HIGH SCHOOL EDITION
Journal of Student Research
Volume 11 Issue 3 (2022)

Deep Neural Network Classifier for Alzheimer's Disease

Jason Lin[1], Hayan Lee[#] and Michael Snyder[#]

[1]Evergreen Valley High School, San Jose, CA, USA
[#]Advisor

ABSTRACT

Alzheimer's disease (AD) is a neurodegenerative disease characterized by dementia and, eventually, a loss of cognitive abilities. Two histopathological features are associated with AD, neurofibrillary tangles, and amyloid-beta plaque. Both contribute to neuron cell death, neuron dysfunction, and AD pathogenesis. Current methods to diagnose AD remain reliant on symptomatic diagnosis with interviews that can be time-consuming, costly, and inaccurate. Alternative methods such as brain imaging are expensive and require extensive laboratory setup for accurate results. Thus molecular-level quantitative approaches are necessary. Omics datasets and machine learning technology advancements have opened new avenues to diagnose AD. This paper proposes using statistical methods such as principal component analysis, t-distributed stochastic neighbor embedding, and Kolmogorov-Smirnov test combined with Benjamini-Hochberg correction through feature selection and dimensionality reduction to isolate significant features associated with AD. Furthermore, we developed machine learning models based on logistic regression, random forest classifier, and deep neural network (DNN) classifier to predict AD diagnosis. Eight unique genes (TGM2, NKIRAS1, SYK, GABARAPL2, ABCC12, NDEL1, TEP1) were identified as significant biomarkers of AD and confirmed previous works identifying prognoses' roles in AD. After extensive hyperparameter tuning, the DNN model showed the best prediction performance for AD diagnosis among the three machine learning algorithms. The DNN model and preprocessed dataset demonstrated a 5-fold cross-validation accuracy of 0.823 and AUC-ROC of 0.940. Its code is publicly available at https://www.kaggle.com/neobrando/ml-dnn.

Introduction

Alzheimer's is a severe, chronic neurodegenerative disease characterized by mild cognitive impairment and slow loss of memory and cognitive ability. Dementia caused by Alzheimer's often interferes with daily function and eventually leads to complete dependency on others (National Institute on Aging, 2021). Alzheimer's is documented as the 7th leading cause of death in the US, and the Alzheimer's association stated deaths per year from Alzheimer's is projected to rise due to the aging population (Alzheimer's Association, 2021).

Alzheimer's disease (AD) is divided into two sub-variants based on onset age. Early Onset AD (EOAD) begins from ages 30-65, and Late Onset AD (LOAD) starts at age 65. EOAD and LOAD are more common in families with a history of AD. 60% of EOAD cases have at least one other AD within their family; of these, 13% inherit the disease in an autosomal dominant manner (Campion et al., 1999) (Brickell et al., 2006).

Alzheimer's disease involves the presence of 2 histopathological features: neurofibrillary tangles and amyloid-beta plaque (Braak & Braak, 1997) (Bekris et al., 2010). Neurofibrillary tangles in AD brains commonly contain hyperphosphorylated tau, which aggregates into an insoluble form (Iwatsubo et al., 1994). High concentrations of these tau tangles cause the death of neurons through the impediment of intracellular nutrient and neurotransmitter transportation (National Institute on Aging, 2017). Despite the presence of tau tangles in AD, the mutation of the gene MAPT, which encodes the main component of neurofibrillary tangles, has not

been genetically linked to AD (Iwatsubo et al., 1994). Amyloid-beta plaque is believed to develop due to mutation of APP, amyloid precursor protein. Specifically, APP-derived 42 amino acid residue is implicated in oligomerization and accumulates in the form of plaques (Goedert & Spillantini, 2006). These amyloid plaques contribute to neurodegeneration in patients with Alzheimer's through direct interference with neuron communication (National Institute on Aging, 2021).

Genetics of Alzheimer's

Twin studies support genetic causation for AD. 79% of AD risk is associated with genetic influence, with a 45% concordance rate of AD among identical male pairs (University of Southern California., 2006). Based on these results, a few genes have been identified as associated with either autosomal dominant or sporadic inheritance of AD: Amyloid Precursor Protein (APP), Presenilin 1 (PSEN1), Presenilin 2 (PSEN2), and Apolipoprotein E (APOE) (Bekris et al., 2010).

Purpose

Recent developments in the use of big data and machine learning algorithms have created the "Omics-Era." (Sancesario & Bernardini, 2018). Furthermore, with investigation in the genetics of AD reopened as a result of scrutiny over the landmark 2006 APP Nature study (Piller, 2022), the use of computational bioinformatics such as genome-wide association studies (GWASs) may identify novel pathways in the development of AD (Bellenguez et al., 2022). In this study, we used statistical analysis methods to determine the significance of the expression of genes in controls and AD patients based on the dataset provided in the paper "Prediction of Alzheimer's disease based on the deep neural network by integrating gene expression and a DNA methylation dataset" (Park, 2021). Then, we created logistic regression, random forest, and deep neural network models to test the applicability of the determined genes in diagnosing AD patients. Developing machine learning models using omics data at the molecular level is necessary to predict AD more quantitatively since conventional AD diagnosis were mostly based on interview and brain imaging, which is labor-intensive and costly. This study investigates multiple techniques of machine learning performance testing as a form of AD diagnosis through a quantitative format.

Methods

Datasets

The datasets used in this study focus on two types of omics datasets: gene expression and DNA methylation titled "allforDNN_ge_sample.tsv" and "allforDNN_me_sample.tsv" respectively (Park et al., 2020). These datasets were provided by the original authors of the paper, "Prediction of Alzheimer's disease based on the deep neural network by integrating gene expression and a DNA methylation dataset," by combining two large-scale gene expression profiles, GSE33000 (Narayanan et al., 2014) and GSE44770 (Zhang, 2013), which focused on tissue samples from the prefrontal cortex area. This combined dataset, "allforDNN_ge_sample.tsv", contains 257 normal and 439 AD samples with 200 gene features, Sample ID, Label_AD, and Label_NO.

The DNA methylation profiles from GSE80970 (Smith, 2018), containing 68 normal and 74 AD samples, were modified into the dataset, "allforDNN_me_sample.tsv" with 500 methylation features, Sample ID, Label_AD, and Label_NO. The GSE dataset's original beta-value was converted into M-values to increase

statistical validity when creating "allforDNN_me_sample.tsv" (Du et al., 2010). All datasets were subsequently normalized in preparation for statistical analysis.

Statistical Analysis

Due to the high number of features within the dataset, statistical analysis via dimension reduction was carried out to reduce the risk of overfitting and feature space. Dimension reduction also provides a means to identify significant gene expression or methylation expression patterns for the onset of AD by eliminating non-significant features from processed datasets (van Driel & Brunner, 2006). To accomplish this, we propose three approaches for dimension reduction: Principal Component Analysis, T-distributed Stochastic Neighbor Embedding, and the Kolmogorov Smirnov Test with False Discovery Rate P-value Correction.

Principal Component Analysis

Principal Component Analysis (PCA) is a regularly used dimension reduction method that searches for linear combinations of features in principal components (PC) and reduces the dimensionality of datasets (Ma & Dai, 2011). In this study, we use the PCA library provided by Sklearn that uses "the LAPACK implementation of the full SVD or a randomized truncated SVD by the method of Halko et al. 2009" (Sklearn, 2009). In implementing PCA, the "svd_solver" option was set to full to select for PC by use of the standard LAPACK solver and postprocessing. In this study, PCA plots were visualized with the plot visualization software provided by plotly (Plotly, n.d.). PCA was preferred over alternatives such as ICAs since no specific independent condition was identified as necessary to be optimized. Rather, since PCA focuses on variance, it was sufficient to indicate features with statistically significant difference of expression between control and AD sets (Ma & Dai, 2011).

T-distributed Stochastic Neighbor Embedding

T-distributed Stochastic Neighbor Embedding (t-SNE) is another dimensionality reduction step commonly used in data analysis pipelines for genetic analysis (Kobak & Berens, 2019). Unlike PCA, t-SNE is a non-linear form of dimensionality reduction and preserves the neighborhood to a point to determine variance. In this study, we used the t-SNE package from Sklearn that implements the t-SNE test by minimizing the Kullbac-Leiber divergence on the gene expression and DNA methylation datasets (Sklearn, 2014). In running the package, the method was set to "exact" to run a slower, exact algorithm without the default Barnes-Hut approximation and increase possible accuracy.

Kolmogorov-Smirnov Test (KS test) and False Discovery Rate p-value correction (FDR)

In this paper, a combination of the Kolmogorov-Smirnov Test (KS test) and False Discovery Rate p-value correction (FDR) were used to determine the existence of significant differences of genetic and methylation feature expression between AD patients and normal controls. KS test has been successfully applied in previous studies in the analysis of disease gene data, including multiple cancers (Su et al., 2017), and is commonly paired with the use of FDR (Rogers & Weiss, 2017). Before using a KS test, we split the gene expression and methylation expression datasets between control and AD patients. Then we implemented the Scipy "ks_2samp" package on both datasets (Scipy, n.d.). With the resulting p-values from the KS test, we used the statsmodel false discovery rate to correct p-value of each gene and methylation feature (Statsmodels, 2019). These p-values were then filtered to create a gene and methylation features list with p-values < 0.05.

Machine Learning Models

Logistic Regression

Logistic Regression is a classifier algorithm which estimates the probability of an event occurring or classification between multiple features. We used the Sklearn Logistic Regression software with the maximum number of interactions modified to 2,000 and using the “SAG" solver (IBM, n.d.) made by Mark Schmidt, Nicolas Le roux, and Fancis Bach in the study “Minimizing Finite Sums with the Stochastic Average Gradient” (Schmidt et al., 2017). The “SAG” solver was used specifically over the default lbfgs since SAG provides faster convergence on the features provided and thus the most accurate model possible with our sample limited dataset.

Random Forest Classifier

The random forest regressor is an estimator algorithm that fits trees to subsamples of the dataset to improve the predictive accuracy of the function and identify significant features through sampling. In this study, we used the RandomForestClassifier available in the Sklearn package (Sklearn, 2018) to create a RandomForestClassifier model. For both the methylation and genetic datasets we replicated the RandomForestClassifier from the study “Prediction of Alzheimer’s disease based on deep neural network by integrating gene expression and DNA methylation dataset” study’s github. The study used the following parameters: “criterion = 'entropy', oob_score = True, n_estimators = 100, n_jobs = -1, random_state = 0, max_depth = 6” (Park et al., 2020). Random Forest is commonly used as an alternative to logistic regression as a standard approach for binary classification (Couronné et al., 2018). The addition of Random Forest provides a secondary confirmation tool for the effectiveness of selected features (Couronné et al., 2018).

Deep Neural Network

A Deep Neural Network is a machine learning method that uses the implementation of layers of individual processing units called neurons to determine weights between features and create a prediction model (Snoek et al., 2012). The DNN model was created from exhaustive hyperparameter search through repeated trials. Thus, the DNN used in this study comprises 8 ReLu (rectified linear) activation function hidden layers with 306 neurons and a sigmoid final layer for classification between normal and AD. The model’s learning rate is set at 0.02 and the dropout rate set to 0.85, and a callback feature was added to terminate the model training if a decline in accuracy due to overfitting was detected. The model of DNN was implemented with API in Google TensorFlow v2 (Martin et al., 2015).

All machine learning models were evaluated primarily using the receiver operating characteristic and the area under the curve (AUC-ROC) approach. Models were also evaluated using validation accuracy to determine their accuracy over a similar dataset compared with the testing set. AUC-ROC is commonly used in the evaluation of medical diagnostic tests due to its evaluated accuracy not being influenced by the decision criterion (Hajian-Tilaki, 2013). Thus, AUC-ROC was ultimately used to evaluate model effectiveness as it is able to assess the inherent ability of a diagnostic test to distinguish between a diseased and control population and serve as an accurate indicator of model effectiveness in this study

Results

PCA

As shown in figure 1, analysis of the genetic PCA plot based on the top 4 PCs with classification based on controls versus AD shows PC1, PC2, PC3, PC4 explain 30.5%, 11.8%, 5.2%, and 5.1% of the variants respectively. The methylation PCA plot shows a much lower variance associated with the first PC: PC1, PC2, PC3, and PC4, explaining 21.3%, 16.0%, 5.4%, and 4.6% of the variance, respectively. The results indicate that a discernable difference in gene expression is observed between normal and AD patients but not the methylation

expression dataset which seems to separate based on an unknown artifact. Thus, the methylation dataset was discarded for the remainder of this study due to the existence of a confounding effect.

t-SNE plots

The t-SNE preprocess helped further reduce the dataset's dimension. As seen in figures 3, similar to the PCA plot, there is a clear separation between normal and AD for the genetic t-SNE plot. Interestingly both the PCA plot and t-SNE plot of the genetic expression dataset suggest that patients of AD have reduced expression of their associated features.

Figure 1. PCA Plot of Principal Components representing the features of Genetic Dataset

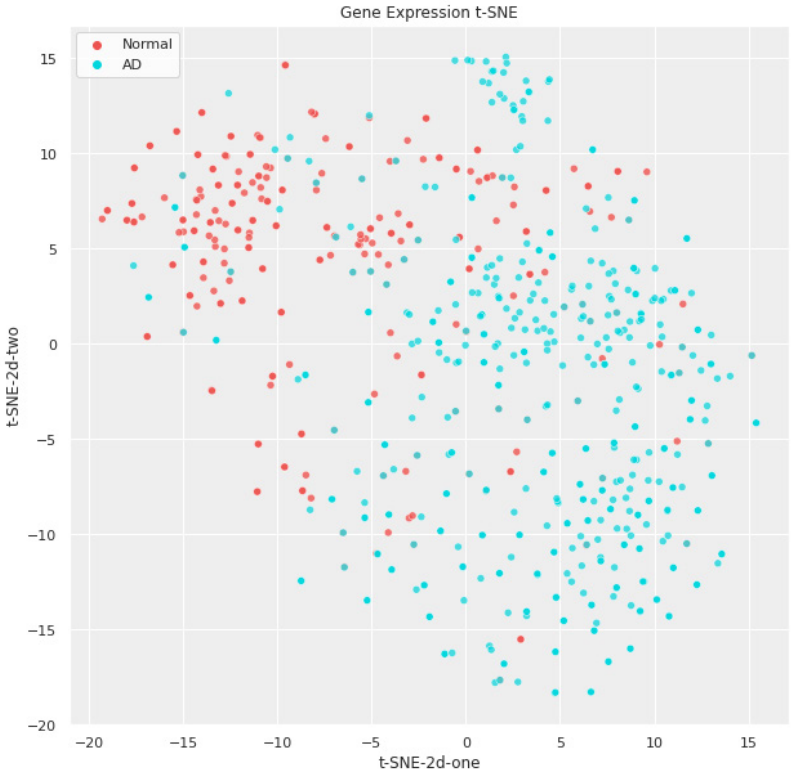

Figure 2. t-SNE plot for Genetic Dataset

KS test and FDR adjusted by Benjamini-Hochberg Procedure

To further reduce the dimension size of the dataset, we separated features based on statistical significance with KS test and FDR. The resulting list for the gene expression dataset contained 177 significant gene features. From the list of significant genes generated, 20 genes with the smallest p-values and the highest stats score were selected. As the example in figure 4 shows, these genes were graphed using histograms comparing gene expression levels between normal and AD patients using matplotlib (Hunter, 2007). FAM131A, a gene associated with Severe Congenital Neutropenia 4 and Uterine Body Mixed Cancer, showed the greatest separation in expression between AD and normal patients (Genecard, n.d.-a).

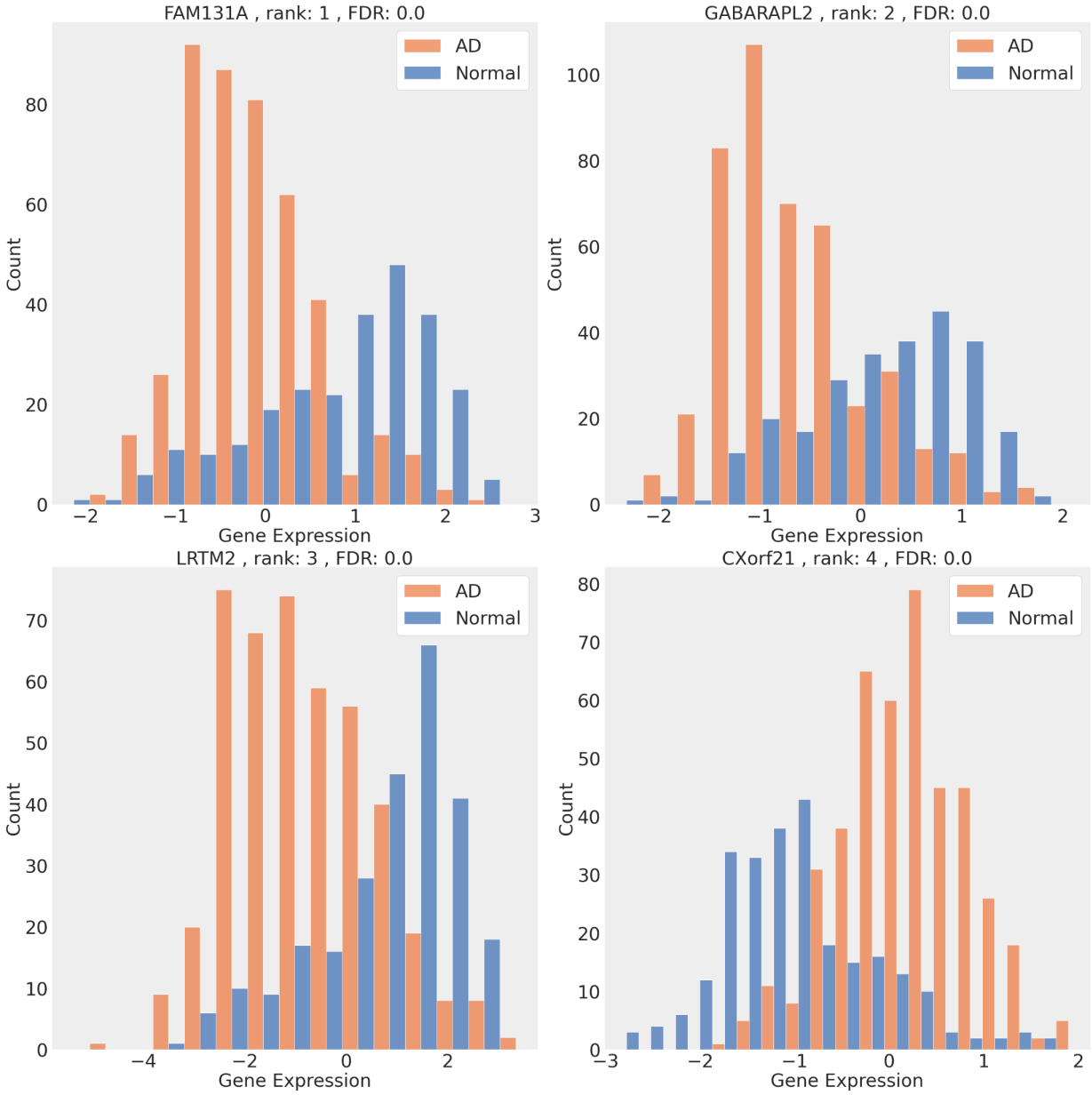

Figure 3. Genetics Dataset Features Histogram, An example of 4 out of the 20 most significant genes was graphed based on gene expression count in a histogram format.

With each statistical analysis step performed, the top 20 most significant genes were collected and analyzed through GeneCard for their attributed functions and related disorders. A sample of this step is shown in Table 1.

Table 1. A compiled list of the top 10 of 20 genes based on the lost p-value and greatest associated variance.

Rank:	Gene Name	Function	Disorders
1	FAM131A	Platelet count and regulation	Severe Congenital Neutropenia 4 Uterine Body Mixed Cancer.
2	GABARAPL2	Golgi traffic Autophagy Mitophagy	Neuronal Ceroid Lipofuscinosis Granulomatous Amebic Encephalitis
3	LRTM2	Heparin and Roundabout binding. Axon guidance Synapse assembly.	Spondylometaphyseal Dysplasia Axial and Preretinal Fibrosis.
4	CXorf21	Innate immune response Toll-like Receptor signaling Lysosomal lumen pH.	Systemic Lupus Erythematosus
5	SPEF1	Filopodia assembly Lamellipodium assembly Negative regulation of cell death.	Hydrocephalus Syndrome 1
6	SYK	Coupling activated immunoreceptors Phagocytosis Epithelial cell growth Tumor suppressor	Immunodeficiency 82 With Systemic Inflammation Arthritis
7	MST150 (SMIM3)	Identical protein binding activity	N/A
8	ZNF544	DNA-binding transcription activator activity, RNA polymerase activity Regulation of transcription	Attention Deficit Hyperactivity Disorder
9	APOL1	Form cholesterol esters Lipid exchange and transport Cholesterol removal	Focal Segmental Glomerulosclerosis 4 Glomerulonephritis
10	RNF135	Protein-protein Protein-DNA interactions	Overgrowth-Macrocephaly-Facial Dysmorphism Syndrome Autism Spectrum Disorder. Covid-19 infection

Note: The data for each gene regarding its function and related disorders are from Genecards, GeneCards – the human gene database, www.genecards.org, Safran M, Rosen N, Twik M, BarShir R, Iny Stein T, Dahary D, Fishilevich S, and Lancet D. The GeneCards Suite Chapter, Practical Guide to Life Science Databases (2022) pp 27-56 [PDF].

Machine Learning Models

We used logistic regression, Random Forest, and DNN to test the effectiveness of the selected features after dimension reduction in successfully diagnosing AD. As the methylation data yielded no significant features, only the genetics dataset was used for this process. Our study uses three models to test the selected features to confirm no one model outperforms on random. The three models are also commonly employed in machine learning bioinformatics and thus would effectively replicate the machine learning conditions of a related experiment (Kong & Yu, 2018) (Inza et al., 2009).

Logistic Regression

Logistic regression is commonly cited as the baseline of the other types of machine learning algorithms due to its simplicity of understanding (Jochen, 2021). Thus, our study used logistic regression as a baseline model to test the effectiveness of the preprocessed dataset. After training the logistic regression model with the genetic dataset filtered for the 20 most significant genes, the maximum cross-validation accuracy received was 0.85 and a minimum of 0.82, with a mean accuracy of 0.841 at a standard deviation of 0.11. The logistic regression was subsequently investigated with the use of a ROC curve. As shown in figure 4, the curve returned an AUC-ROC of 0.83.

Random Forest

After training and testing through the 5-fold cross validation of the gene dataset, the Random Forest models had a maximum cross-validation accuracy of 0.863, minimum of 0.813, mean of 0.838, and standard deviation of 0.017.

The Random Forest trained models were then subsequently graphed using ROC. As shown in figure 4, the resulting AUC-ROC was measured to be 0.86, greater than the mean cross-validation accuracy of 0.838 but less than the maximum cross-validation accuracy of 0.863.

Deep Neural Network (DNN)

As outlined in the methods, the genetics dataset is separated into test and training groups by the partition-TrainTest_ML_for_CV_DNN package provided by the aforementioned study's Github. Each model's dataset was partitioned through the use of 5-fold cross validation to create an additional validation set based on an unused split dataset. During the training and testing phase, batch size of 256 with 100 epochs per generation. After training, the models were evaluated based on cross-validation accuracy across a validation dataset. The maximum validation accuracy of the DNN models produced through this replication was 0.850, with a minimum of 0.791 and an average of 0.8232. The trained models were plotted with a ROC curve to determine their AUC measurement and determine the models' effectiveness. The maximum obtained AUC for the ROC curve of the DNN models was an area of 0.940: the highest out of any other model.

Performance Comparison

As shown in Figure 4, upon examination of each model together, we found that on the AUC-ROC, DNN was the best algorithm. Overall, averaging across each model's results concluded a mean validation accuracy of around 0.834 and an average AUC of 0.877. These results support that the genes determined based on the statistical analysis with PCA, t-SNE, and KS test FDR show a level of significance and some predictive value in correlation with AD to be explored further (Barkved, 2022) (Google, 2019).

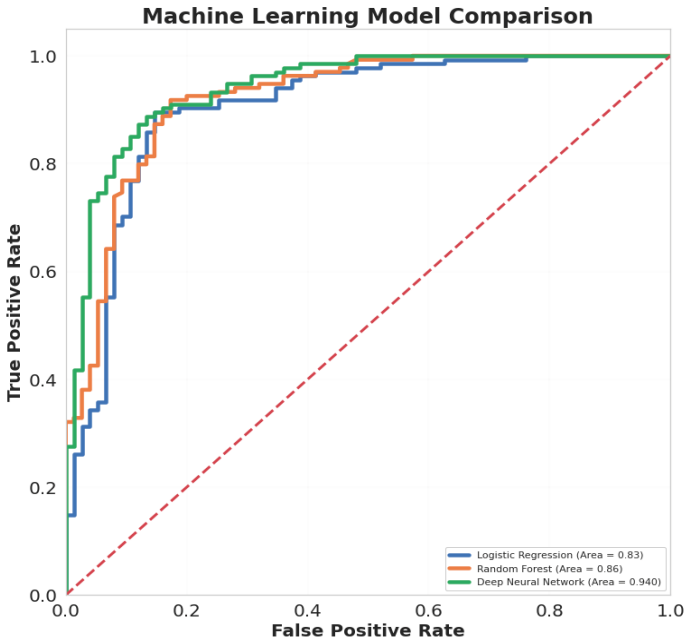

Figure 4. Combined ROC graphs between Logistic Regression, Random Forest Plot, and DNN.

Discussion

The top 20 genes were analyzed through GeneCard for their individual functions, disorders, and potential links with Alzheimer's. Of the 20 genes listed, previous studies concerning AD prognosis and genetics identified eight genes as having a strong correlation to AD.

GABARAPL2 (ATG8) was strongly involved in vesicle elongation and autophagosome assembly. This is hypothesized to interact with the p62 protein in forming LC3-bound autophagosome membranes associated with AD (Caberlotto et al., 2019) (Weidberg et al., 2011).

LRTM2 was confirmed as a potential biomarker of AD using LASSO logistic regression in a previous study (Yu et al., 2021). Furthermore, a study in Biorxiv by Chanabasayya Vastrad documented an upregulation of the gene in AD patients (Vastrad & Vastrad, 2021).

SYK has been implicated in AD due to its effects on tau hyperphosphorylation and the regulation of beta-amyloid production and clearance through its effects on the blood-brain barrier. Transgenic mice overexpressing SYK developed excess amyloid beta accumulation while inhibiting SYK-reduced beta-amyloid levels and tau hyperphosphorylation (Paris et al., 2014). In another study, SYK is known to be recruited by stress granules in microglial cells, which promote inflammation and are associated with AD (Ghosh & Geahlen, 2015).

TEP1, also known as Telomerase Associated Protein 1, encodes for telomerase that repairs the cell's shortening of telomeres from cell division (Harrington et al., 1997). TEP1 has been shown to prevent apoptosis and thus prevent neurodegeneration associated with AD (Zhu, 2001). TEP1 and longer telomeres have also been associated with AD due to having a positive correlation with APOE epsilon 4 (Wikgren, 2010).

NKIRAS1 is known to regulate NF-κB activity (Genecard, n.d.-c). The NF-κB pathway is associated with AD due to its mediation of brain inflammation (Feng et al., 2017). Stimuli activate NF-κB, which in turn regulates expressions of isoforms of SET, which is directly implicated in AD pathogenesis. Sirtuin deacetylates, which down-regulate NF-κB, have also been shown to reduce the effects of aging and AD progression (Natoli, 2009).

ABCC12, also known as ATP Binding Cassette Subfamily C Member 12, is associated with the ATP Binding cassette (ABC) transports which mediate transport across cellular membranes. Studies have implicated ABC's role in AD due to their role in detoxification and neuroprotection of brains (Pereira et al., 2017). ABCC1, in the same subfamily as ABCC12, has been implicated in amyloid beta transport. Mice with knockouts of ABCC1 had up to a 14-fold increase in amyloid-beta levels, and studies have shown ABCC1 activation in APPPS1, AD mice can reduce alpha beta levels by up to 80% (Aykac & Sehirli, 2021) (Krohn et al., 2011).

TGM2, also known as transglutaminase type 2, is found to be highly expressed in AD due to the formation of MAM (mitochondria-associated ER membrane) under conditions of high glucose concentration (D'Eletto et al., 2018). Significant increases in MAM function have been reported in AD patients (Area-Gomez et al., 2012) and implicated in calcium abnormalities associated with AD pathogenesis (Bellenguez et al., 2022). TGM2 reduction through the use of urolithin A and other substances has already been proposed to prevent DM-associated AD by reducing MAM and mitochondrial calcium (Lee et al., 2021).

NDEL1 regulates neural stem cell apoptosis, differentiation, and proliferation rate (Zhang et al., 2022). NDEL1 is also shown to have interactions with miR-103-3p, which are theorized to play a role in AD due to its suppression of cells in AD patients (Yang et al., 2018).

Table 2. Summarized table between associated functions and genes connected with AD pathogenesis or causation.

MAM (Mitochondrial) Interactions	Brain Inflammation	Cell Death	Tau Phosphorylation	Amyloid Beta Proteins	Interaction with confirmed AD genes
TGM2	NKIRAS1 SYK	GABARAPL2 SYK NKIRAS1 ABCC12 NDEL1	SYK	SYK ABCC12	TEP1

As shown in table 2, many of the significant genes share functions. This could imply that the same mechanism or pathway in the progression of AD could result from multiple different genetic and environmental pathways. The prominence in cell death relation with the genes identified is commonly associated with beta-amyloid build-up and tau phosphorylation, further promoting the notion of a single mechanism through multiple means of causation (Carter & Lippa, 2001).

Conclusion

In this study, we proposed using a multi-omics dataset to identify novel genes and methylations that significantly contribute to AD. We subsequently proposed the use of statistical analysis such as PCA, TSNE, and KS test-FDR, to reduce the dimension of the dataset and isolate significant features. Due to the insignificant variance found by statistical analysis on the methylation dataset, the methylation dataset was removed from further analysis. To demonstrate the effectiveness of the preprocessed genes, we used three machine learning algorithms logistic regression, random forest, and DNN to train on the selected significant genes and test for accuracy. Our DNN model was the most effective diagnostic model with a reported AUC-ROC of 94.0%. The 8 genes we've identified: TGM2, NKIRAS1, SYK, GABARAPL2, ABCC12, NDEL1, and TEP1, all show promise as alternative genes to be addressed by AD treatments.

The limitation of our study is that we only used one multi-omics data set with limited sample sizes, gene count, and methylation. As a result, many potential genes and correlations that exist would be lost within the possibility of overfitting or simple lack of data. The dropping of methylation data further limited the significance of the results of this study due to loss of the possible interactions between the datasets. To overcome these limitations, the 5-fold cross validation and cross validation process was employed to reduce overfitting and maximize the effectiveness of the given genetic dataset.

Omics models combined with machine learning serve an important advantage of being capable of quickly identifying novel biomarkers or drugs for the treatment of AD (Sancesario & Bernardini, 2018). Thus, in the future, we propose further studies to determine the mechanisms behind the correlations proposed by this paper between the extracted genes and AD, particularly in the form of experimentation through the possible use of gene therapy or other epidemiological study techniques. Furthermore, we intend to use a wider breadth of the multi-omics datasets with more patients and features for a more in-depth analysis of the causations of AD and to further test our study's AD diagnostic models.

Acknowledgments

This research was mentored and supported by Dr. Hayan Lee, post doctorate of the Snyder Lab at Stanford University. All above packages were suggested and employed under her tutelage for the results produced.

References

Alzheimer's Association. (2021). 2021 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *17*(3). https://doi.org/10.1002/alz.12328

Area-Gomez, E., del Carmen Lara Castillo, M., Tambini, M. D., Guardia-Laguarta, C., de Groof, A. J. C., Madra, M., Ikenouchi, J., Umeda, M., Bird, T. D., Sturley, S. L., & Schon, E. A. (2012). Upregulated function of mitochondria-associated ER membranes in Alzheimer disease. *The EMBO Journal*, *31*(21), 4106–4123. https://doi.org/10.1038/emboj.2012.202

Aykac, A., & Sehirli, A. Ö. (2021). The Function and Expression of ATP-Binding Cassette Transporters Proteins in the Alzheimer's Disease. *Global Medical Genetics*, *08*(04), 149–155. https://doi.org/10.1055/s-0041-1735541

Barkved, K. (2022, March 9). *How To Know if Your Machine Learning Model Has Good Performance | Obviously AI*. Www.obviously.ai. https://www.obviously.ai/post/machine-learning-model-performance#:~:text=But%20in%20our%20opinion%2C%20anything

Battineni, G., Chintalapudi, N., Amenta, F., & Traini, E. (2020). A Comprehensive Machine-Learning Model Applied to Magnetic Resonance Imaging (MRI) to Predict Alzheimer's Disease (AD) in Older Subjects. *Journal of Clinical Medicine*, *9*(7), 2146. https://doi.org/10.3390/jcm9072146

Bekris, L. M., Yu, C.-E., Bird, T. D., & Tsuang, D. W. (2010). Review Article: Genetics of Alzheimer Disease. *Journal of Geriatric Psychiatry and Neurology*, *23*(4), 213–227. https://doi.org/10.1177/0891988710383571

Bellenguez, C., Küçükali, F., Jansen, I. E., Kleineidam, L., Moreno-Grau, S., Amin, N., Naj, A. C., Campos-Martin, R., Grenier-Boley, B., Andrade, V., Holmans, P. A., Boland, A., Damotte, V., van der Lee, S. J., Costa, M. R., Kuulasmaa, T., Yang, Q., de Rojas, I., Bis, J. C., & Yaqub, A. (2022). New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature Genetics*. https://doi.org/10.1038/s41588-022-01024-z

Braak, H., & Braak, E. (1997). Frequency of Stages of Alzheimer-Related Lesions in Different Age Categories. *Neurobiology of Aging*, *18*(4), 351–357. https://doi.org/10.1016/s0197-4580(97)00056-0

Brickell, K. L., Steinbart, E. J., Rumbaugh, M., Payami, H., Schellenberg, G. D., Van Deerlin, V., Yuan, W., & Bird, T. D. (2006). Early-Onset Alzheimer Disease in Families With Late-Onset Alzheimer Disease. *Archives of Neurology*, *63*(9), 1307. https://doi.org/10.1001/archneur.63.9.1307

Caberlotto, L., Nguyen, T.-P., Lauria, M., Priami, C., Rimondini, R., Maioli, S., Cedazo-Minguez, A., Sita, G., Morroni, F., Corsi, M., & Carboni, L. (2019). Cross-disease analysis of Alzheimer's disease and type-2 Diabetes highlights the role of autophagy in the pathophysiology of two highly comorbid diseases. *Scientific Reports*, *9*(1), 3965. https://doi.org/10.1038/s41598-019-39828-5

Campion, D., Dumanchin, C., Hannequin, D., Dubois, B., Belliard, S., Puel, M., Thomas-Anterion, C., Michon, A., Martin, C., Charbonnier, F., Raux, G., Camuzat, A., Penet, C., Mesnage, V., Martinez, M., Clerget-Darpoux, F., Brice, A., & Frebourg, T. (1999). Early-Onset Autosomal Dominant Alzheimer Disease: Prevalence, Genetic Heterogeneity, and Mutation Spectrum. *The American Journal of Human Genetics*, *65*(3), 664–670. https://doi.org/10.1086/302553

Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., Bennett, C., Hawken, S., McInnes, M., Magwood, O., Sheikh, Y., & Holzinger, A. (2022). Deep ROC Analysis and AUC as Balanced Average Accuracy to Improve Model Selection, Understanding and Interpretation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–1. https://doi.org/10.1109/TPAMI.2022.3145392

Carter, J., & Lippa, C. (2001). β-Amyloid, Neuronal Death and Alzheimers Disease. *Current Molecular Medicine*, *1*(6), 733–737. https://doi.org/10.2174/1566524013363177

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2264-5

D'Eletto, M., Rossin, F., Occhigrossi, L., Farrace, M. G., Faccenda, D., Desai, R., Marchi, S., Refolo, G., Falasca, L., Antonioli, M., Ciccosanti, F., Fimia, G. M., Pinton, P., Campanella, M., & Piacentini, M. (2018). Transglutaminase Type 2 Regulates ER-Mitochondria Contact Sites by Interacting with GRP75. *Cell Reports*, *25*(13), 3573-3581.e4. https://doi.org/10.1016/j.celrep.2018.11.094

Zhang B, Gaiteri C, Bodea LG, Wang Z et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell 2013 Apr 25;153(3):707-20. PMID: 23622250

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., & Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, *11*(1). https://doi.org/10.1186/1471-2105-11-587

Feng, Y., Li, X., Zhou, W., Lou, D., Huang, D., Li, Y., Kang, Y., Xiang, Y., Li, T., Zhou, W., & Song, W. (2017). Regulation of SET Gene Expression by NFkB. *Molecular Neurobiology*, *54*(6), 4477–4485. https://doi.org/10.1007/s12035-016-9967-2

Genecard. (n.d.-a). *FAM131A Gene - GeneCards | F131A Protein | F131A Antibody*. Www.genecards.org. Retrieved August 3, 2022, from https://www.genecards.org/cgi-bin/carddisp.pl?gene=FAM131A#diseases

Genecard. (n.d.-b). *FAM234B Gene - GeneCards | F234B Protein | F234B Antibody*. Www.genecards.org. https://www.genecards.org/cgi-bin/carddisp.pl?gene=FAM234B&keywords=KIAA1467#diseases

Genecard. (n.d.-c). *NKIRAS1 Gene - GeneCards | KBRS1 Protein | KBRS1 Antibody*. Www.genecards.org. Retrieved August 3, 2022, from https://www.genecards.org/cgi-bin/carddisp.pl?gene=NKIRAS1

Ghosh, S., & Geahlen, R. L. (2015). Stress Granules Modulate SYK to Cause Microglial Cell Dysfunction in Alzheimer's Disease. *EBioMedicine*, *2*(11), 1785–1798. https://doi.org/10.1016/j.ebiom.2015.09.053

Goedert, M., & Spillantini, M. G. (2006). A century of Alzheimer's disease. *Science (New York, N.Y.)*, *314*(5800), 777–781. https://doi.org/10.1126/science.1132814

Google. (2019). *Classification: Accuracy | Machine Learning Crash Course*. Google Developers. https://developers.google.com/machine-learning/crash-course/classification/accuracy

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian Journal of Internal Medicine, 4(2), 627–635. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/

Harrington, L., McPhail, T., Mar, V., Zhou, W., Oulton, R., Program, A. E., Bass, M. B., Arruda, I., & Robinson, M. O. (1997). A Mammalian Telomerase-Associated Protein. *Science*, *275*(5302), 973–977. https://doi.org/10.1126/science.275.5302.973

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, *18*(1). https://doi.org/10.1186/s13059-017-1215-1

IBM. (n.d.). *What is Logistic regression? | IBM*. Www.ibm.com. https://www.ibm.com/topics/logistic-regression#:~:text=Logistic%20regression%20estimates%20the%20probability

Inza, I., Calvo, B., Armañanzas, R., Bengoetxea, E., Larrañaga, P., & Lozano, J. A. (2009). Machine Learning: An Indispensable Tool in Bioinformatics. *Methods in Molecular Biology*, 25–48. https://doi.org/10.1007/978-1-60327-194-3_2

Iwatsubo, T., Odaka, A., Suzuki, N., Mizusawa, H., Nukina, N., & Ihara, Y. (1994). Visualization of Aβ42(43) and Aβ40 in senile plaques with end-specific Aβ monoclonals: Evidence that an initially deposited species is Aβ42(43). *Neuron*, *13*(1), 45–53. https://doi.org/10.1016/0896-6273(94)90458-8

J. D. Hunter, "Matplotlib: A 2D Graphics Environment," in Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, May-June 2007, doi: 10.1109/MCSE.2007.55.

Johnson, P., Vandewater, L., Wilson, W., Maruff, P., Savage, G., Graham, P., Macaulay, L. S., Ellis, K. A., Szoeke, C., Martins, R. N., Rowe, C. C., Masters, C. L., Ames, D., & Zhang, P. (2014). Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease. *BMC Bioinformatics*, *15*(Suppl 16), S11. https://doi.org/10.1186/1471-2105-15-s16-s11

Kaitlin, Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Science Review*, *1*(3), 9. https://scholar.smu.edu/cgi/viewcontent.cgi?article=1041&context=datasciencereview#:~:text=variables%20exceeds%20the%20number%20of

Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-13056-x

Kong, Y., & Yu, T. (2018). A Deep Neural Network Model using Random Forest to Extract Feature Representation for Gene Expression Data Classification. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-34833-6

Krohn, M., Lange, C., Hofrichter, J., Scheffler, K., Stenzel, J., Steffen, J., Schumacher, T., Brüning, T., Plath, A.-S., Alfen, F., Schmidt, A., Winter, F., Rateitschak, K., Wree, A., Gsponer, J., Walker, L. C., & Pahnke, J. (2011). Cerebral amyloid-β proteostasis is regulated by the membrane transport protein ABCC1 in mice. *Journal of Clinical Investigation*, *121*(10), 3924–3931. https://doi.org/10.1172/jci57867

Lee, H. J., Jung, Y. H., Choi, G. E., Kim, J. S., Chae, C. W., Lim, J. R., Kim, S. Y., Yoon, J. H., Cho, J. H., Lee, S.-J., & Han, H. J. (2021). Urolithin A suppresses high glucose-induced neuronal amyloidogenesis by modulating TGM2-dependent ER-mitochondria contacts and calcium homeostasis. *Cell Death & Differentiation*, *28*(1), 184–202. https://doi.org/10.1038/s41418-020-0593-1

Mark Schmidt, Nicolas Le Roux, Francis Bach. Minimizing Finite Sums with the Stochastic Average Gradient. Mathematical Programming, Springer Verlag, 2017, 162 (1-2), pp.83-112. ff10.1007/s10107- 016-1030-6ff. Ffhal-00860051v2f

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke,

Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

Ma, S., & Dai, Y. (2011). Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics*, *12*(6), 714–722. https://doi.org/10.1093/bib/bbq090

Narayanan M, Huynh JL, Wang K, Yang X et al. Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases. Mol Syst Biol 2014 Jul 30;10:743. PMID: 25080494

National Institute on Aging. (2017, May 16). *What Happens to the Brain in Alzheimer's Disease?* National Institute on Aging. https://www.nia.nih.gov/health/what-happens-brain-alzheimers-disease#:~:text=These%20tangles%20block%20the%20neuron

National Institute on Aging. (2021, July 8). *Alzheimer's Disease Fact Sheet*. National Institute on Aging. https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet

Natoli, G. (2009). When Sirtuins and NF-κB Collide. *Cell*, *136*(1), 19–21. https://doi.org/10.1016/j.cell.2008.12.034

Paris, D., Ait-Ghezala, G., Bachmeier, C., Laco, G., Beaulieu-Abdelahad, D., Lin, Y., Jin, C., Crawford, F., & Mullan, M. (2014). The Spleen Tyrosine Kinase (Syk) Regulates Alzheimer Amyloid-β Production and Tau Hyperphosphorylation. *The Journal of Biological Chemistry*, *289*(49), 33927–33944. https://doi.org/10.1074/jbc.M114.608091

Park, C. (2021, March 20). *DNN_for_ADprediction/dataset at master · ChihyunPark/DNN_for_ADprediction*. GitHub. https://github.com/ChihyunPark/DNN_for_ADprediction/tree/master/dataset

Park, C., Ha, J., & Park, S. (2020). Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. *Expert Systems with Applications*, *140*, 112873. https://doi.org/10.1016/j.eswa.2019.112873

Piller, C. (2022, July 21). Potential fabrication in research images threatens key theory of Alzheimer's disease. Www.science.org. https://www.science.org/content/article/potential-fabrication-research-images-threatens-key-theory-alzheimers-disease

Pereira, C. D., Martins, F., Wiltfang, J., da Cruz e Silva, O. A. B., & Rebelo, S. (2017). ABC Transporters Are Key Players in Alzheimer's Disease. *Journal of Alzheimer's Disease*, *61*(2), 463–485. https://doi.org/10.3233/jad-170639

Plotly. (n.d.). *Plotly Python Graphing Library*. Plotly.com. https://plotly.com/python/

Rogers, A., & Weiss, S. (2017). *False Discovery Rate - an overview | ScienceDirect Topics*. Www.sciencedirect.com. https://www.sciencedirect.com/topics/neuroscience/false-discovery-rate

Sancesario, G. M., & Bernardini, S. (2018). Alzheimer's disease in the omics era. *Clinical Biochemistry*, *59*, 9–16. https://doi.org/10.1016/j.clinbiochem.2018.06.011

Scipy. (n.d.). *scipy.stats.ks_2samp — SciPy v1.9.0 Manual*. Docs.scipy.org. Retrieved August 2, 2022, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html#scipy.stats.ks_2samp

Sklearn. (2014). *sklearn.manifold.TSNE — scikit-learn 0.21.3 documentation*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

Sklearn. (2018). *3.2.4.3.2. sklearn.ensemble.RandomForestRegressor — scikit-learn 0.20.3 documentation*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

sklearn.decomposition.PCA — scikit-learn 0.20.3 documentation. (2009). Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Smith RG, Hannon E, De Jager PL, Chibnik L et al. Elevated DNA methylation across a 48-kb region spanning the HOXA gene cluster is associated with Alzheimer's disease neuropathology. Alzheimers Dement 2018 Dec;14(12):1580-1588. PMID: 29550519

Snoek, J., Larochelle, H., & Adams, R. P. (2012). *Practical Bayesian Optimization of Machine Learning Algorithms*. Neural Information Processing Systems; Curran Associates, Inc. https://papers.nips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html

Statsmodels. (2019). *statsmodels.stats.multitest.fdrcorrection — statsmodels*. Www.statsmodels.org. https://www.statsmodels.org/stable/generated/statsmodels.stats.multitest.fdrcorrection.html

Su, Q., Wang, Y., Jiang, X., Chen, F., & Lu, W. (2017). A Cancer Gene Selection Algorithm Based on the K-S Test and CFS. *BioMed Research International*, *2017*, 1–6. https://doi.org/10.1155/2017/1645619

Tabe-Bordbar, S., Emad, A., Zhao, S. D., & Sinha, S. (2018). A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models. *Scientific Reports*, *8*. https://doi.org/10.1038/s41598-018-24937-4

University of Southern California. (2006, February 7). *Alzheimer's Found To Be Mostly Genetic: Largest Twin Study Ever Undertaken Confirms Highest Estimates Of Genetic Risk*. ScienceDaily. http://www.sciencedaily.com/releases/2006/02/060206232300.htm

van Driel, M. A., & Brunner, H. G. (2006). Bioinformatics methods for identifying candidate disease genes. *Human Genomics*, *2*(6), 429. https://doi.org/10.1186/1479-7364-2-6-429

Vastrad, B., & Vastrad, C. (2021). *Bioinformatics analyses of significant genes, related pathways and candidate prognostic biomarkers in Alzheimer's disease*. https://doi.org/10.1101/2021.05.06.442918

Weidberg, H., Shvets, E., & Elazar, Z. (2011). Biogenesis and Cargo Selectivity of Autophagosomes. *Annual Review of Biochemistry*, *80*(1), 125–156. https://doi.org/10.1146/annurev-biochem-052709-094552

Wikgren, M. et al. APOE epsilon4 is associated with longer telomeres, and longer telomeres among epsilon4 carriers predicts worse episodic memory. Neurobiol. Aging (2010). doi:10.1016/j.neurobiolaging.2010.03.004

Wilhelm, Jochen. (2021). Re: Can logistic regression be used as the initial baseline or something to start with for any data classification system?. Retrieved from: https://www.researchgate.net/post/Can_logistic_regression_be_used_as_the_initial_baseline_or_something_to_start_with_for_any_data_classification_system/60fc0ec263ef9768526143fe/citation/download.

Yang, H., Wang, H., Shu, Y., & Li, X. (2018). miR-103 Promotes Neurite Outgrowth and Suppresses Cells Apoptosis by Targeting Prostaglandin-Endoperoxide Synthase 2 in Cellular Models of Alzheimer's Disease. *Frontiers in Cellular Neuroscience*, *12*, 91. https://doi.org/10.3389/fncel.2018.00091

Yu, W., Yu, W., Yang, Y., & Lü, Y. (2021). Exploring the Key Genes and Identification of Potential Diagnosis Biomarkers in Alzheimer's Disease Using Bioinformatics Analysis. *Frontiers in Aging Neuroscience*, *13*. https://doi.org/10.3389/fnagi.2021.602781

Zhang, X.-H., Jin, G.-H., Li, W., Wang, S.-S., Shan, B.-Q., Qin, J.-B., Zhao, H.-Y., Tian, M.-L., He, H., & Cheng, X. (2022). miR-103-3p targets Ndel1 to regulate neural stem cell proliferation and differentiation. *Neural Regeneration Research*, *17*(2), 401. https://doi.org/10.4103/1673-5374.317987

Zhu, H., Fu, W. & Mattson, M. P. The Catalytic Subunit of Telomerase Protects Neurons Against Amyloid β-Peptide-Induced Apoptosis. J. Neurochem. 75, 117–124 (2001).Google Scholar