlighting the disease's complexity (Chen, 2018). Imaging technologies, cognitive level identification, and various fluid biomarkers are now used to diagnose AD (Reitz, 2015; Blennow and Zetterberg, 2018; Sun et al., 2018). It is becoming more apparent that AD is a disease with a complex regulatory network that is becoming increasingly complex (Veitch et al., 2019). As a result, more precise diagnostic and treatment targets for AD are urgently needed. The rapid advancement of microarray and high-throughput sequencing technologies in the last decade has suggested a reliable and widespread method for decoding inherited and epigenetic determinants of disease. At the same time, it also provides a lot of evidence for the diagnosis and treatment of various diseases (Kulasingam and Diamandis, 2008). Although genetic risk markers have been identified that can be used to predict and diagnose AD, their power may be limited because of the complexity of the genetic structure (Zhu et al., 2020). In diagnostic models, the use of multiple biomarkers has been shown to improve success rates significantly (Vilhjálmsson et al., 2015). In recent years, the primary difficulty in constructing a classification model based on

gene expression data has been choosing the most significant index or feature for classification. This problem can be solved using a variety of machine learning techniques (Kursa, 2014; Tian et al., 2020; Xie et al., 2020). These algorithms have made significant contributions to the classification of gene expression data, disease detection, cell migration, and microbiome research when used alone or in combination (Hsieh et al., 2011; Kong and Yu, 2018; Zhang et al., 2018; Janßen et al., 2019).

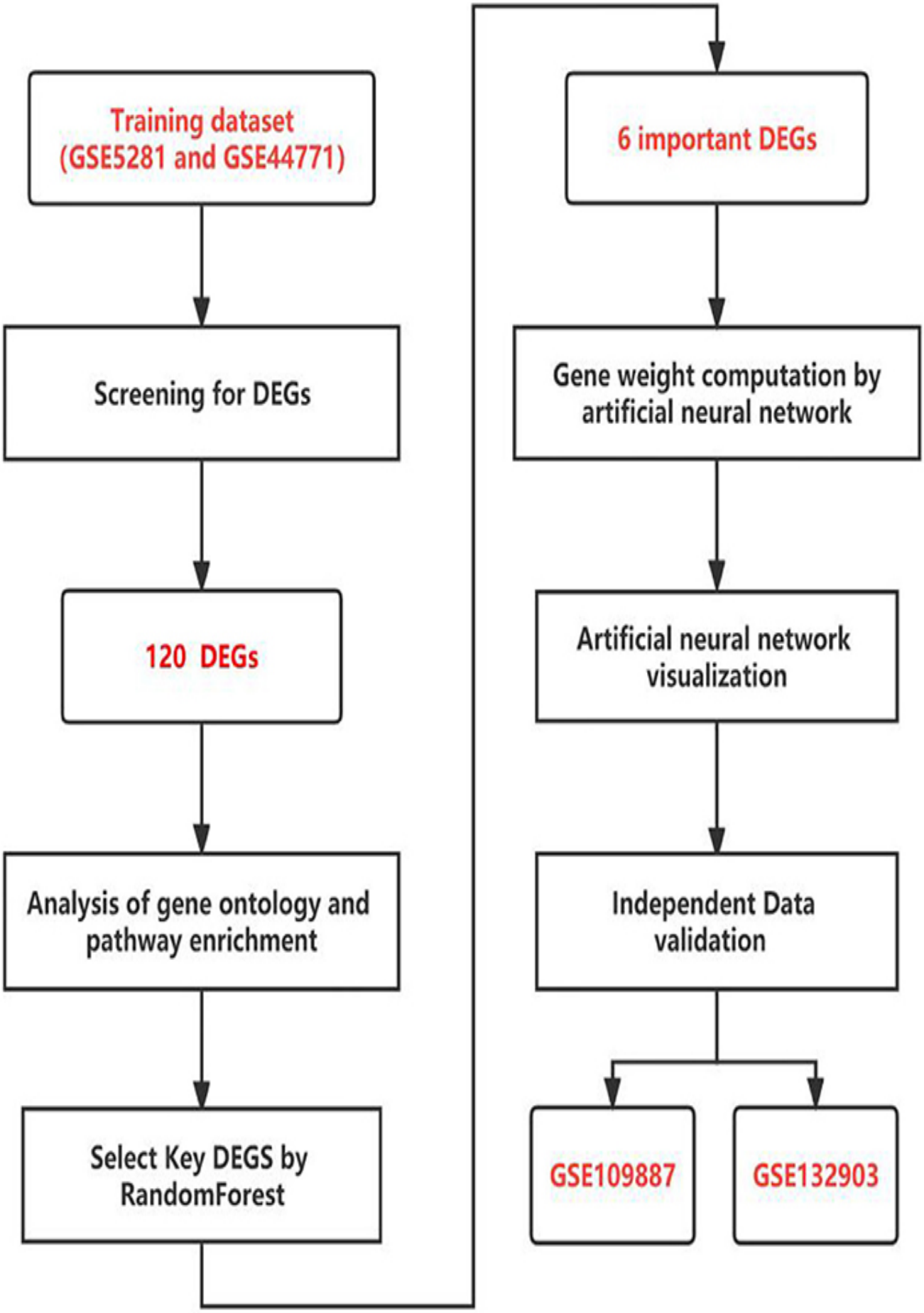
Using the key genes screened from datasets in the GEO database, we created an AD diagnosis model. It was first determined which genes were most important for AD classification using RF. A genetic diagnostic model for AD was then built using these key genes by artificial neural networks. We evaluated the performance of the diagnosis model with independent validation datasets to confirm its accuracy and performance.

Materials and Methods

Study Design

For the differentially expressed genes (DEGs) screening, the GSE5281 dataset was merged with the GSE44771 dataset as the training dataset (step 1). We went on to analyse gene ontology and pathway enrichment (step 2). Then, we screened the key genes using RF classification (step 3). Following the computation of gene weights (step 4), an ANN model was developed (step 5). In the end, GSE109887 and GSE132903 datasets were used to conduct further validation (step 6). All statistics are computed by R software version 4.1.3. Figure 1 depicts the entire research flow.

FIGURE 1



Enlarge this image.

Data Selection and Processing

Datasets in this study were obtained from the GEO database, which stores information about how genes are

expressed using high-throughput methods. It was created by the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi). The keywords “AD, normal” or “AD, health” were used in this study to conduct a broad search through the NCBI database platform. The type of datasets we chose was expression profiling by array, and the type of organisms was homo sapiens. The sample size of the dataset is greater than 60. We used the ComBat function in R package sva (Varma, 2020) to remove the batch effect of data from different platforms. The log2-transformed quantile-normalized signal intensity of these datasets was rectified, and the corrected results were outputted.

Screening for DEGs

Using traditional Bayesian data analysis, the R package limma (Ritchie et al., 2015) was utilized to screen DEGs of the training dataset. Adjusted *P* values less than 0.05 and logFoldChang (logFC) greater than 1 were established as the significance criteria for DEGs. The DEGs heatmap was created using the R package pheatmap. The volcano plot was created using the R package ggplot2 (Ito and Murphy, 2013).

Analysis of Gene Ontology and Pathway Enrichment

Gene ontology and pathway analysis are utilized for the purpose of interpreting gene expression data. An online comprehensive gene set enrichment web tool, EnrichR (https://maayanlab.cloud/EnrichR), was used in our study to conduct gene ontology and pathway enrichment analyses. Gene ontology, including biological processes, cellular components, and molecular functions, was analyzed using EnrichR. In addition, we used KEGG pathway 2021, WikiPathways 2021, and Retcome 2016 as classification sources for pathways to identify gene common pathways. EnrichR used the logarithm of the *P*-value and the z-score to create a combined score. We ranked them in order of the combined score and showed them in bar charts.

Random Forest Screening for Key Genes

We screened the key genes using random forest by R package random Forest (R project, 2022). In order to determine the lowest error rate and best stability tree number as the optimal parameter, each error rate for 1–200 trees was calculated. After that, a random forest was used to screen key genes, and the Gini coefficient method was used to calculate the dimensional significance value. The AD key genes for ANN model development were selected from the top 30 DEGs with a significance value greater than 6. The key genes in the training dataset were put into new groups based on their unsupervised hierarchical clusters, and the heatmap was generated using the R package pheatmap (Hu, 2021).

Artificial Neural Network for Building an AD Classification Model

First, the DEG expression data was converted to a Gene Score table based on the expression level. A comparison was made between the median of all sample expression values and the expression value of a single gene in a given sample. If the expression value of the up-regulated gene is greater than 0, it will be given a 1; otherwise, it will be given a 0. Likewise, if a down-regulated gene's expression value is higher, it will be given a value of 0; otherwise, it will be given a value of 1. AD was the outcome variable, and cases were assigned a 1 while controls were assigned a 0. The R package neuralnet (Beck, 2018) was used to create an ANN model based on the Gene Score table we constructed. The model parameter was set to 5 hidden layers. R package Caret (Nachid and Boussiala, 2021) was used to calculate 5-fold cross-validation of the ANN model in order to optimize the model and reduce overfitting. The confusion matrix function calculated the accuracy of the results. Using the R package pROC (Robin et al., 2011), we calculated the areas under the receiver operating characteristic curve (AUC).

Verification Using Validation Datasets

On two separate validation datasets (GSE109887 and GSE132903), the ANN model was tested for effectiveness verification. The AUC was calculated using the R package pROC.

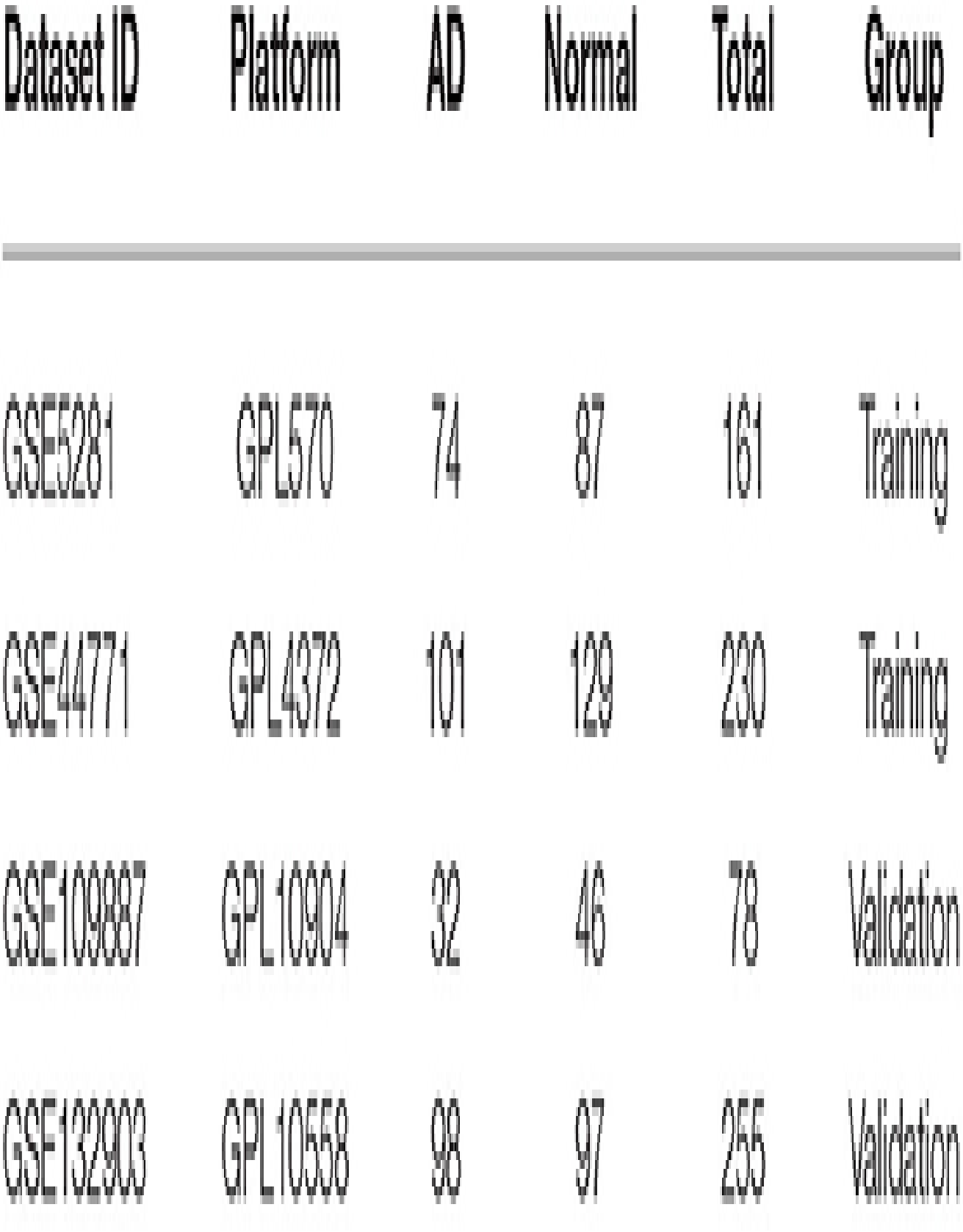
Results

Identification of DEGs

GSE5281 was a dataset including 74 AD samples and 87 control samples. Brain samples were collected from three Alzheimer's Disease Centers. Gene expression was analyzed using Affymetrix U133 Plus 2.0. GSE44771 was a dataset including 101 AD samples and 129 control samples. Brain samples were collected through the Harvard

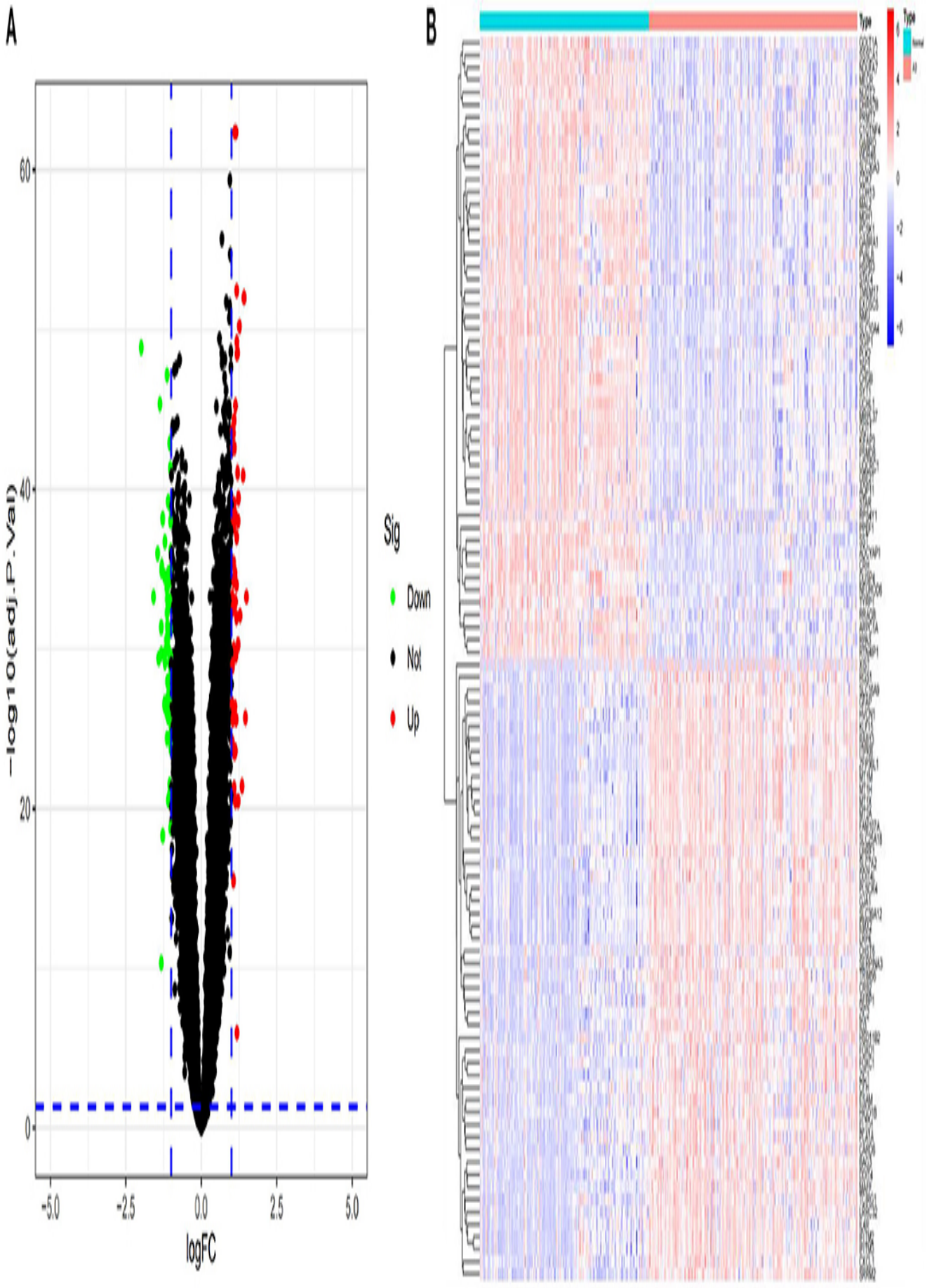
Brain Tissue Resource Center. Gene expression was analyzed using Rosetta/Merck Human 44k 1.1 microarray. GSE109887 was a dataset including 32 AD samples and 46 control samples. Brain and blood samples were collected through University Medical Center Göttingen. Gene expression was analyzed using Illumina HumanHT-12 v4 BeadChip. GSE132903 was a dataset including 98 AD samples and 97 control samples. Brain samples were collected through America Translational Genomic Research Institute. Gene expression was analyzed using Illumina Human HT-12 v4 arrays. Details about four datasets are shown in Table 1. Two datasets (GSE5281 and GSE44771) were combined to create a training dataset with a large sample size. Meanwhile, GSE109887 and GSE132903 were set as validation datasets. The training dataset was screened and eventually identified 120 significant DEGs related to AD based on logFC>1 and adjusted *P*-value <0.05. A volcano map was used to depict the expression status of all DEGs in the training dataset (Figure 2A). The difference between upregulated and downregulated genes were distinct. Using the heatmap, we can see which of the DEGs have the most upregulated gene expression compared to the control group (Figure 2B).

TABLE 1



Enlarge this image.

FIGURE 2

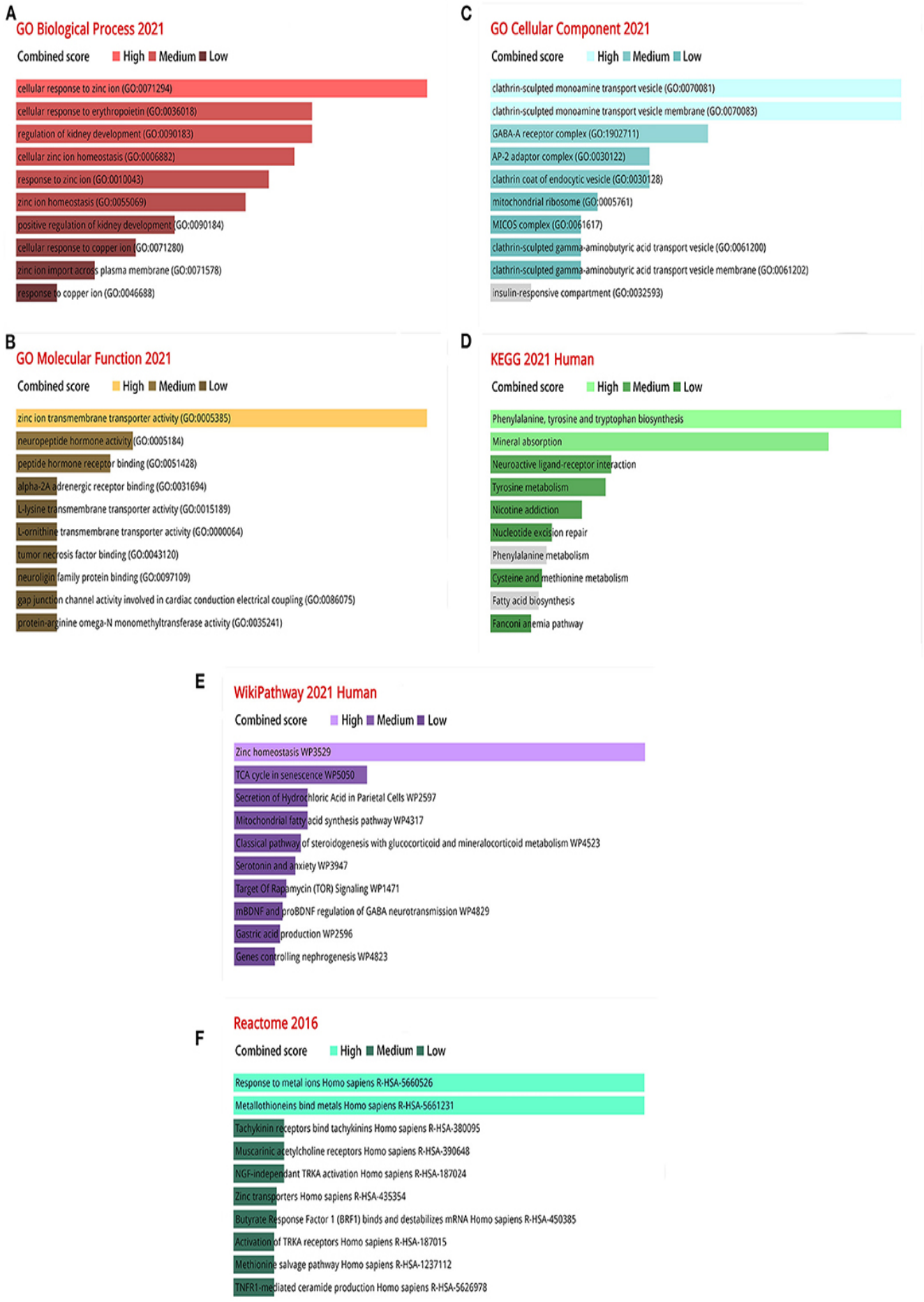


Enlarge this image.

Analysis of Gene Ontology and Pathway Enrichment

We analyzed the ontology and pathway enrichment for the 120 DEGs. For the Biological Process subsection, the

results demonstrate that the DEGs were significantly enhanced in cellular response to zinc ion. Molecular function subsection data indicated a zinc ion transmembrane transporter activity involved in the DEGs. The Cellular Component analysis revealed that clathrin-sculpted monoamine transport vesicle played a significant role. It showed the Phenylalanine, tyrosine and tryptophan biosynthesis, Zinc homeostasis and Response to metal ions interaction with the most important genes according to the KEGG, WikiPathway and Reactome pathway. The combined scores rank for GO terms and analysis results from various pathway databases are shown in Figure 3.
FIGURE 3



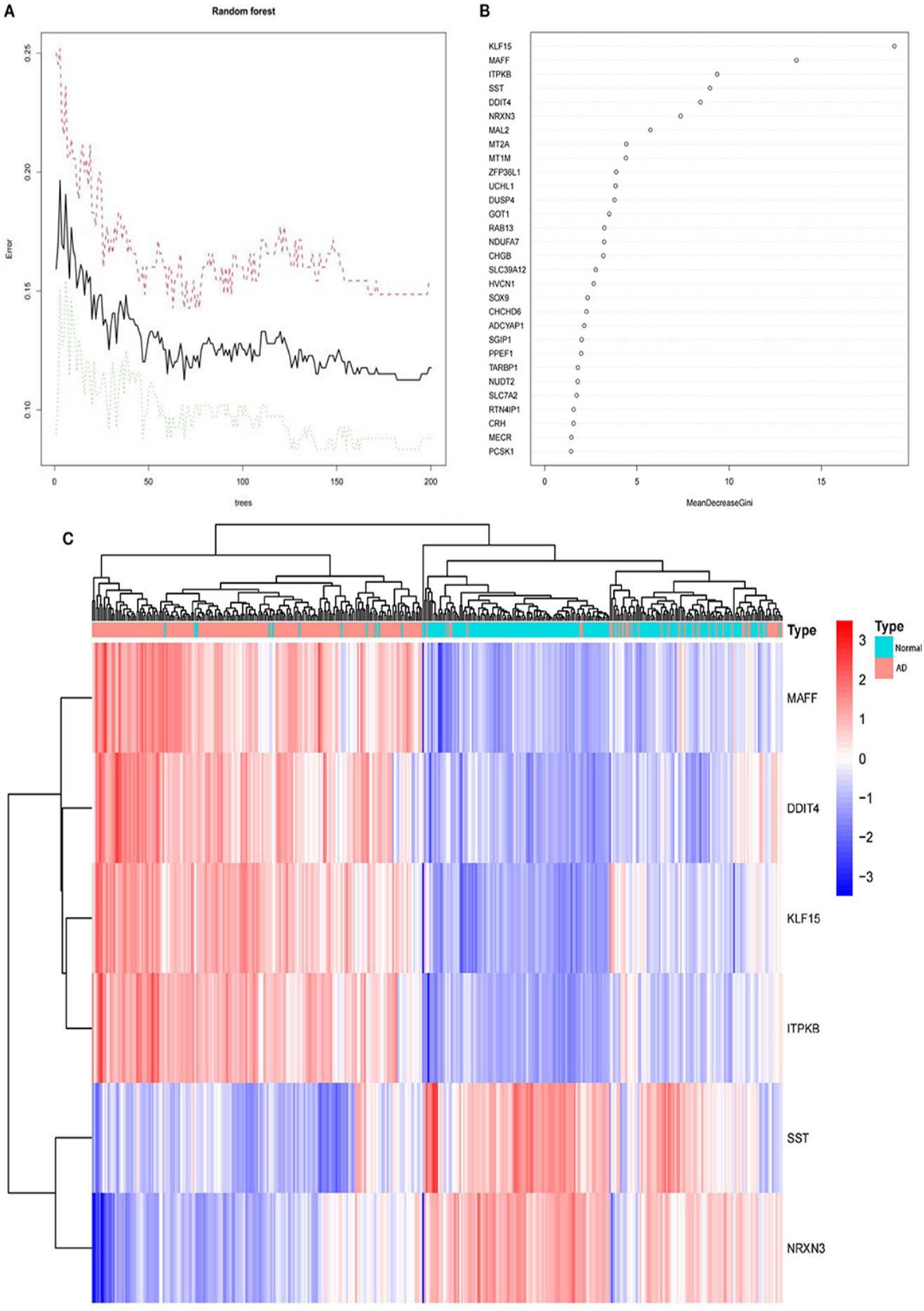
Enlarge this image.

Random Forest Screening for Key Genes

To obtain key genes, we fed the 120 DEGs listed above into the RF classifier. Based on the correlation plot between

the number of RF trees and model error (Figure 4A), we chose 190 trees as the final model's parameter. We then identified six genes with a significance >6 as candidate genes for further analysis. According to Figure 4B, KLF15 was the most significant variable, followed by MAFF, ITPKB, SST, DDIT4, and NRXN3. Figure 4C show that in 120 DEGs from the training dataset, the six genes were able to identify AD samples. MAFF, DDIT4, KLF15, and ITPKB genes were a group of genes whose expression was low in normal samples and high in AD samples. On the other hand, SST and NRXN3 belonged to a different cluster. In normal samples, they were expressed at high levels, but in AD samples, they were expressed at low levels.

FIGURE 4



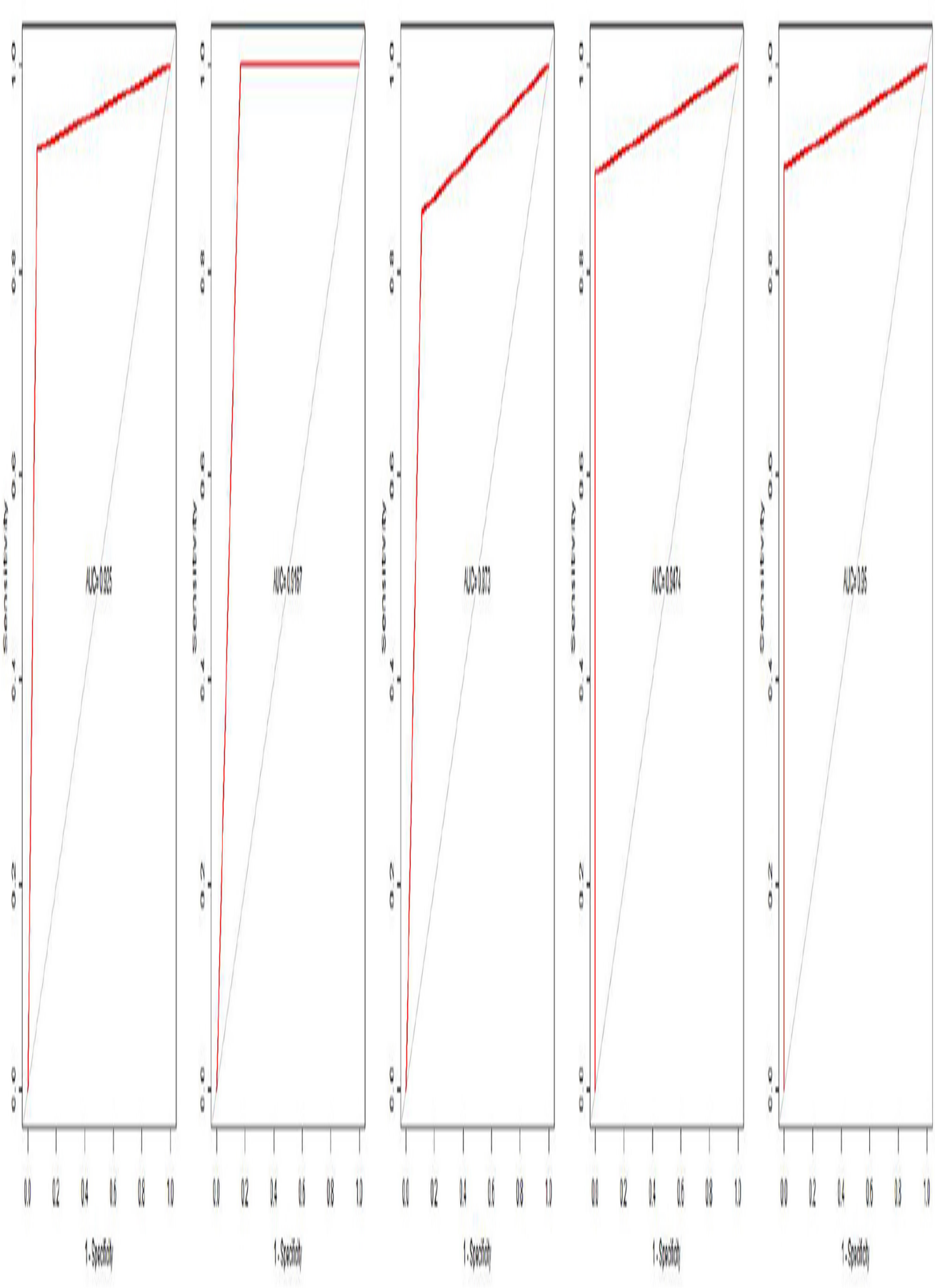
Enlarge this image.

Construction of the ANN Model

We got a Gene Score table with 6 lines of samples, 391 columns, and a column for the AD outcome variable

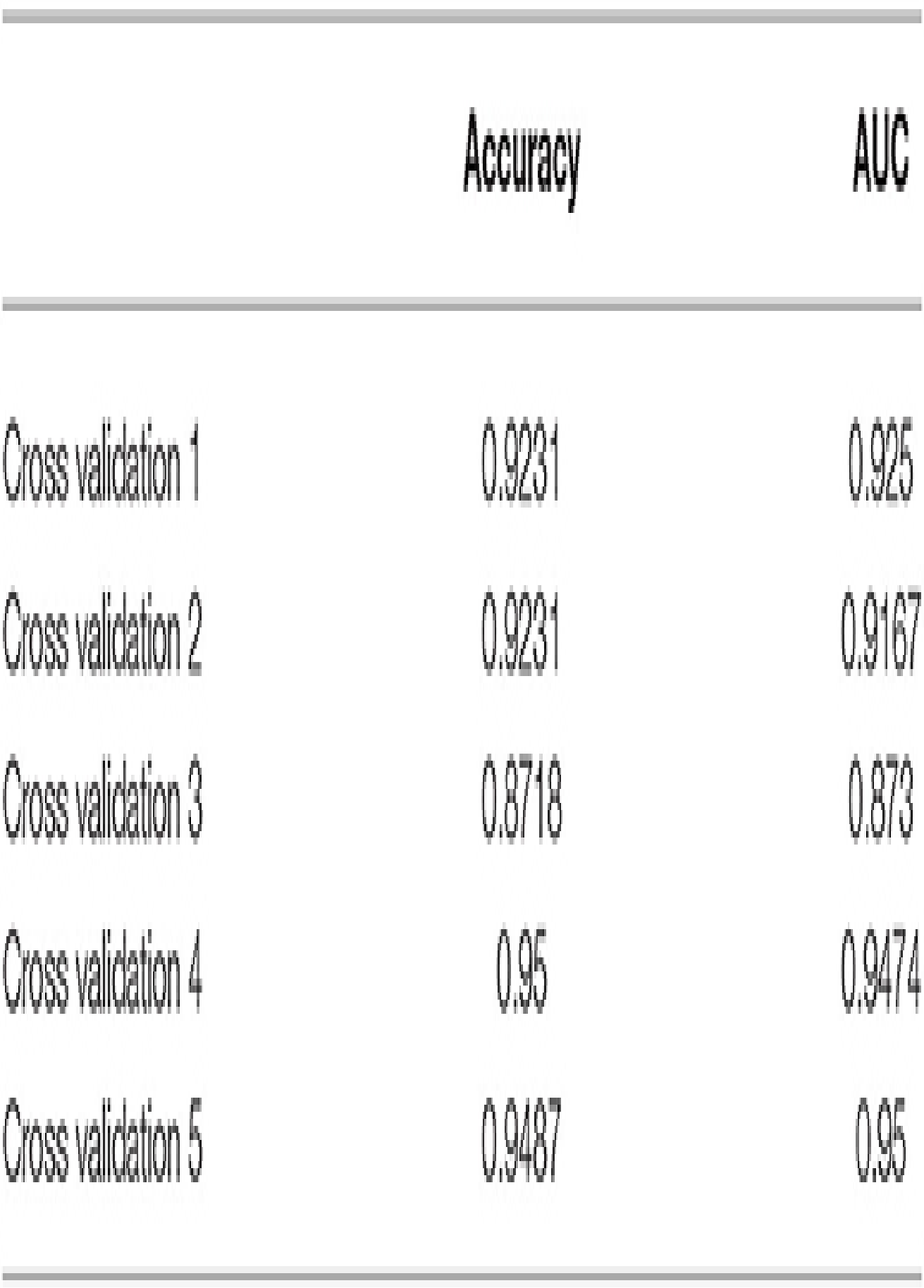
(case/control). We built an ANN model based on the Gene Score table. Six input layers, five hidden layers, and two output layers were set for the ANN. Each result of the 5-fold cross-validation is presented by ROC curves (Figure 5), while the accuracy is shown in Table 2. The model's reliability was demonstrated by the fact that the average AUC of the 5-fold cross-validation results exceeded 0.90. Finally, we built an ANN model for classifying gene expression data between AD and control samples based on the information presented above (Figure 6). The overall AUC of this model is 0.953, and its accuracy is 0.914 (Figure 7A).

FIGURE 5



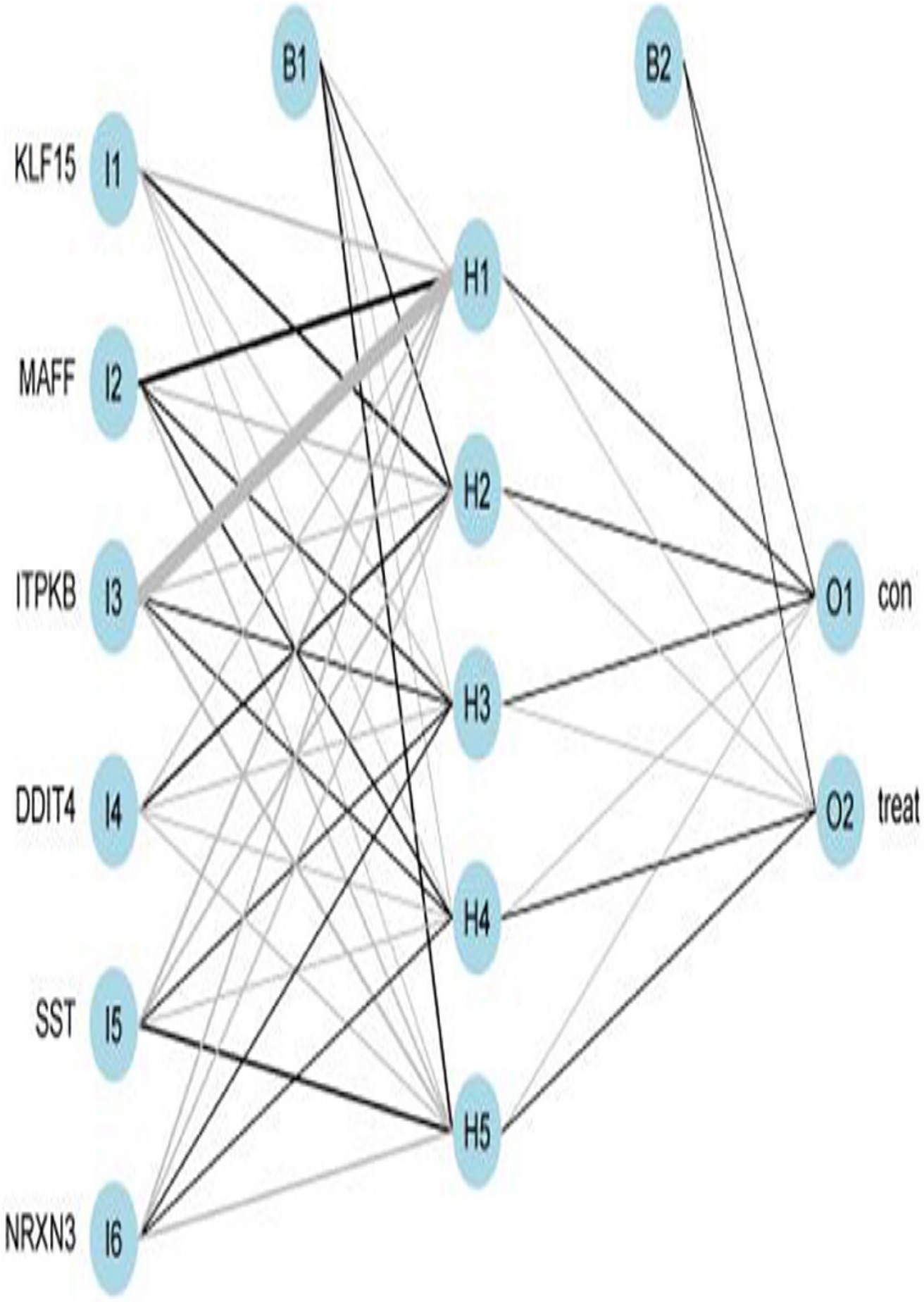
Enlarge this image.

TABLE 2



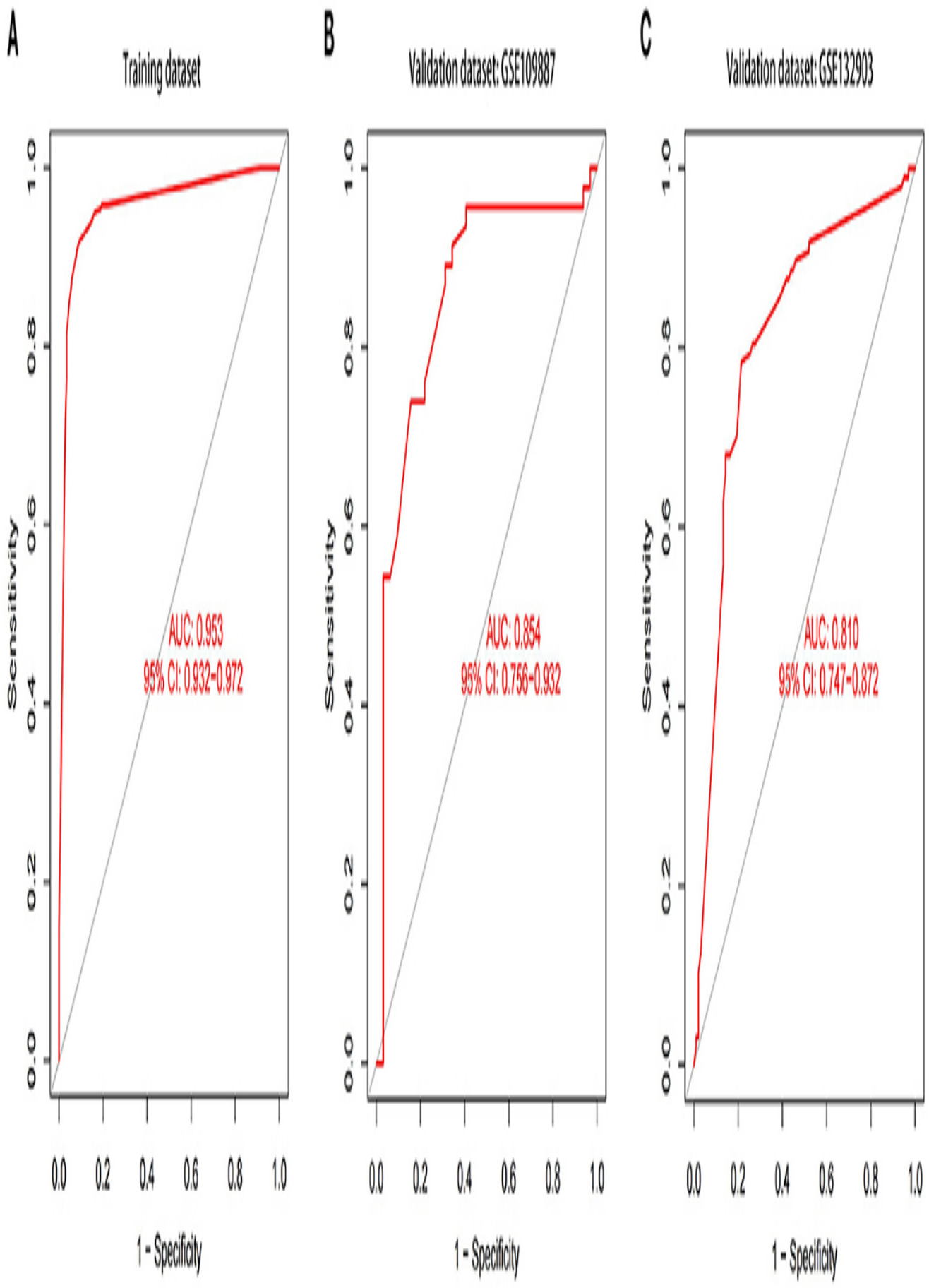
Enlarge this image.

FIGURE 6



Enlarge this image.

FIGURE 7



Enlarge this image.

Validation of the ANN Model

The model's prediction accuracy was 0.854 in GSE109887 and 0.810 in GSE132903, indicating that the ANN is

stable in diagnosing AD (Figure 7). These findings demonstrate that we successfully developed an AD diagnostic model based on the differential gene expression of AD and normal samples.

Discussion

Over the last century, advances in AD research have led to the development of increasingly effective treatments (Sun et al., 2018). However, the specific mechanisms of AD development remain unknown. It is almost impossible to make an early clinical diagnosis of AD because the symptoms overlap with those of other neuropathological diseases. Identifying critical diagnostic and prognostic biomarkers for AD remains critical. Advancements in machine learning and public gene expression data make it feasible to infer biomarkers for disease diagnosis and prognosis (Ramakrishnan et al., 2019).

In our study, we combined an AD diagnostic model with random forest and an artificial neural network that could distinguish AD samples from normal samples. Diagnostic evidence for diseases like AD is being bolstered by advances in high-speed bioinformatics. To identify DEGs of AD, we first combined two GEO datasets (GSE5281 and GSE44771). Then analyzed the gene ontology and pathway enrichment. According to the GO and pathway enrichment analysis, the DEGs are related to a vast array of GO terms and pathways, reflecting the pathogenesis' dynamics and complexity. There are already many studies supporting our findings. Prior research has suggested a connection between zinc ion and the occurrence of AD. The new research has uncovered a list of essential zinc ion transmembrane transporters whose mRNA or protein levels were found to be abnormally altered at various stages of AD (Xu et al., 2019). Changing zinc levels, especially at the synapses, have been suggested as a possible cause of cognitive changes that come with aging and AD (Hancock et al., 2014). Aged brains have been predicted to have less efficient homeostasis mechanisms and molecules for zinc ions (Bertoni-Freddari et al., 2006). The best way to understand an organism's internal changes is to conduct a pathway analysis. The disruption of phenylalanine metabolism in the hippocampus could be an important factor in the progression of AD (Liu et al., 2021). In AD, the peripheral modulation of tyrosine phosphorylation signaling could be investigated as a potential diagnostic marker (Mallozzi et al., 2020). It is possible that the pathogenesis of AD is influenced by immune activation-induced tryptophan degradation (Widner et al., 2000). Dyshomeostasis of zinc in the brain contributes to AD. Excess zinc is toxic to neuronal cells (Li and Wang, 2016). Homeostasis of metal ion levels is essential for normal physiological processes. Researchers have discovered a link between AD and an imbalance in the metal ions in the brain (Wang L. et al., 2020).

Further performance of RF classification screened out 6 key genes, namely, KLF15, MAFF, ITPKB, SST, DDIT4, and NRXN3. Previous research has supported our findings. Kruppel Like Factor 15 (KLF15) is a member of the Sp/KLF family of zinc-finger transcription factors. This family has been linked to controlling many cellular processes, such as cell growth, differentiation, normal development, and even cancer. It inhibits the growth of neurons (Otteson et al., 2004; Wang X. et al., 2020). MAF BZIP Transcription Factor F (MAFF) is upregulated in all tissues in AD. It can potentiate antioxidation inhibition and may be a potential therapeutic target in AD (Wang et al., 2017; Wang X. et al., 2020). Inositol (1,4,5) trisphosphate 3-kinase B (ITPKB) is an essential regulator in AD that plays a role in the apoptosis of neuronal cells, the processing of APP and the phosphorylation of tau (Stygelbout et al., 2014). Somatostatin (SST) receptor levels are lower in AD. SST-releasing neurons are often found near plaques. Its' expression levels decline with age (Beal et al., 1985; Roberts et al., 1985; Saito et al., 2005; Koivisto et al., 2007; Xue et al., 2009; Lau et al., 2017). Upregulation of DNA damage-inducible transcript 4 (DDIT4), a stress-regulated protein, can cause neuronal trigger death. It has been identified as a biomarker for AD (Pérez-Sisqués et al., 2021). Neurexin 3 (NRXN3) is a type of presynaptic adhesion molecule that regulates neurotransmitter release and specifies neuron synapses. In AD patients, NRXN3 expression is reduced. Dysregulation of presynaptic NRXN3 expression and splicing may promote neuron inflammation in the AD brain (Hishimoto et al., 2019). These studies demonstrated that the 6 key genes could be used as key biomarkers of AD.

The highlight of our study is the innovative combination of RF and ANN methods which yielded excellent results in terms of predictive power. Several other diseases, including ulcerative colitis, heart failure, and polycystic ovary syndrome, have already benefited from this innovative research technique (Li et al., 2020; Tian et al., 2020; Xie et

al., 2020). Prior to this, a few AD prediction models based on methylated gene biomarkers had been developed (Ren et al., 2020; Mahendran and PM, 2022). However, some problems exist in these studies, such as small sample size or general prediction effect of the established models. Our model performed better on the validation datasets GSE109887 and GSE132903, with AUC of 0.854 and 0.810, indicating it is more suitable for AD classification. Even so, there are some limitations in our research. Although we used two datasets with more samples to build a model, it is still not a big data sample for machine learning, and we can include more research data in the training set in the future. Overfitting in machine learning is objective and cannot be eliminated, even if we use 5-fold cross-validation in the modeling process to minimize overfitting. Checking for overfitting is helpful, but it does not solve the problem. This means that although we get a good model effect on the validation set, the actual generalization ability may not be good due to the appearance of noise in reality. So this still means that we need to include more research data to test the reliability of the model in the future.

Conclusions

To summarize, our thorough examination of AD datasets from GEO revealed KLF15, MAFF, ITPKB, SST, DDIT4, and NRXN3 as potential diagnostic biomarkers. Based on machine learning algorithms employing RF and ANN, a diagnostic model for AD was created that demonstrated excellent prediction performance.

Data Availability Statement

The datasets presented in this study can be found in online repositories. The names of the repository and accession numbers can be found below: https://www.ncbi.nlm.nih.gov/geo/. GEO accession numbers: GSE5281, GSE44771, GSE109887 and GSE132903.

Author Contributions

DS: data collation and drafting the manuscript. HP: technical review and revision of data analysis. ZW: project administration and funding support. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the National Key R&D Program of China and Ministry of Science and Technology of China, Grant Number 2018YFC2002504.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aisen, P. S., Cummings, J., Jack, C. J., Morris, J. C., Sperling, R., Frölich, L., et al. (2017). On the path to 2025: Understanding the Alzheimer's disease continuum. *Alzheimers Res. Ther.* 9, 60. doi: 10.1186/s13195-017-0283-5

Beal, M. F., Mazurek, M. F., Tran, V. T., Chattha, G., Bird, E. D., and Martin, J. B. (1985). Reduced numbers of somatostatin receptors in the cerebral cortex in Alzheimer's disease. *Science.* 229, 289–291. doi: 10.1126/science.2861661

Beck, M. W.. (2018). NeuralNetTools: Visualization and analysis tools for neural networks. *J. Stat. Softw.* 85, 1–20. doi: 10.18637/jss.v085.i11

Bertoni-Freddari, C., Mocchegiani, E., Malavolta, M., Casoli, T., Di Stefano, G., and Fattoretti, P. (2006). Synaptic and mitochondrial physiopathologic changes in the aging nervous system and the role of zinc ion homeostasis. *Mech. Ageing Dev.* 127, 590–596. doi: 10.1016/j.mad.2006.01.019

Blennow, K., and Zetterberg, H. (2018). Biomarkers for Alzheimer's disease: Current status and prospects for the future. *J. Intern. Med.* 284, 643–663. doi: 10.1111/joim.12816

Chen, Y. G.. (2018). Research progress in the pathogenesis of alzheimer's disease. *Chin. Med. J.* 131, 1618–1624. doi: 10.4103/0366-6999.235112

Delaby, C., Alcolea, D., Hirtz, C., Vialaret, J., Kindermans, J., Morichon, L., et al. (2022). Blood amyloid and tau biomarkers as predictors of cerebrospinal fluid profiles. *J. Neural Transm.* 129, 231–237. doi: 10.1007/s00702-022-02474-9
Hancock, S. M., Finkelstein, D. I., and Adlard, P. A. (2014). Glia and zinc in ageing and Alzheimer's disease: a mechanism for cognitive decline? *Front. Aging Neurosci.* 6, 137. doi: 10.3389/fnagi.2014.00137
Hishimoto, A., Pletnikova, O., Lang, D. L., Troncoso, J. C., Egan, J. M., and Liu, Q. R. (2019). Neurexin 3 transmembrane and soluble isoform expression and splicing haplotype are associated with neuron inflammasome and Alzheimer's disease. *Alzheimers Res. Ther.* 11, 28. doi: 10.1186/s13195-019-0475-2
Hsieh, C. H., Lu, R. H., Lee, N. H., Chiu, W. T., Hsu, M. H., and Li, Y. C. (2011). Novel solutions for an old disease: Diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. *Surgery*. 149, 87–93. doi: 10.1016/j.surg.2010.03.023
Hu, K.. (2021). Become competent in generating RNA-Seq heat maps in one day for novices without prior r experience. *Methods Mol. Biol*. 2239, 269–303. doi: 10.1007/978-1-0716-1084-8_17
Ito, K., and Murphy, D. (2013). Application of ggplot2 to pharmacometric graphics. *CPT Pharmacom. Syst. Pharmacol*. 2, e79. doi: 10.1038/psp.2013.56
Janßen, R., Zabel, J., von Lukas, U., and Labrenz, M. (2019). An artificial neural network and Random Forest identify glyphosate-impacted brackish communities based on 16S rRNA amplicon MiSeq read counts. *Mar. Pollut. Bull.* 149, 110530. doi: 10.1016/j.marpolbul.2019.110530
Koivisto, A. M., Tapaninen, T., Hiltunen, M., and Soininen, H. (2007). Somatostatin genetic variants modify the risk for Alzheimer's disease among Finnish patients. *J. Neurol.* 254, 1504–1508. doi: 10.1007/s00415-007-0539-2
Kong, Y., and Yu, T. (2018). A deep neural network model using random forest to extract feature representation for gene expression data classification. *Sci. Rep*. 8, 16477. doi: 10.1038/s41598-018-34833-6
Kulasingam, V., and Diamandis, E. P. (2008). Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nat. Clin. Pract. Oncol*. 5, 588–599. doi: 10.1038/ncponc1187
Kursa, M. B.. (2014). Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics*. 15, 8. doi: 10.1186/1471-2105-15-8
Lau, A., Bourkas, M., Lu, Y. Q. Q., Ostrowski, L. A., Weber-Adrian, D., Figueiredo, C., et al. (2017). Functional amyloids and their possible influence on Alzheimer disease. *Discoveries*. 5, e79. doi: 10.15190/d.2017.9
Li, H., Lai, L., and Shen, J. (2020). Development of a susceptibility gene based novel predictive model for the diagnosis of ulcerative colitis using random forest and artificial neural network. *Aging*. 12, 20471–20482. doi: 10.18632/aging.103861
Li, L. B., and Wang, Z. Y. (2016). Disruption of brain zinc homeostasis promotes the pathophysiological progress of Alzheimer's disease. *Histol. Histopathol.* 31, 623–627. doi: 10.14670/HH-11-737
Liu, P., Yang, Q., Yu, N., Cao, Y., Wang, X., Wang, Z., et al. (2021). Phenylalanine metabolism is dysregulated in human hippocampus with alzheimer's disease related pathological changes. *J. Alzheimers Dis.* 83, 609–622. doi: 10.3233/JAD-210461
Mahendran, N., and PM, D. R. V. (2022). A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease. *Comput. Biol. Med.* 141, 105056. doi: 10.1016/j.compbiomed.2021.105056
Mallozzi, C., Crestini, A., D'Amore, C., Piscopo, P., Cappella, M., Perrone, F., et al. (2020). Activation of tyrosine phosphorylation signaling in erythrocytes of patients with alzheimer's disease. *Neuroscience*. 433, 36–41. doi: 10.1016/j.neuroscience.2020.02.050
Nachid, M., and Boussiala, M. (2021). Machine Learning_Iris_caret Package. Researchgate. doi: 10.13140/RG.2.2.12576.71683. Available online at: https://www.researchgate.net/publication/353846714_Machine_Learning_Iris_caretPackage
Otteson, D. C., Liu, Y., Lai, H., Wang, C., Gray, S., Jain, M. K., et al. (2004). Kruppel-like factor 15, a zinc-finger transcriptional regulator, represses the rhodopsin and interphotoreceptor retinoid-binding protein promoters.

Invest. Ophthalmol. Vis. Sci. 45, 2522–2530. doi: 10.1167/iovs.04-0072
Pérez-Sisqués, L., Sancho-Balsells, A., Solana-Balaguer, J., Campoy-Campos, G., Vives-Isern, M., Soler-Palazón, F., et al. (2021). RTP801/REDD1 contributes to neuroinflammation severity and memory impairments in Alzheimer's disease. *Cell Death Dis.* 12, 616. doi: 10.1038/s41419-021-03899-y
R project. (2022). *Breiman and Cutler's Random Forests for Classification and Regression*. Available online at: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf (accessed May 23, 2022).
Ramakrishnan, A., Pardes, A., Lynch, W., Molaro, C., and Platt, M. L. (2019). A machine learning approach to identifying objective biomarkers of anxiety and stress. *bioRxiv*. doi: 10.1101/745315
Reitz, C.. (2015). Genetic diagnosis and prognosis of Alzheimer's disease: Challenges and opportunities. *Expert Rev. Mol. Diagn.* 15, 339–348. doi: 10.1586/14737159.2015.1002469
Ren, J., Zhang, B., Wei, D., and Zhang, Z. (2020). Identification of Methylated Gene Biomarkers in Patients with Alzheimer's Disease Based on Machine Learning. *Biomed. Res. Int.* 2020, 1–11. doi: 10.1155/2020/8348147
Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47. doi: 10.1093/nar/gkv007
Roberts, G. W., Crow, T. J., and Polak, J. M. (1985). Location of neuronal tangles in somatostatin neurones in Alzheimer's disease. *Nature*. 314, 92–94. doi: 10.1038/314092a0
Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., and Müller, M. (2011). PROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* 12, 77. doi: 10.1186/1471-2105-12-77
Saito, T., Iwata, N., Tsubuki, S., Takaki, Y., Takano, J., Huang, S. M., et al. (2005). Somatostatin regulates brain amyloid beta peptide Abeta42 through modulation of proteolytic degradation. *Nat. Med.* 11, 434–439. doi: 10.1038/nm1206
Scheltens, P., Blennow, K., Breteler, M. M., de Strooper, B., Frisoni, G. B., Salloway, S., et al. (2016). Alzheimer's disease. *Lancet*. 388, 505–517. doi: 10.1016/S0140-6736(15)01124-1
Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C. E., et al. (2021). Alzheimer's disease. *Lancet*. 397, 1577–1590. doi: 10.1016/S0140-6736(20)32205-4
Stygelbout, V., Leroy, K., Pouillon, V., Ando, K., D'Amico, E., Jia, Y., et al. (2014). Inositol trisphosphate 3-kinase B is increased in human Alzheimer brain and exacerbates mouse Alzheimer pathology. *Brain*. 137, 537–552. doi: 10.1093/brain/awt344
Sun, B. L., Li, W. W., Zhu, C., Jin, W. S., Zeng, F., Liu, Y. H., et al. (2018). Clinical Research on Alzheimer's Disease: Progress and Perspectives. *Neurosci. Bull.* 34, 1111–1118. doi: 10.1007/s12264-018-0249-z
Tian, Y., Yang, J., Lan, M., and Zou, T. (2020). Construction and analysis of a joint diagnosis model of random forest and artificial neural network for heart failure. *Aging (Albany NY)*. 12, 26221–26235. doi: 10.18632/aging.202405
Varma, S.. (2020). Blind estimation and correction of microarray batch effect. *PLoS ONE*. 15, e231446. doi: 10.1371/journal.pone.0231446
Veitch, D. P., Weiner, M. W., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., et al. (2019). Understanding disease progression and improving Alzheimer's disease clinical trials: Recent highlights from the Alzheimer's Disease Neuroimaging Initiative. *Alzheimers Dement.* 15, 106–152. doi: 10.1016/j.jalz.2018.08.005
Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97, 576–592. doi: 10.1016/j.ajhg.2015.09.001
Wang, L., Yin, Y. L., Liu, X. Z., Shen, P., Zheng, Y. G., Lan, X. R., et al. (2020). Current understanding of metal ions in the pathogenesis of Alzheimer's disease. *Transl. Neurodegener.* 9, 10. doi: 10.1186/s40035-020-00189-z
Wang, Q., Li, W. X., Dai, S. X., Guo, Y. C., Han, F. F., Zheng, J. J., et al. (2017). Meta-Analysis of parkinson's disease and alzheimer's disease revealed commonly impaired pathways and dysregulation of NRF2-Dependent genes. *J. Alzheimers Dis.* 56, 1525–1539. doi: 10.3233/JAD-161032
Wang, X., Zhang, Y., Wan, X., Guo, C., Cui, J., Sun, J., and Li, L. (2020). Responsive expression of MafF to -Amyloid-induced oxidative stress. *Dis. Markers*. 2020, 8861358. doi: 10.1155/2020/8861358

Widner, B., Leblhuber, F., Walli, J., Tilz, G. P., Demel, U., and Fuchs, D. (2000). Tryptophan degradation and immune activation in Alzheimer's disease. *J. Neural. Transm.* 107, 343–353. doi: 10.1007/s007020050029
Xie, N. N., Wang, F. F., Zhou, J., Liu, C., and Qu, F. (2020). Establishment and analysis of a combined diagnostic model of polycystic ovary syndrome with random forest and artificial neural network. *Biomed. Res. Int.* 2020, 2613091. doi: 10.1155/2020/2613091
Xu, Y., Xiao, G., Liu, L., and Lang, M. (2019). Zinc transporters in Alzheimer's disease. *Mol. Brain*. 12, 106. doi: 10.1186/s13041-019-0528-2
Xue, S., Jia, L., and Jia, J. (2009). Association between somatostatin gene polymorphisms and sporadic Alzheimer's disease in Chinese population. *Neurosci. Lett.* 465, 181–183. doi: 10.1016/j.neulet.2009.09.002
Zhang, Z., Chen, L., Humphries, B., Brien, R., Wicha, M. S., Luker, K. E., et al. (2018). Morphology-based prediction of cancer cell migration using an artificial neural network and a random decision forest. *Integr. Biol.* 10, 758–767. doi: 10.1039/C8IB00106E
Zhu, M., Jia, L., Li, F., and Jia, J. (2020). Identification of KIAA0513 and other hub genes associated with alzheimer disease using weighted gene coexpression network analysis. *Front. Genet.* 11, 981. doi: 10.3389/fgene.2020.00981

AuthorAffiliation
Dazhong Sun, Haojun Peng and Zhibing Wu*
* The First Clinical Medical School, Guangzhou University of Chinese Medicine, Guangzhou, China

DETAILS

Subject:	Biomarkers; Computational neuroscience; Gene expression; Datasets; Neurodegenerative diseases; Alzheimer's disease; Protein gene product 9.5; Neural networks; Alzheimers disease
Identifier / keyword:	Alzheimer's disease; random forest; artificial neural network; GEO; Diagnostic model
Publication title:	Frontiers in Aging Neuroscience; Lausanne
Publication year:	2022
Publication date:	Jun 30, 2022
Section:	ORIGINAL RESEARCH article
Publisher:	Frontiers Research Foundation
Place of publication:	Lausanne
Country of publication:	Switzerland, Lausanne
Publication subject:	Gerontology And Geriatrics, Medical Sciences--Psychiatry And Neurology
ISSN:	16634365
e-ISSN:	16634365
Source type:	Scholarly Journal

Language of publication:	English
Document type:	Journal Article
Publication history :	
Online publication date:	0001-01-01
Milestone dates:	2022-04-16 (Received); 2022-06-01 (Accepted); 2022-06-30 (Published)
Publication history :	
First posting date:	01 Jan 0001
DOI:	https://doi.org/10.3389/fnagi.2022.921906
ProQuest document ID:	2682564611
Document URL:	https://www.proquest.com/scholarly-journals/establishment-analysis-combined-diagnostic-model/docview/2682564611/se-2
Copyright:	© 2022. This work is licensed under http://creativecommons.org/licenses/by/4.0/ (the “License”). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.
Last updated:	2022-07-01
Database:	Publicly Available Content Database