

Hybrid Feature Selection Method for Predicting Alzheimer's Disease Using Gene Expression Data

Aliaa El-Gawady^{1,*}, BenBella S. Tawfik¹ and Mohamed A. Makhlof^{1,2}

¹Department of Information Systems, Faculty of Computers and Informatics, Suez Canal University, Ismailia, 41522, Egypt

²Faculty of Computer Science, Nahda University, Beni Suef, Egypt

*Corresponding Author: Aliaa El-Gawady. Email: alia_saad@ci.suez.edu.eg

Received: 26 July 2022; Accepted: 03 October 2022

Abstract: Gene expression (GE) classification is a research trend as it has been used to diagnose and prognosis many diseases. Employing machine learning (ML) in the prediction of many diseases based on GE data has been a flourishing research area. However, some diseases, like Alzheimer's disease (AD), have not received considerable attention, probably owing to data scarcity obstacles. In this work, we shed light on the prediction of AD from GE data accurately using ML. Our approach consists of four phases: preprocessing, gene selection (GS), classification, and performance validation. In the preprocessing phase, gene columns are preprocessed identically. In the GS phase, a hybrid filtering method and embedded method are used. In the classification phase, three ML models are implemented using the bare minimum of the chosen genes obtained from the previous phase. The final phase is to validate the performance of these classifiers using different metrics. The crux of this article is to select the most informative genes from the hybrid method, and the best ML technique to predict AD using this minimal set of genes. Five different datasets are used to achieve our goal. We predict AD with impressive values for MultiLayer Perceptron (MLP) classifier which has the best performance metrics in four datasets, and the Support Vector Machine (SVM) achieves the highest performance values in only one dataset. We assessed the classifiers using seven metrics; and received impressive results, allowing for a credible performance rating. The metrics values we obtain in our study lie in the range [.97, .99] for the accuracy (Acc), [.97, .99] for F1-score, [.94, .98] for kappa index, [.97, .99] for area under curve (AUC), [.95, 1] for precision, [.98, .99] for sensitivity (recall), and [.98, 1] for specificity. With these results, the proposed approach outperforms recent interesting results. With these results, the proposed approach outperforms recent interesting results.

Keywords: Gene expression; gene selection; machine learning; classification; Alzheimer's disease



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1 Introduction

Dementia and memory loss are commonly caused by AD. It seems that in the mid of 1906, AD is first recognized. It has been discovered as the principal reason for death. AD begins slowly and slowly exacerbates over time [1]. By 2050, approximately 152 million people worldwide will be detected with AD [2]. It is expected that [3] one person out of every eighty-five will suffer from AD by 2050. AD has various symptoms; the most prevalent is the inability of remembering recent occurrences. Difficulties with self-motivation, language, orientation, memory, self-care, mood, and behavior are some indications of advanced AD [4]. As the status of AD patients degenerates, they start to keep away from their families and society. Progressively, Body functions are lost, which finally results in death. Despite progress might happen quickly or slowly, the average life expectancy after an AD diagnosis is 3 to 9 years [5]. As a result, early identification of AD has the potential to save lives, which is where the current research comes into play.

So far, most of the recent studies on AD diagnosis have been performed using neuropsychological testing and brain magnetic resonance imaging (MRI). Because it was challenging to sample the posterior brains of normal and AD individuals, molecular recognition of AD is inadequate. Recent trials have thankfully yielded large-scale omics data for numerous brain locations. Using this data, it is now quite simple to construct prediction approaches, like those described in this study, in which ML techniques are used to diagnose AD earliest opportunity. These treatments can be advantageous for the patient because they are simple and affordable. In certain cases, they have even been demonstrated to predict AD better than physicians [6]. This fact has prompted a lot of studies into utilizing ML to diagnose AD using medical data in various ways, such as MRI.

In the following articles, the authors used MRI in the diagnosis of AD: an approach based on SVM with recursive feature elimination to choose the informative features is employed in [7]. They allege to have better performance in classifying normal (N), mild cognitive impairment (MCI), and AD cases (subjects or instances). The authors in [8] used MRI images with an unsupervised convolutional neural network (CNN) for AD feature extraction. In addition, they classify AD, MCI, and N by employing k-means clustering. The output achieves an accuracy of 95.52% for AD vs. MCI and 90.63% for MCI vs. N in one slice of data. For TOP data, the method achieves an accuracy of 97.01% for AD vs. MCI, and an accuracy of 92.6% for MCI vs. N. Techniques of the type Resting-state functional MRI and deep learning (DL) have been used in [9] to classify AD and its stages. To apply transfer learning with and without fine-tuning, they used an expanded network architecture. While off-the-shelf and fine-tuned models, they produce impressive outcomes with high accuracy. The study in [10] classified AD and MCI cases from N cases using CNN. They investigate the effects of including information from diffusion tensor imaging and MRI. They achieved AUC, specificity, and Acc results of 0.93%, 91.7%, and 88.9%, respectively, for N vs. AD classification, and 0.68%, 81.8%, and 71.1%, respectively, for N vs. MCI classification.

Interesting results have been obtained in [11], where MRI is used to classify stable MCI, converter MCI, N, and AD cases. For dimensionality reduction, they used partial least squares and Analysis of variance (ANOVA) for features selection. For classification, random forest (RF) is used, obtaining an accuracy of 56.25% for a test set of 160 cases. In [12], mobility data and DL models are used to identify the stage of AD patients by processing a time series for each patient. In addition, a CNN model is used to identify the patterns which detect AD stages. The results outperform those of traditional supervised learning models. In [13] the authors used MRI images to check the presence of AD and to assess the degree of AD. An evolutionary algorithm is applied to extract specific information for AD classification. Two experiments are executed; the first one is to classify images into two classes

moderate demented and non-demented. The second experiment is to classify images into three classes moderate demented, mild demented, and non-demented. Their approach achieved 100% and 100% for Acc and F1-score, respectively for the first experiment. For the second experiment, they achieved 91.49%, and 91.49% for Acc and F1-score, respectively.

In [14] ML framework is used to automatically diagnose neurodegenerative diseases, focusing on AD and frontotemporal dementia. The authors suggested an explainable artificial intelligence tool to assist clinicians in diagnosis because it includes all the stages necessary to analyze these datasets, including data preparation, feature selection using an evolutionary approach, and modeling of the diseases listed. It was evident how the suggested framework enables fluid processing of the cognitive and image assessments, with a significant decrease in the number of features required for the diagnosis, and a significant increase in classification accuracy.

GE data-based research enables the diagnosis of numerous diseases through the powerful technology of Deoxyribonucleic Acid (DNA) microarrays [15]. Thousands of genes' expression levels are provided in this data [16]. The gene expression level is signified by some distinct messenger molecules ribonucleic acid (mRNA) in the cell. Utilizing this level, the diseases can be detected and the best treatment options and alterations in other processes can be discovered [17]. In this direction, blood-derived gene expression biomarkers in [18] are used to differentiate AD cases from N cases. XGBoost is used as a classifier and successfully identifies AD by including associated mental and geriatric health issues. However, the model's sensitivity must be improved to create a more accurate blood test for AD. Three datasets, AD Neuroimaging Initiative (ADNI), AddNeuroMed1 (ANM1), and ANM2 have been used in [19] to classify AD from N. To select the most relevant genes, various GS methods like transcription factor, convergent functional genomics (CFG), hub genes, and variational autoencoder have been employed. Five classifiers, SVM, L1-regularized LR (L1-LR), RF, deep neural network (DNN), and logistic regression (LR) have been implemented for classification purposes. They achieved AUC results as follows 87.4% for ANM1, 80.4% for ANM2, and 65.7% for ADNI. Moreover, the blood genes' biological functions relevant to AD have been analyzed and the blood bio-signature has been compared with the brain bio-signature. Several 1291 brain genes have been selected from a GE dataset with 2021 blood genes selected from the other datasets, reporting that 140 genes are common among them.

The authors in [20] identify GE from a blood dataset, exploring the correlation between an AD patient's blood and brain genes. They discover 789 differentially expressed genes in both the brain and blood. The least absolute shrinkage and selection operator (LASSO) regression is utilized as a GS technique. They used logistic ridge regression (RR), RF, and SVM approaches in the classification stage. With 78.1% for Acc, they successfully distinguish AD cases from N cases. To identify prospective diagnostic biomarkers of AD, the authors in [21] used multiple brain regions. GE data from six brain regions are used to identify AD biomarkers. For the selection of the relevant genes, a t-test is employed. To identify biomarkers and gauge their use for clinical diagnosis, Significance tests have been performed. GE and DNA methylation datasets have been integrated in [22] for the prediction of AD using DNN. Moreover, they used t-stochastic nearest neighbor (t-SNE) and principal component analysis (PCA) techniques for gene selection. They verify the effectiveness of their method by comparing its Acc and AUC with those of conventional ML models, like SVM, RF, and Naive Bayes (NB). They achieve 82.3% Acc and 79.7% AUC.

Blood GE data from dementia case registry (DCR) cohorts and the ANM has been used in [23]. For GS, they used recursive feature elimination (RFE) and for the classification task, they used RF.

In addition, they used ANM1 for training the classifier and used for testing integration of ANM2 and DCR. Their methods achieve 65.7% for Acc and 72.4% for AUC.

Table 1 presents a summary of recent studies that have been presented to diagnose AD. This table illustrates the main challenge in analyzing GE data: the imbalance between the number of genes and cases (the number of genes is significantly more than the number of cases). The Table lists the number of genes utilized in each experiment at the beginning and after the GS phase. We conclude that the selection of genes is heavily influenced by the dataset and model and doesn't have any discernible pattern or guideline. In other words, depending on the ML model employed, each experiment may choose a different number for the most significant genes and arrive at a different Acc value.

Table 1: summary of some recent studies using various GS techniques and ML models to predict AD from GE data

GEO ID	Cases number	Genes number	GS method	Selected genes no.	ML model	Performance metrics	Work
GSE5281	N:74, AD:87	23,643	t-test	1,001	SVM	AUC: 0.894	[21]
GSE33000+	N:257, AD:439	19,488	PCA	35	RF	AUC: 0.531, Acc: 0.624	[22]
GSE44770			t-SNE	35	SVM	AUC: 0.511, Acc: 0.632	
GSE63060	N:104, AD:145	7,584	CFG	353	DNN	AUC: 0.874	[19]
GSE63061	N:134, AD:139	6,154	CFG	188	SVM	AUC: 0.804	
ADNI	N:136, AD:63	3,897	CFG	922	DNN	AUC: 0.657	
GSE63060+	N:182, AD:245	16,928	LASSO	3,601	SVM	AUC: 0.859, Acc: 0.781	[20]
GSE63061							
GSE63060+	N:118, AD:118	261	RFE	12	RF	AUC: 0.724, Acc: 0.657	[23]
DCR							

This paper proposes an approach to predict AD based on GE data, consisting of steps. In the beginning, we preprocess each dataset to prepare the datasets for manipulation. Then, we use a filtering method to evaluate a dataset's genes. Next, we rank and select the genes that have the highest values. Then we use an embedded method to select the most relevant and significant genes. Finally, we feed the selected genes into various ML techniques and track their classification results. The technique with the greatest performance is chosen to be used in the AD prediction system in the future, which is our ultimate goal.

The structure of the article follows the following order. In Section 2, we present the materials and methods of the approach used in our investigations. Section 3 is devoted to the experimental work. Section 4 is used for the discussion and concluding remarks.

2 Materials and Methods

This section presents our suggested strategy for GS and AD classification as depicted in [Fig. 1](#). Below, we present the details of the suggested strategy.

2.1 Datasets

The experimental work is applied using five well-known gene expression datasets. The datasets are composed of multiple human brain tissues for DNA microarray data. They are obtained from the National Center for Biotechnology Information-Gene Expression Omnibus (NCBI-GEO) database [24]. The access numbers of the used datasets are GSE5281, GSE118553, GSE132903, GSE48350, and GSE36980. GSE5281 dataset contains 161 cases from various six brain regions, i.e., medial temporal gyrus (MTG), hippocampus (HIP), entorhinal cortex (EC), primary visual cortex (VCX), superior

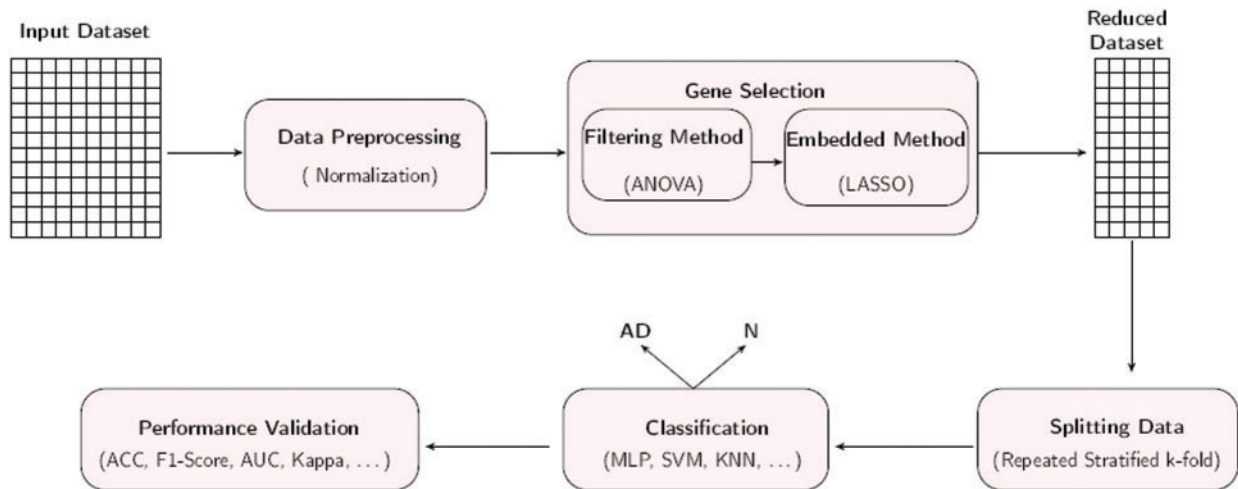


Figure 1: Workflow of the proposed approach for AD prediction

frontal gyrus (SFG), and posterior cingulate (PC). GSE118553 contains 267 cases from four different brain regions, frontal cortex (FC), temporal cortex (TC), EC, and cerebellum (CR). GSE132903 dataset contains 195 cases from the MTG brain region. GSE48350 dataset contains 253 cases from different four brain regions, i.e., EC, HIP, SFG, and post-central gyrus (PCG). GSE36980 dataset contains 80 cases from three different brain regions HIP, TC, and FC. The five datasets are summarized in [Table 2](#).

Table 2: Summary of the five datasets used in the present study

Dataset ID	Brain regions	No. of genes	AD cases	Normal cases	Total no. of cases
GSE5281	EC, HIP, MTG, PC, SFG, and VCX	23,516	87	74	161
GSE118553	EC, TC, FC, and CR	31,331	167	100	267
GSE132903	MTG	31,331	97	98	195
GSE48350	EC, HIP, SFG, and PCG	23,516	79	174	253
GSE36980	HIP, TC, and FC	24,533	33	47	80

2.2 Preprocessing

Preprocessing is essential for handling gene expression data because it prepares the datasets for manipulation. The gene values are normalized to avoid large differences among the genes. The min-max approach is used for normalization, and rescales the values range to the interval [0,1] for each individual gene. Denote by C the set of cases described by G as a set of genes. The normalized gene value $\hat{\Lambda}_{c_i, g_j}$, $c_i \in C$, and $g_j \in G$, of some gene value Λ_{c_i, g_j} is defined by

$$\hat{\Lambda}_{c_i, g_j} = \frac{\Lambda_{c_i, g_j} - \min_{c_k \in C} (\Lambda_{c_i, g_j})}{\max_{c_k \in C} (\Lambda_{c_i, g_j}) - \min_{c_k \in C} (\Lambda_{c_i, g_j})} \quad (1)$$

where $\min_{c_k \in C} (\Lambda_{c_i, g_j})$ and $\max_{c_k \in C} (\Lambda_{c_i, g_j})$ are the minimum and maximum values of gene $g_j \in G$ over all cases $c_k \in C$.

2.3 Gene Selection

GS is considered the most challenging task in the gene expression data analysis because compared to the number of genes the number of cases is substantially smaller. Table 2 illustrates this disparity. It's a crucial step to select genes from the raw gene expression datasets that are relevant to AD prediction. Meanwhile, including insignificant and redundant genes can significantly negatively impact classification accuracy. As a result, we devote great attention to gene selection in our research. There are three forms of gene selection methods: filter, wrapper, and embedded methods. We employ a hybrid method for this phase to take advantage of both the filtering and the embedded methods. Filtering methods are good for a theoretical framework; it helps to understand the structure of the data. It is suitable to be used for larger datasets, as it is fast to be computed. Additionally, the embedded method is also fast and more accurate, considering the interaction of the genes, and is less prone to overfitting. We start by looking at each gene's importance in AD prediction using a filter-based method, ANOVA, and order the genes according to their ability for AD prediction. Then we select the most relevant and informative genes using an embedded method LASSO. Finally, we assess the importance of genes according to each of the three ML classification techniques that we have used. We can detect the most relevant and essential genes as well as the best accurate technique for predicting AD at the end of these two phases.

- **ANOVA-F statistic:**

ANOVA is an efficient tool usually used to gauge the significance of differentiating between the means of two random variables [11]. In the current work, the two variables are a gene and the main output. One metric of the ANOVA family is the F statistic. For a binary dataset with two classes, 1 and 2. We calculate the F statistic of a gene as follows. First, determine the sum of squares and the degrees of freedom, then employ the formula below (see [11]).

$$F = \frac{((\rho_1 - 1) + (\rho_2 - 1))[\rho_1 (\bar{x} - \bar{x}^*)^2 + \rho_2 (\bar{x} - \bar{x}^{**})^2]}{(\sum_{k=1}^{\rho_1} (x^{*,k} - \bar{x}^*)^2 + \sum_{k=1}^{\rho_2} (x^{**,k} - \bar{x}^{**})^2)} \quad (2)$$

with ρ_1 as the number of class 1 cases, and ρ_2 as the number of class 2 cases. The overall average of the gene is denoted by \bar{x} , \bar{x}^* denotes the average of the class 1 gene values, \bar{x}^{**} is the average of the class 2 gene values, $x^{*,k}$ the k value, with class 1 gene, and $x^{**,k}$ the average of the values, with class 2 gene. The gene is significant for determining the class, AD or N if the F statistic has a larger value and vice versa.

- **LASSO:**

A potent tool with two primary functions: regularization and feature selection. Also known as a form of regression method that uses regularization to best fit a generalized linear model. LASSO shatters the regression coefficient to zero for the variable with the smallest impact on the model, based on the principle of penalizing the regression model (L_1 -norm). Consequently,

the LASSO regression model has the best feature selection ability [25].

$$\omega^{Lasso} = \arg_{\omega} \min \left\{ \frac{1}{2} \sum_{i=1}^N \left(y_i - \omega_0 - \sum_{j=1}^p x_{ij} w_j \right)^2 + \lambda \sum_{j=1}^p |w_j| \right\} \quad (3)$$

where λ indicates the penalty parameter, which is frequently obtained using cross-validation, that defines the shrinkage proportion. For instances where the data has a small sample size and a high dimension, like in our case, LASSO maintains a great performance. Additionally, when a small number of variables account for the bulk of the data (have a big coefficient); the remaining variable's power for prediction is relatively low. Additionally, LASSO has some important benefits, including when features have several collinearities, LASSO effectively addresses this and produces highly independent features. Also, LASSO proved to be computationally less expensive.

2.4 Classification

The classification phase in our approach follows the GS phase, where we identify the most relevant genes to AD prediction. Generally speaking, any ML model can be used to predict AD, but in the experiments below, we'll focus on the three ML classifiers that proved to be the most effective in this task: MLP, SVM, and k-nearest neighbor (KNN). In the present work, the classification is a binary task, we classify AD cases from N cases. The three classifiers used in our experiment will be described as follows.

- **MLP:**

MLP is a feed-forward neural network containing one or more hidden layers that are directed in the following order: input, hidden, and output. The rectified linear unit (ReLU) or sigmoid functions are the activation functions used by MLP. The sigmoid function returns a value in the range [0,1], enabling the neural network to classify the data smartly. However, there is a downside to this sigmoid function feature, i.e., with deeper networks; the function's output is strongly biased towards the extremes of the range. To overcome this issue, the ReLU function was introduced. It gives 0 for an input value smaller than 0 but remains the original input value if it is larger than 0. The ReLU function is defined by

$$g(x) = x^+ \equiv \max(0, x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \quad (4)$$

where x is the input to a neuron. It is well known that some optimizers that improve and stabilize the learning rates of MLP include momentum, nesterovated gradient, stochastic gradient descent, and adaptive moment estimation (Adam). Adam is applied in our study, because of its low memory needs, great computational efficiency, and scalability with big datasets [26]. Since a learning rate of 0.01 is known to help preventing underfitting, it is the default value for this parameter, which regulates the step size in weight updates.

- **SVM:**

SVM is a well-known supervised ML technique that classifies data as follows. First, by nonlinearly transforming the data to higher dimension gene spaces. Second, present a linear optimal hyperplane or a decision boundary to detach the points of different classes (see [27]). SVM seeks to maximize distances between the hyperplane and closest training data points. The SVM classifier (the hyperplane) is denoted by

$$w^T \cdot \gamma(t) + a = 0 \quad (5)$$

w is a weight vector, $\gamma(t)$ is a nonlinear mapping and a is the bias. The optimal hyperplane is defined by finding w and a which minimize the function

$$\frac{1}{2} w^T \cdot w + B \sum_{i=1}^n v_i \quad (6)$$

with some slack variables $v_i > 0$, n the number of cases, and B some factor.

- **KNN:**

KNN algorithm is a famous nonparametric technique used in regression or classification (see [28]). Based on the inter-sample similarity seen in the training set, KNN intuitively categorizes unlabeled samples. When the number of neighbors is given as a parameter, a small value causes the model's decision boundary to be complicated and subsequently overfit, whereas a large value causes the decision boundary to be simple and underfit. Consequently, choosing a suitable value for this parameter is crucial. The core parameter of the KNN in this investigation was set to the value demonstrating the maximum performance, `n_neighbors`, individually for each dataset.

2.5 Performance Metrics

The last phase is to evaluate the performance of the classification techniques. To widen the assessment scope, we used for performance evaluation seven metrics: Acc, F1-score, Kappa index, precision, sensitivity (recall), specificity, and AUC.

To calculate these metrics, we first compose for each classification experiment a confusion matrix, having the number of the cases classified and whether the classification is true or false. Let TP be the number of true positive cases, TN the true negatives, FP the false positives, and FN the false negatives. Then, accuracy is defined as

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

Note that the denominator represents all Predictions.

The F1-Score is given by

$$\text{F1-score} = \frac{TP}{TP + 0.5(FP + FN)} \quad (8)$$

Kappa index is given by

$$\text{Kappa} = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)} \quad (9)$$

Precision is defined as

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

Sensitivity (Recall) is denoted by

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (11)$$

Specificity is defined as

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (12)$$

3 Experimental Work and Results

This section represents the results of our experimentation to check the validity of the proposed approach. The experiments were carried out using five different datasets consisting of multi-tissue GE profiles from various human brain regions as per the Algorithm below.

All the datasets have the property that the number of genes is significantly more than the number of cases as shown in Table 2. The code was run in Python version 3.7.3 with the Scikit-learn packages, it was executed on an Intel (R) Core (TM) i7-8550U CPU, 8 GB RAM, and 64-bit OS Win 10 configuration. Using this method, the most informative and pertinent genes for AD were chosen, and the other genes that did not contribute to the accuracy of a predictive model or would produce unfavorable outcomes were eliminated. In the beginning, the datasets are preprocessed to facilitate handling GE data. In predicting AD, every gene was evaluated for its relevance using the ANOVA filter metric as mentioned above. Genes are ranked in descending order and we select the highest 1000 genes, these numbers of genes are chosen according to the previous experiment we have done in [29] and achieved great results. Then an embedded method, LASSO, is used to learn which genes best contribute to the model's accuracy.

Algorithm 1: Best classifier identification and GS using hybrid filter method and embedded method.

Input: U // AD gene expression dataset
 $G = \{g_1, g_2, \dots, g_{|G|}\}$ //Set of genes
Output: The best classifier with the lowest number of genes
//Preprocessing:
1 Normalize the dataset U as per Eq. (1).
//Gene relevance evaluation:
2 **for** $j = 1$ **to** $|G|$ **do**
3 Calculate for gene g_j its F statistic metric F_{g_j} as per Eq. (2).
4 **end**
//Gene Selection–filter method: Rank the genes based on the ANOVA metric.
5 Sort the genes in descending order based on their F values, placing the sorted genes in the array G_F .
6 Select the highest 1000 genes and input them to LASSO.
//Gene Selection– embedded method: Select the minimum number of relevant genes.
//Tuning λ in the LASSO penalty.
7 **for** $\lambda \in \{.1, .01, .001, .0001\}$ **do**
8 For given λ simply minimize the cost function to find the weights ω as per Eq. (3).
9 Then compute the norm of ω (excluding ω_0).
10 **end**
11 Validate each classifier with the gene subset and calculate the metrics Acc, F1-score, Kappa index, precision, sensitivity (recall), specificity, and the AUC.
12 Reporting the validation

Different numbers of genes are selected from the datasets. These genes are compared with the list of the genes described in the well-known database AlzGene [30], which comprises 695 significant AD genes derived from 1395 research, and other eminent studies. The final number of the selected genes and the most significant genes, which are considered biomarkers of AD, are represented in Table 3. According to the used datasets, we have a kind of imbalanced data, where the number of AD cases is not equal to the normal cases. To overcome such an issue, we split the datasets using a repeated stratified k-fold cross-validation approach, with $k = 10$ and the number of repetitions is 30 (total of 300 times). This approach is suited for small-sized datasets, which are processed multiple times and report the mean performance across all folds and all repeats. It has the benefit of improving the estimate of the mean model performance.

Table 3: Summary of the number of the selected genes from the five datasets, and some of the significant genes reported in the AlzGene database and other eminent studies

Dataset	Selected genes ANOVA+LASSO	Significant genes	Reported in references
GSE5281	98 genes	IRF6	[30]
		HFE	[30]
		ND2	[31]
		COX2	[30]
GSE118553	152 genes	CAT	[30]
		ANK3	[30]
GSE132903	107 genes	RPF1	[32]
GSE48350	92 genes	THRA	[30]
		ND2	[31]
		COX2	[30]
GSE36980	57 genes	RPF1	[32]
		TXNIP	[33]

Every classifier is employed on the selected genes, and the seven performance metrics, (Acc, F1-score, Kappa index, precision, sensitivity (recall), specificity, and the AUC), are evaluated. Finally, the average of 300 results for each performance metric is recorded. We observed that the MLP model achieved the best results in four datasets. In the datasets GSE5281, GSE118553, and GSE132903, MLP contained five hidden layers, each hidden layer consisted of 4 neurons and used the ReLU as the activation function, and Adam as the gradient descent algorithm. The initial learning rate was 0.01 and it is executed over 500 epochs. In GSE48350 datasets MLP contained ten hidden layers and the remaining hyperparameters are the same in the other datasets. SVM achieved the best results in only one dataset, GSE36980. Figs. 2–5 represent the results of our experiment for four metrics, Acc, F1-score, Kappa index, and AUC, whose equations are explained above.

Table 4 represents the average results of the remaining metrics, precision, sensitivity (recall), and specificity. Each figure represents the average values of this metric for three classifiers, MLP, SVM, and KNN, and each selected subset of genes for the five datasets.

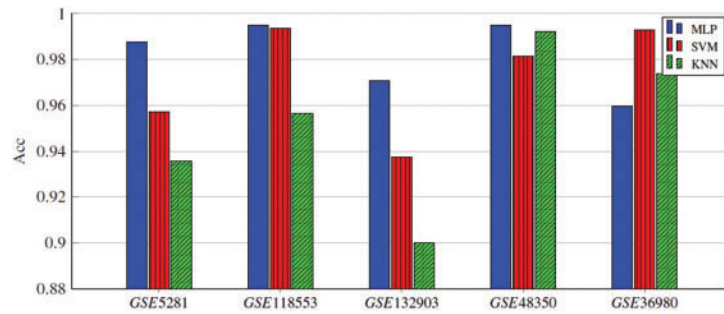


Figure 2: Acc of all three ML models applied for five gene subsets

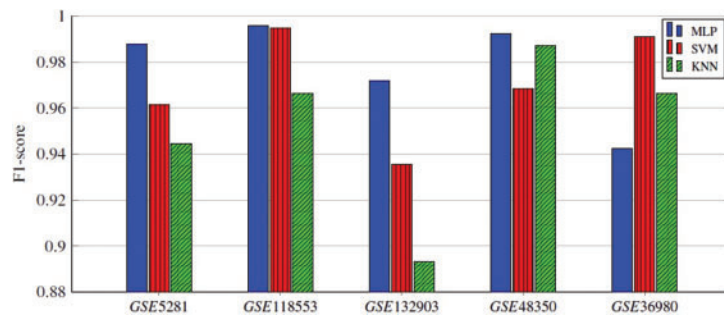


Figure 3: F1-score of all three ML models applied for five gene subsets

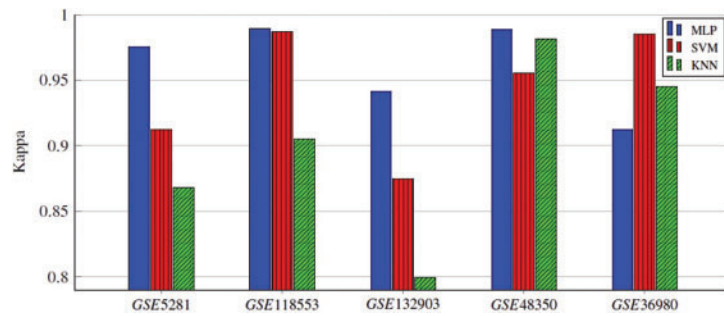


Figure 4: Kappa index of all three ML models applied for five gene subsets

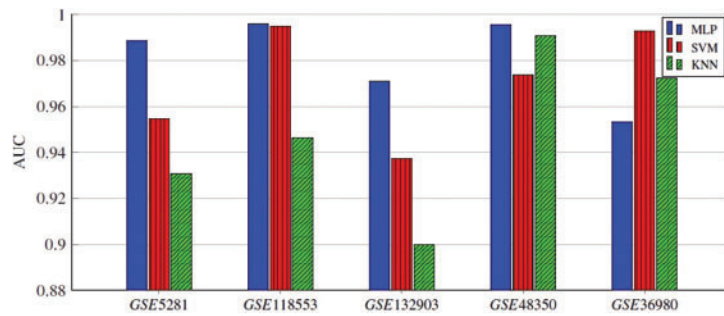


Figure 5: AUC of all three ML models applied for five gene subsets

Table 4: Summary of some performance metrics (Precision, Recall, and Specificity)

Dataset	Classifier	Precision	Recall	Specificity
GSE5281	MLP	0.9990	0.9784	0.9784
	SVM	0.9448	0.9822	0.9822
	KNN	0.9076	0.9888	0.9888
GSE118553	MLP	1	0.9918	0.9918
	SVM	1	0.9898	1
	KNN	0.9483	0.9868	1
GSE132903	MLP	0.9596	0.9877	1
	SVM	0.9504	0.9275	1
	KNN	0.9347	0.8649	1
GSE48350	MLP	0.9884	0.9971	0.9971
	SVM	0.9883	0.9535	0.9535
	KNN	0.9886	0.9873	0.9873
GSE36980	MLP	0.9923	0.9133	0.9133
	SVM	0.9911	0.9933	0.9933
	KNN	0.9783	0.9638	0.9638

4 Conclusion

We presented an approach for predicting a disease that receives insufficient attention in the literature—Alzheimer’s disease, using gene expression data. Compared with recently published competitive approaches, the proposed strategy has been demonstrated to accurately and efficiently predict AD using GE data. We used a hybrid method for GS to decide which genes are most important for AD prediction. In our experiments, five different GE datasets are used. Firstly, we employ a filtering method to rank all the genes according to their importance and select the highest ones. Secondly, we use an embedded method to reduce the number of genes by selecting the most relevant and informative genes. Third, three different classifiers are utilized and the best of them is what achieves the highest performance. It turns out that the MLP model has reached the highest performance and outperforms SVM and KNN in classifying four datasets using the smallest number of genes. While, SVM results outperform MLP and KNN in classifying only one dataset.

The approach is accurate and reliable for classifying AD cases from N cases using the smallest number of genes. Its flexibility characterizes it. It can use any ML classification model, filtering, and embedded method. It can also be applied to other GE-based diseases. It has been proved that it outperforms the state of the art in terms of achieving superior performance with fewer genes. Although the reliability of this work, has one constraint, it is used only for binary classification. Potential future work could solve this constraint by developing a framework capable of predicting AD phases.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

- [1] M. S. Tan, P. -L. Cheah, A. -V. Chin, L. -M. Looi and S. -W. Chang, "A review on omics-based biomarkers discovery for Alzheimer's disease from the bioinformatics perspectives: Statistical approach vs machine learning approach," *Computers in Biology and Medicine*, vol. 139, no. 2020, pp. 104947, 2021.
- [2] F. Chen, J. Bai, S. Zhong, R. Zhang, X. Zhang *et al.*, "Molecular signatures of mitochondrial complexes involved in Alzheimer's disease via oxidative phosphorylation and retrograde endocannabinoid signaling pathways," *Oxidative Medicine and Cellular Longevity*, vol. 2022, no. 4, pp. 1–12, 2022.
- [3] S. Wang, P. Phillips, Y. Sui, B. Liu, M. Yang *et al.*, "Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling," *Journal of Medical Systems*, vol. 42, no. 5, pp. 1–11, 2018.
- [4] L. Liu, S. Zhao, H. Chen and A. Wang, "A new machine learning method for identifying Alzheimer's disease," *Simulation Modelling Practice and Theory*, vol. 99, no. 2, pp. 102023, 2019.
- [5] W. L. Member, Y. Zhao, X. Chen, Y. Xiao and Y. Qin, "Detecting Alzheimer's disease on small dataset: A knowledge transfer perspective," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 1234–1242, 2018.
- [6] S. Alam and G. -R. Kwon, "Alzheimer disease classification using KPCA, LDA, and multi-kernel learning SVM," *International Journal of Imaging Systems and Technology*, vol. 27, no. 2, pp. 133–143, 2017.
- [7] B. Richhariya, M. Tanveer and A. H. Rashid, "Diagnosis of Alzheimer's disease using universum support vector machine based recursive feature elimination (USVM-RFE)," *Biomedical Signal Processing and Control*, vol. 59, pp. 101903, 2020.
- [8] X. Bi, S. Li, B. Xiao, Y. Li, G. Wang *et al.*, "Computer aided Alzheimer's disease diagnosis by an unsupervised deep learning technology," *Neurocomputing*, vol. 392, no. 1, pp. 296–304, 2020.
- [9] F. Ramzan, M. U. G. Khan, A. Rehmat, S. Iqbal, T. Saba *et al.*, "A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and Residual Neural Networks," *Journal of Medical Systems*, vol. 44, no. 2, pp. 37, 2020.
- [10] E. N. M. Id, A. M. Eldeib, I. A. Y. Id and Y. M. Kadah, "Alzheimer's disease diagnosis from diffusion tensor images using convolutional neural networks," *PLoS One*, vol. 15, no. 3, pp. 1–16, 2020.
- [11] J. Ramírez, J. Górriz, A. Ortiz, F. Martínez-Murcia, F. Segovia *et al.*, "Ensemble of random forests one vs. rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares," *Journal of Neuroscience Methods*, vol. 302, no. 6, pp. 47–57, 2018.
- [12] S. Bringas, S. Salomón, R. Duque, C. Lage and J. Luis, "Alzheimer's disease stage identification using deep learning models," *Journal of Biomedical Informatics*, vol. 109, no. 9, pp. 103514, 2020.
- [13] I. De Falco, G. De Pietro and G. Sannino, "A two-step approach for classification in Alzheimer's disease," *Sensors*, vol. 22, no. 11, pp. 3966, 2022.
- [14] F. García-Gutierrez, J. Díaz-Álvarez, J. A. Matias-Guiu, V. Pytel, J. Matías-Guiu *et al.*, "GA-MADRID: Design and validation of a machine learning tool for the diagnosis of Alzheimer's disease and frontotemporal dementia using genetic algorithms," *Medical & Biological Engineering & Computing*, vol. 60, no. 9, pp. 2737–2756, 2022.
- [15] S. M. Ayyad, A. I. Saleh and L. M. Labib, "Gene expression cancer classification using modified K-Nearest Neighbors technique," *BioSystems*, vol. 176, no. 23, pp. 41–51, 2019.
- [16] C. D. A. Vanitha, D. Devaraj and M. Venkatesulu, "Gene expression data classification using support vector machine and mutual information-based gene selection," *Procedia Computer Science*, vol. 47, no. 457, pp. 13–21, 2015.
- [17] S. M. Ayyad, A. I. Saleh and L. M. Labib, "A new distributed feature selection technique for classifying gene expression data," *International Journal of Biomathematics*, vol. 12, no. 4, pp. 1950039, 2019.
- [18] H. Patel, R. Iniesta, D. Stahl, R. J. Dobson and S. J. Newhouse, "Working towards a blood-derived gene expression biomarker specific for Alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 74, no. 2, pp. 545–561, 2020.
- [19] T. Lee and H. Lee, "Prediction of Alzheimer's disease using blood gene expression data," *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020.

- [20] X. Li, H. Wang, J. Long, G. Pan, T. He *et al.*, “Systematic analysis and biomarker study for Alzheimer’s disease,” *Scientific Reports*, vol. 8, no. 1, pp. 1–14, 2018.
- [21] L. Wang and Z. Liu, “Detecting diagnostic biomarkers of Alzheimer’s disease by integrating gene expression data in six brain regions,” *Frontiers in Genetics*, vol. 10, pp. 157, 2019.
- [22] C. Park, J. Ha and S. Park, “Prediction of Alzheimer’s disease based on deep neural network by integrating gene expression and DNA methylation dataset,” *Expert Systems with Applications*, vol. 140, no. 4, pp. 112873, 2020.
- [23] N. Voyle, A. Keohane, S. Newhouse, K. Lunnon, C. Johnston *et al.*, “A pathway based classification method for analyzing gene expression for Alzheimer’s disease diagnosis,” *Journal of Alzheimer’s Disease*, vol. 49, no. 3, pp. 659–669, 2016.
- [24] “Gene Expression Omnibus.” [Online]. Available: <https://www.ncbi.nlm.nih.gov/geo/> (accessed on 24 September 2022).
- [25] S. Zhang, F. Zhu, A. Q. Yu and X. Zhu, “Identifying DNA-binding proteins based on multi-features and LASSO feature selection,” *Biopolymers*, vol. 112, no. 2, pp. e23419, 2021.
- [26] J. Kim, Y. Yoon, H. Park and Y. Kim, “Comparative study of classification algorithms for various DNA microarray data,” *Genes*, vol. 13, no. 3, pp. 494, 2022.
- [27] C. C. Aggarwal, *Machine Learning for Text*, 2nd ed., Switzerland: Springer Nature Switzerland AG 2022, 2022.
- [28] S. Wan, Y. Liang, Y. Zhang, M. Guizani, “Deep multi-layer perceptron classifier for behavior analysis to estimate Parkinson’s disease severity using smartphones,” *IEEE Access*, vol. 6, pp. 36825–36833, 2018.
- [29] A. El-gawady, M. A. Makhlof, B. S. Tawfik and H. Nassar, “Machine learning framework for the prediction of Alzheimer’s disease using gene expression data based on efficient gene selection,” *Symmetry*, vol. 14, no. 3, pp. 491, 2022.
- [30] “Alzgene.” [Online]. Available: <https://http/www.alzgene.org/> (accessed on 25 September 2022).
- [31] Y. E. Cruz-Rivera, J. Perez-Morales, Y. M. Santiago, V. M. Gonzalez, L. Morales *et al.*, “A selection of important genes and their correlated behavior in Alzheimer’s disease,” *Journal of Alzheimer’s Disease*, vol. 65, no. 1, pp. 193–205, 2018.
- [32] H. Li, G. Hong, M. Lin, Y. Shi, L. Wang *et al.*, “Identification of molecular alterations in leukocytes from gene expression profiles of peripheral whole blood of Alzheimer’s disease,” *Scientific Reports*, vol. 7, no. 1, pp. 1–10, 2017.
- [33] J. Ren, B. Zhang, D. Wei and Z. Zhang, “Identification of methylated gene biomarkers in patients with Alzheimer’s disease based on machine learning,” *BioMed Research International*, vol. 2020, no. 4, pp. 1–11, 2020.