OXFORD

Briefings in Bioinformatics, 2024, **25(4)**, bbae291
https://doi.org/10.1093/bib/bbae291
Problem Solving Protocol

AITeQ: a machine learning framework for Alzheimer's prediction using a distinctive five-gene signature

Ishtiaque Ahammad[1,‡], Anika Bushra Lamisa[1,‡], Arittra Bhattacharjee[1,‡], Tabassum Binte Jamal[1], Md. Shamsul Arefin[2], Zeshan Mahmud Chowdhury[1], Mohammad Uzzal Hossain[1], Keshob Chandra Das[3], Chaman Ara Keya[2], Md. Salimullah[3,*]

[1]Bioinformatics Division, National Institute of Biotechnology, Ganakbari, Ashulia, Savar, Dhaka 1349, Bangladesh
[2]Department of Biochemistry and Microbiology, North South University, Bashundhara, Dhaka 1229, Bangladesh
[3]Molecular Biotechnology Division, National Institute of Biotechnology, Ganakbari, Ashulia, Savar, Dhaka 1349, Bangladesh
*Corresponding author. Molecular Biotechnology Division, National Institute of Biotechnology, Ganakbari, Ashulia, Savar, Dhaka 1349, Bangladesh. E-mail: dgnib.gov.bd@gmail.com
‡Ishtiaque Ahammad, Anika Bushra Lamisa and Arittra Bhattacharjee contributed equally to this work.

Abstract

Neurodegenerative diseases, such as Alzheimer's disease, pose a significant global health challenge with their complex etiology and elusive biomarkers. In this study, we developed the Alzheimer's Identification Tool (AITeQ) using ribonucleic acid-sequencing (RNA-seq), a machine learning (ML) model based on an optimized ensemble algorithm for the identification of Alzheimer's from RNA-seq data. Analysis of RNA-seq data from several studies identified 87 differentially expressed genes. This was followed by a ML protocol involving feature selection, model training, performance evaluation, and hyperparameter tuning. The feature selection process undertaken in this study, employing a combination of four different methodologies, culminated in the identification of a compact yet impactful set of five genes. Twelve diverse ML models were trained and tested using these five genes (*CNKSR1*, *EPHA2*, *CLSPN*, *OLFML3*, and *TARBP1*). Performance metrics, including precision, recall, F1 score, accuracy, Matthew's correlation coefficient, and receiver operating characteristic area under the curve were assessed for the finally selected model. Overall, the ensemble model consisting of logistic regression, naive Bayes classifier, and support vector machine with optimized hyperparameters was identified as the best and was used to develop AITeQ. AITeQ is available at: https://github.com/ishtiaque-ahammad/AITeQ.

Keywords: AITeQ; Alzheimer's disease; machine learning; transcriptomics; differentially expressed genes

Introduction

Millions across the world are affected by Alzheimer's disease (AD) that leads to cognitive impairments. For timely and effective treatment, early and accurate diagnosis of the disease is very important. Traditional diagnostic approaches often rely on clinical symptoms and neuroimaging, which might not capture the molecular intricacies of the disease. Ribonucleic acid-sequencing (RNA-seq), a high-throughput sequencing technique, offers a comprehensive snapshot of the transcriptome and enables the identification of gene expression alterations associated with neurodegeneration [1].

Machine learning (ML) algorithms have demonstrated remarkable potential in analyzing large-scale, complex datasets like RNA-seq data. By integrating ML techniques, researchers can identify disease-specific gene expression signatures, classify patient samples, and predict disease progression [2]. ML models learn from patterns within the data and can uncover subtle relationships that might elude traditional statistical methods. Selection of important genes from RNA-seq data is an application of supervised ML techniques [3].

Identifying reliable biomarkers is a critical step in disease diagnosis and prognosis. ML models can aid in the discovery of potential biomarkers by pinpointing genes consistently associated with disease states. Since numerous RNA-seq studies are based on the comparison between cases and controls, one such study focused on the development of a logistic regression model where disease state was described as a function of RNA-seq reads [4]. The support vector machine (SVM) was also used for the early detection for both prediction and classification of AD [5]. Another

Ishtiaque Ahammad is a scientific officer at National Institute of Biotechnology.
Anika Bushra Lamisa is a post-graduate research fellow at National Institute of Biotechnology.
Arittra Bhattacharjee is a scientific officer at National Institute of Biotechnology.
Tabassum Binte Jamal is a scientific officer at National Institute of Biotechnology.
Md. Shamsul Arefin is a graduate student at North South University.
Zeshan Mahmud Chowdhury is a scientific officer at National Institute of Biotechnology.
Mohammad Uzzal Hossain is a scientific officer at National Institute of Biotechnology.
Keshob Chandra Das is a principal scientific officer at National Institute of Biotechnology.
Chaman Ara Keya is a professor at the Department of Biochemistry and Microbiology, North South University.
Md. Salimullah is the director general (additional charge) and chief scientific officer at National Institute of Biotechnology.
Received: September 15, 2023. **Revised:** May 23, 2024. **Accepted:** June 4, 2024

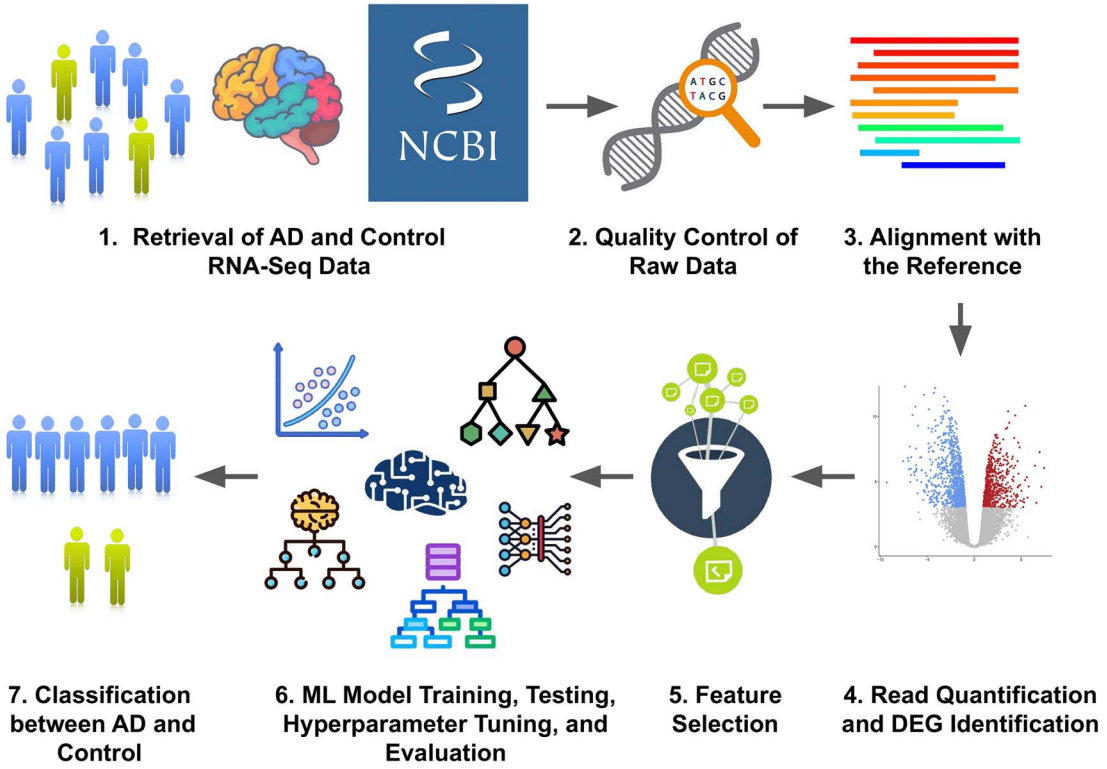

Figure 1. Workflow of the study. RNA-seq data of AD and control were retrieved from NCBI. The raw reads were subjected to quality control using FastQC and subsequently aligned with the human reference genome (GRCh38.p13) using HISAT2. The quantification of reads was performed using the featureCounts algorithm, while the identification of DEGs was conducted using the DESeq2 statistical tool. Feature selection was carried out using four methods. It was followed by 13 ML model training, testing, hyperparameter tuning, and evaluation.

study revealed the efficacy of the decision tree algorithm for construction of classifiers that can classify different AD genes [6]. Random forest model was also implemented to predict the individualized conversion from mild cognitive impairment stage to AD [7]. A more robust multi stage classifier-based approach consisting of K-nearest neighbor (KNN), SVM, and naive Bayes classifier was reported to be able to efficiently classify AD [8]. For biomarker-based early diagnosis of AD with high classification accuracy, gradient boosting algorithm was also used [9]. Analyzing single-cell RNA-seq data from patients with AD and healthy individuals using extreme gradient boosting (XGBoost) revealed genes with diagnostic potential [10]. A meta-analysis and ML-based integrative study identified differentially expressed microRNAs in blood as potential biomarkers for AD using adaptive boosting (Adaboost) [11]. Light gradient boosting machine (LightGBM) was used for feature selection to detect AD from circulating non-coding RNAs [12].

Predictive modeling for AD detection is common using radiomics data. Radiomics has demonstrated promising outcomes in the diagnosis of AD. Nevertheless, relying solely on imaging is insufficient for the detection of AD, and frequent radiological examinations may result in further health complications [13, 14]. Hence, multiple methods of detection would be more robust than using a single method.

In light of these advancements, we aimed to analyze publicly available AD-associated gene expression data and build a gene signature-based ML framework that can differentiate AD from control. For this purpose, several sophisticated feature selection methods and ML algorithms were utilized following the identification of differentially expressed genes (DEGs). Findings from this study are likely to contribute to the better understanding of the genes most crucial for AD and utilize them as biomarkers.

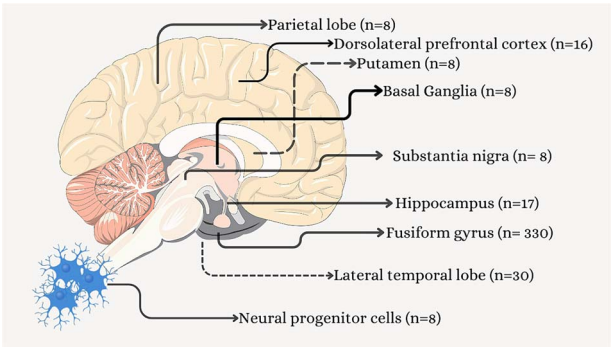

Figure 2. Regions of the brain from where the RNA-seq datasets were generated (with sample size *n*).

Materials and methods

A visual representation of the workflow followed in the study is illustrated in Figure 1.

Data retrieval and preprocessing

The NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) database was used to obtain the RNA-seq datasets from nine projects [15]. The NCBI BioProject IDs and the sample source for these projects are PRJNA675355 (source: Putamen), PRJNA683625 (source: neural progenitor cells), PRJNA767074 (source: hippocampus), PRJNA796229 (source: substantia nigra, parietal lobe, hippocampus, basal ganglia), PRJNA279526 (source: hippocampus), PRJNA232669 (source: dorsolateral prefrontal cortex), PRJNA377568 (source: fusiform gyrus), PRJNA413568 (source: lateral temporal lobe), and PRJNA516886 (source: fusiform gyrus; Fig. 2). A table containing more detailed information (sample

size, counts of AD and healthy subjects, gender distribution, age range, and brain region) of each project has been included in Supplementary Table S1.

Combining all datasets, the total number of samples were 433 individuals, of whom 293 were diagnosed with AD and 140 were healthy controls. The RNA-seq data analysis workflow consisted of several steps. At first, the raw read quality was checked using FastQC [16]. Next, the alignment of the reads to the *Homo sapiens* GRCh38.p13 reference genome was carried out using HISAT2 [17]. The mapped reads were then distributed to genomic features. Finally, gene expression was quantified using FeatureCounts [18]. The count table was separated with a ratio of 80:20 with random shuffle and stratification where 80% data were kept for training and the further analysis, whereas 20% data were used as unseen test dataset. On the training dataset, the DESeq2 statistical tool was utilized to identify DEGs [19]. In order to adjust the *P*-values and ascertain the reliability of the identified DEGs, the false discovery rate method was employed [20]. Between the control and AD groups, the fold change (FC) of each gene was calculated. Genes with a *P*-adjusted value of $<.01$ and a Log2FC value $> |0.5|$ were considered as significant DEGs [19]. The normalized and variant stabilized count of these significant DEGs were used as the features for ML. Moreover, limma::removeBatchEffect() function was separately applied on train and test datasets to remove the batch effects [21]. The normalized, variance stabilized, and batch effect removed datasets were used for feature selection.

Oversampling technique for the minority class

To overcome the data imbalance, synthetic minority oversampling technique (SMOTE) was applied on the training dataset. SMOTE created synthetic samples combining real points in the feature space to provide new minority class data [22].

Feature selection for ML models

This study utilized a comprehensive array of feature selection strategies to unravel the most important features needed for training various ML models. The determination of feature importance was conducted through the application of four separate methodologies, namely random forest classifier [23], gradient boosting classifier [24], recursive feature elimination [25], and LassoCV [26]. Feature selection was performed only on the training set to avoid information leakage. In our study, we utilized the scikit-learn "SelectFromModel" function of the RandomForestClassifier and GradientBoostingClassifier algorithms to assess the relative importance of each feature in the model. The recursive feature elimination technique entails iteratively eliminating features with the least significance by employing a linear regression model. Furthermore, the LassoCV technique employed a Lasso linear regression model to award significance scores to features according to their coefficients. These strategies, when used together, enabled the identification of important features from our dataset. A Venn diagram was constructed with the top 10 features identified by each approach, and the set of features that were found to be common to all methods were selected. Subsequently, the selected features were utilized to build and refine ML models for AD classification.

ML model training

Scaling of features is an important part of data preprocessing in most ML methodologies. In this study, the input features were scaled utilizing the "StandardScaler" function from the preprocessing module in the scikit-learn toolkit [27]. The mean and standard deviation of the training dataset was applied on the test dataset for standard scaling. Afterward, the test dataset was utilized to evaluate the performance of the models that were trained on the training dataset. The training process involved the utilization of 13 ML models, namely logistic regression, SVM, decision tree, random forest, naive Bayes, KNN, gradient boosting, Adaboost, XGBoost, LightGBM and multilayer perceptron (MLP) classifier, Ensemble Model 1 (logistic regression + naive Bayes classifier + SVM + MLP classifier with soft voting), and Ensemble Model 2 (logistic regression + naive Bayes classifier + SVM with soft voting).

Logistic regression

In ML, logistic regression is an algorithm that is frequently used for solving regression tasks where the dependent variable is categorical in nature. It predicts the probability of the dependent variables by estimating the coefficients of the independent variables in the ML model [28].

SVM

SVM is a powerful ML model, which is used in both classification and regression domains. Recognition of the hyperplane that achieves the maximum separation between two classes is the primary goal of SVM. Identification of such hyperplanes relies upon the identification of the support vectors [29].

Decision tree

Decision tree is an ML model where each internal node of the tree is equivalent to a choice made based on a particular attribute, and each leaf node corresponds to an output of classification or regression. The algorithm iteratively divides the dataset into smaller subsets. It continues to look for the feature that contains the most significant information, until a predetermined output is found [30].

Random forest

Random forest is a notable ensemble learning strategy that is utilized for not just classification and regression but also feature selection. In case of ensemble learning, numerous decision trees are put to use for enhanced accuracy and generalization [23].

Naive Bayes

Naive Bayes is a Bayes' theorem-based probabilistic model that calculates the likelihood of a class from a given set of features. It assumes that the features are independent of each other while assigning them a class label, thereby getting the name "naive" [31].

KNN

KNN is a nonparametric method that is mainly used to decipher problems involving classification and regression. KNN functions through the identification of neighboring data points based on their similarity [32].

Gradient boosting

Gradient boosting exhibits remarkable efficacy in making predictions from intricate datasets, such as RNA-seq data. It is an ensemble method that iteratively combines numerous weak learners in order to generate strong learners which can eventually make accurate predictions [33].

Adaboost

The Adaboost algorithm takes an iterative approach to modify the weights assigned to the training data, with the objective of

focusing on the misclassified cases in each iteration. During each iteration, Adaboost uses a weak learner to train on a certain subset of the training data. It takes into account its classification error and assigns a weight to each training example. The weights assigned to misclassified examples are augmented, while the weights assigned to correctly classified examples are diminished. This technique is iteratively implemented until a predetermined outcome is satisfied [34].

XGBoost

XGBoost algorithm enhances the conventional gradient boosting approach by integrating well established regularization methods such as L1 and L2 regularization, to minimize the possibility of model overfitting. Additionally, XGBoost employs a novel approach to estimate the second-order gradient of the loss function. Thus, it enhances both the speed of convergence and the accuracy of the model to solve regression and classification tasks [35].

LightGBM

The LightGBM is a framework that utilizes a collection of weak learners, most commonly in the form of decision trees, with the objective of constructing a strong learner. It operates by iteratively including additional models into its ensemble learning approach, with a primary objective of lowering the gradient of the loss function [36].

MLP classifier

MLP classifiers are artificial neural networks with fully connected neurons and activation functions. MLP classifiers can differentiate data that are not linearly separable [37].

Ensemble modeling

Two ensemble models were constructed. The first one incorporated logistic regression, naive Bayes classifier, SVM, MLP classifier (Ensemble Model 1). The second one incorporated logistic regression, naive Bayes classifier and SVM (Ensemble Model 2). The ensemble models also employed a soft voting algorithm to merge predictions from the classifiers, leveraging the probability of each prediction.

Hyperparameter tuning

Hyperparameter tuning is a crucial step in optimizing the performance of ML models and was an integral component of the current study. The primary objective of hyperparameter tuning is to identify the optimal configuration of hyperparameters that maximizes the models' performance. In this study, we utilized the scikit-learn library in Python to conduct comprehensive hyperparameter tuning for all ML models. Our hyperparameter tuning process involved utilizing GridSearchCV to systematically traverse the hyperparameter space [38]. The best-performing hyperparameters were chosen based on the results of the search, ultimately enhancing the generalization ability of our models and ensuring their robustness to overfitting.

K-fold cross-validation with hyperparameter tuned ensemble model

Cross-validation is an important method in ML, as it provides a more reliable estimate of the success of the model on unseen data as opposed to a single train-test split. It has the ability to remove the variability that might arise as a result of using a single partition of the data for testing. After training 13 previously mentioned models, the "StratifiedKFold" function from scikit-learn was used to perform a 10-fold cross-validation by concatenating training and test dataset [39]. During each iteration, one-fold was used as the validation set. The remaining nine-folds were used for training. In each of 10 iterations, the performance of the model was evaluated by calculating accuracy, Matthew's correlation coefficient (MCC), Area under the receiver operating characteristic curve (AUC–ROC), and F1 score. The average of these 10 results was calculated to get an overall measure of how the models were likely to work on unseen data.

Establishment of Alzheimer's Identification Tool

The full Alzheimer's Identification Tool (AITeQ) documentation containing the instructions on how to run the tool for AD prediction can be found at https://github.com/ishtiaque-ahammad/AITeQ.

The entire experimental setup has been summed up in Figure 3.

Results

Eighty-seven DEGS were identified

The quality assessment of the raw-sequencing data was conducted for a total of 433 raw sequences, revealing that all of them were of high quality. After aligning the reads to the human reference genome, a total of 62 702 genes were discovered. These genes were then subjected to differential expression analysis in the quantification step. A comprehensive analysis revealed that a total of 87 genes had differential expression in samples obtained from patients with AD under P-adjusted value of $<.01$ and a Log2FC value $> |0.5|$ parameters.

Ensemble Model 2 exhibited best overall performance

Top important features were identified by each of the four feature selection tools (Table 1). Among these features, five genes were found to be commonly identified by all four tools (Fig. 4). These five genes (*CNKSR1*, *EPHA2*, *CLSPN*, *OLFML3*, and *TARBP1*) were finally selected as features to be used for training 13 ML models. After systematic exploration of a wide range of hyperparameters in order to find the optimal combination for each of the 13 ML models, the best hyperparameter values obtained have been summarized in Table 2. The performance of the models was evaluated based on accuracy (Fig. 5), MCC (Fig. 6), AUC–ROC (Fig. 7), F1 score for non-AD (Fig. 8), and F1 score for AD (Fig. 9) before and after hyperparameter tuning. Supplementary Table S2 contains the values of these performance metrics in tabular format. Kruskal–Wallis rank sum test was used for calculating the statistical significance of the differences in performance (accuracy, MCC, AUC–ROC, and F1 score) of the considered classifiers. However, the differences in performance were not statistically significant according to the Kruskal–Wallis rank sum test. The result has been included in Supplementary Table S3. After hyperparameter tuning, the Ensemble Model 2 exhibited the best overall performance (Accuracy = 0.74, MCC = 0.41, AUC–ROC = 0.73, F1 score_non_AD = 0.59, F1 score_AD = 0.81).

K-fold cross-validation with hyperparameter tuned Ensemble Model 2

Cross-validation is an important method in ML as it provides a more reliable estimate of the success of the model on unseen data as opposed to a single train-test split. It has the ability

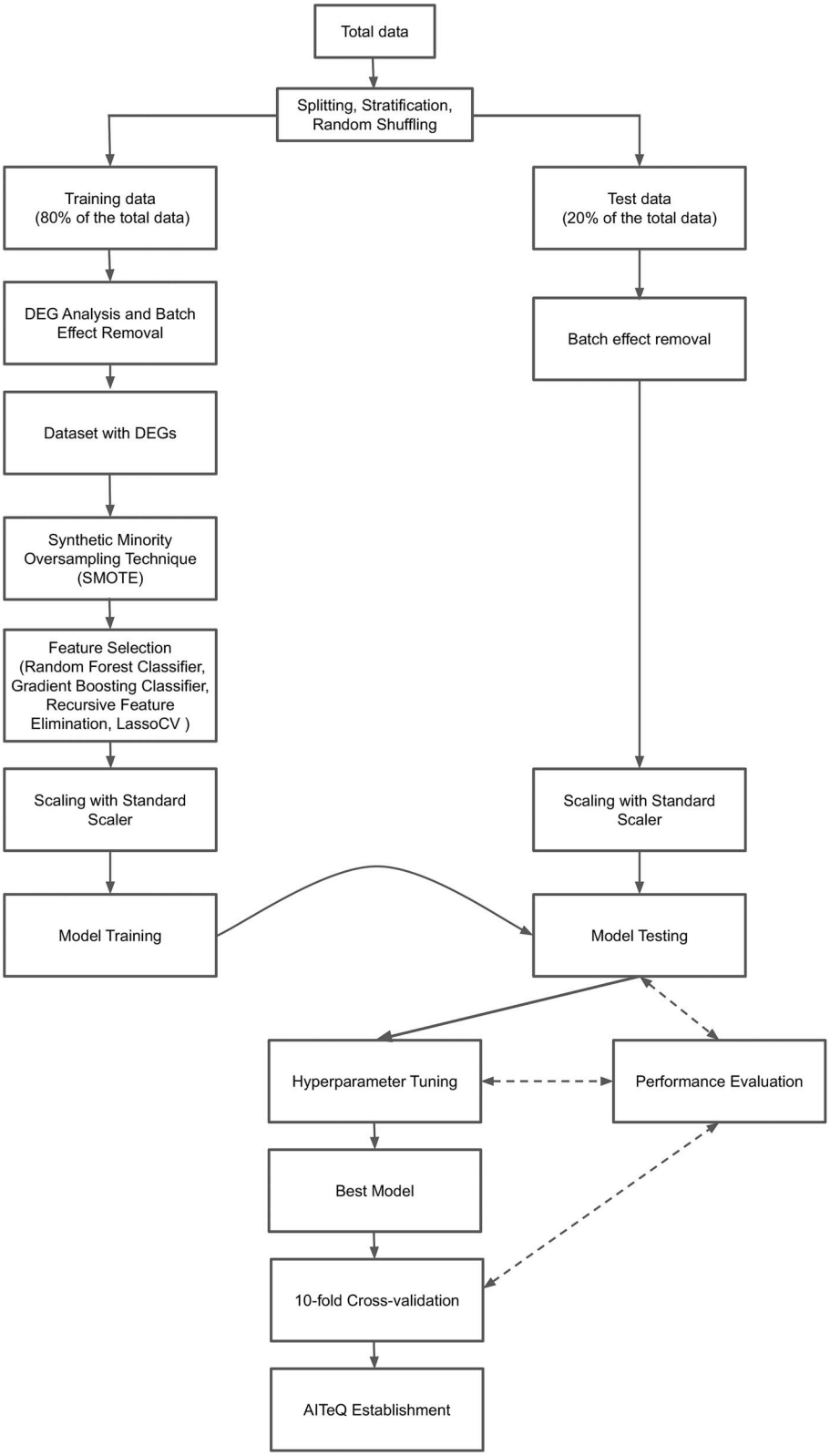

Figure 3. The experiment setup. After splitting the total data into training and test data, they followed separate courses. The training data were subjected to DEG analysis, batch effect removal, SMOTE, feature selection, and standard scaling before model training, while the test data underwent batch effect removal (independently from training data) and standard scaling before model testing. The trained models were then applied on the test data. AITeQ was established after the tested models went through hyperparameter tuning, selection of best model, and 10-fold cross-validation. Performance evaluation was carried out at three different stages (before and after hyperparameter tuning and during 10-fold cross-validation) in order to gain important feedback and continue on to the next stage of the workflow.

Table 1. Selected features or genes (Ensembl ID) from random Forest classifier, gradient boosting classifier, recursive feature elimination, and LassoCV

Random forest	Gradient boosting classifier	Recursive feature elimination	LassoCV
ENSG00000059588	ENSG00000059588	ENSG00000059588	ENSG00000059588
ENSG00000092607	ENSG00000092607	ENSG00000092853	ENSG00000092607
ENSG00000092853	ENSG00000092853	ENSG00000116774	ENSG00000092853
ENSG00000116679	ENSG00000116254	ENSG00000142615	ENSG00000116774
ENSG00000116774	ENSG00000116774	ENSG00000142627	ENSG00000122224
ENSG00000116824	ENSG00000117592	ENSG00000142675	ENSG00000127472
ENSG00000117091	ENSG00000122224	ENSG00000157064	ENSG00000142627
ENSG00000122224	ENSG00000123080	ENSG00000157978	ENSG00000142675
ENSG00000123080	ENSG00000142615	ENSG00000184371	ENSG00000162571
ENSG00000127472	ENSG00000142627	ENSG00000235777	ENSG00000183298
ENSG00000142615	ENSG00000142675		ENSG00000203859
ENSG00000142627	ENSG00000143631		ENSG00000231615
ENSG00000142675	ENSG00000157064		ENSG00000232878
ENSG00000143119	ENSG00000157978		
ENSG00000143631	ENSG00000162618		
ENSG00000157064	ENSG00000181656		
ENSG00000157978	ENSG00000215808		
ENSG00000158014	ENSG00000225675		
ENSG00000172260	ENSG00000227056		
ENSG00000181656	ENSG00000227466		
ENSG00000183298	ENSG00000227741		
ENSG00000183317	ENSG00000228187		
ENSG00000184371	ENSG00000231364		
ENSG00000187513	ENSG00000231615		
ENSG00000197106	ENSG00000232650		
ENSG00000203859	ENSG00000233623		
ENSG00000215808	ENSG00000235777		
ENSG00000215874	ENSG00000236290		
ENSG00000223489	ENSG00000284696		
ENSG00000225087	ENSG00000117592		
ENSG00000225675			
ENSG00000226759			
ENSG00000227056			
ENSG00000227741			
ENSG00000228057			
ENSG00000230523			
ENSG00000230817			
ENSG00000231364			
ENSG00000231615			
ENSG00000232650			
ENSG00000232878			
ENSG00000233623			
ENSG00000235777			
ENSG00000236290			
ENSG00000237505			
ENSG00000270911			
ENSG00000284696			

to remove the variability that might arise as a result of using a single partition of the data for testing. The "StratifiedKFold" function from scikit-learn was used to perform a 10-fold cross-validation by concatenating training and test dataset using the hyperparameter tuned Ensemble Model 2. During each iteration, one-fold was used as the validation set. The remaining nine-folds were used for training. After each of the 10 iterations, 10 individual accuracy, MCC, AUC–ROC, and F1 scores were obtained based on how well the models performed on the validation set. The average of these 10 accuracy, MCC, AUC–ROC and F1 scores, was calculated to get an overall measure of how the models were likely to work on overall data (Fig. 10). The raw values of each fold of cross-validation have been included in Supplementary Table S4.

AITeQ implementation

The structure of the final AITeQ ensemble model is described in Figure 11. AITeQ documentation can be found at https://github.com/ishtiaque-ahammad/AITeQ. The tool can be used directly through the Google colab platform [40].

Discussion

Integration of ML methods with transcriptomics data processing has been reported to benefit the understanding of complicated neurodegenerative illnesses like AD. Along these lines, the current study aimed at analyzing RNA-seq data using ML algorithms to predict AD. The findings from this study will contribute to the

Table 2. Best hyperparameter values for ML models following tuning

ML model	Hyperparamters	Selected value
Logistic regression	C	0.01
SVM	C	0.1
	Gamma	0.001
Decision tree	Max depth	10
	Min samples_leaf	1
	Min samples split	2
Random forest	N estimators	300
	Max depth	None
Naive Bayes	Var smoothing	1e-09
KNN	N neighbors	3
	p	1
	Weights	distance
Gradient boosting	Learning rate	0.01
	Max depth	6
	N estimators	300
Adaboost	Learning rate	0.5
	N estimators	150
XGBoost	Max depth	7
	N estimators	100
LightGBM	Learning rate	0.3
	Max depth	6
	N estimators	200
MLP	Activation	relu
	hidden_layer_sizes	150
	max_iter	1500
	Solver	lbfgs
Ensemble Model 1 (logistic regression + naive Bayes classifier + SVM + MLP classifier with soft voting)	logistic__model__C	0.01
	svm__model__C	0.1
	svm__model__gamma	0.001
	mlp_activation	relu
	mlp_hidden_layer_sizes	150
	mlp_max_iter	1500
	mlp_solver	lbfgs
	nbc_Var smoothing	1e-09
Ensemble Model 2 (logistic regression + naive Bayes classifier + SVM with soft voting)	lgr_model__C	0.001
	nbc_model__var_smoothing	1e-09
	svm_model_C	0.1
	svm_model__gamma	0.1

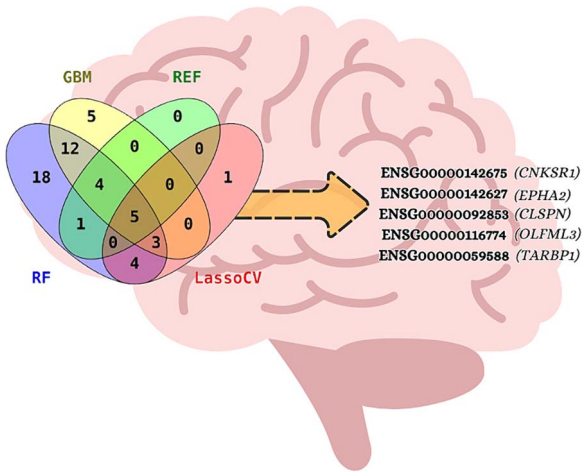

Figure 4. A Venn diagram of features (genes) selected by four distinct feature selection algorithms—random Forest classifier, gradient boosting classifier, recursive feature elimination, and LassoCV. Five genes were unanimously predicted by all four methods.

ongoing efforts for early and precise diagnosis of AD by utilizing a refined five-gene signature as an accurate predictor of the disease.

The work relied heavily on the thorough analysis of RNA-seq data from publicly available datasets in NCBI. Quality evaluation, read alignment, and quantification constituted parts of the preprocessing processes were essential for generating valid inputs for the ML models in the subsequent step. The complex transcriptomic aberrations associated with AD were highlighted by the finding that over 87 genes undergo differential expression in individuals with the condition.

One of the most crucial aspects of this study was the selection of features (genes) while developing a robust predictive model for AD. Using a combination of techniques, such as the random forest classifier, gradient boosting classifier, recursive feature elimination, and LassoCV, five genes were consistently determined to be important across all employed techniques. Implementing multiple methods improved the credibility of the gene signature, resulting in a more dependable method for predicting AD. This was in accordance with a number of previously conducted ML studies that utilized feature selection to solve classification

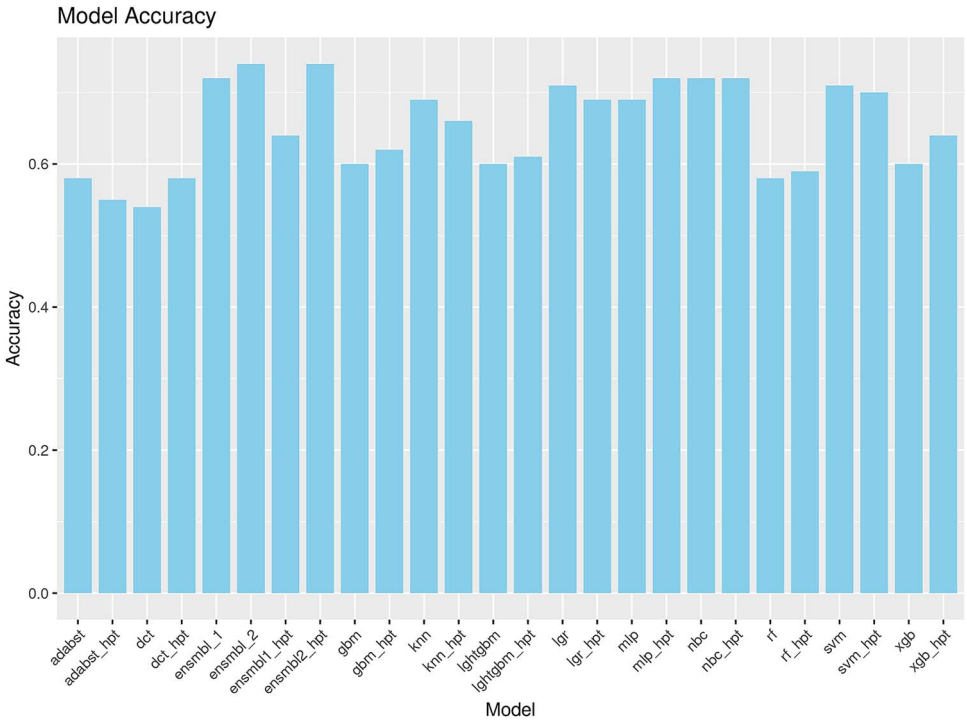

Figure 5. Accuracy of different models before hyperparamter tuning lgr (logistic regression), rf (random forest), nbc (naive Bayes classifier), xgboost (extreme gradient boosting), adaboost (adaptive boosting), dct (decision tree), lghtgbm (light gradient boosting machine), gbm (gradient boosting machine), knn (k-nearest neighbor), svm (support vector machine), mlp (multilayer perceptron), ensmbl1 (lgr + nbc + svm + mlp with soft voting), ensmbl2 (lgr + nbc + svm with soft voting). Accuracy of different models after hyperparamter tuning (hpt) lgr_hpt, rf_hpt, nbc_hpt, xgboost_hpt, adaboost_hpt, dct_hpt, lghtgbm_hpt, gbm_hpt, knn_hpt, svm_hpt, mlp_hpt, ensmbl1_hpt (lgr + nbc + svm + mlp with soft voting), ensmbl2_hpt (lgr + nbc + svm with soft voting).

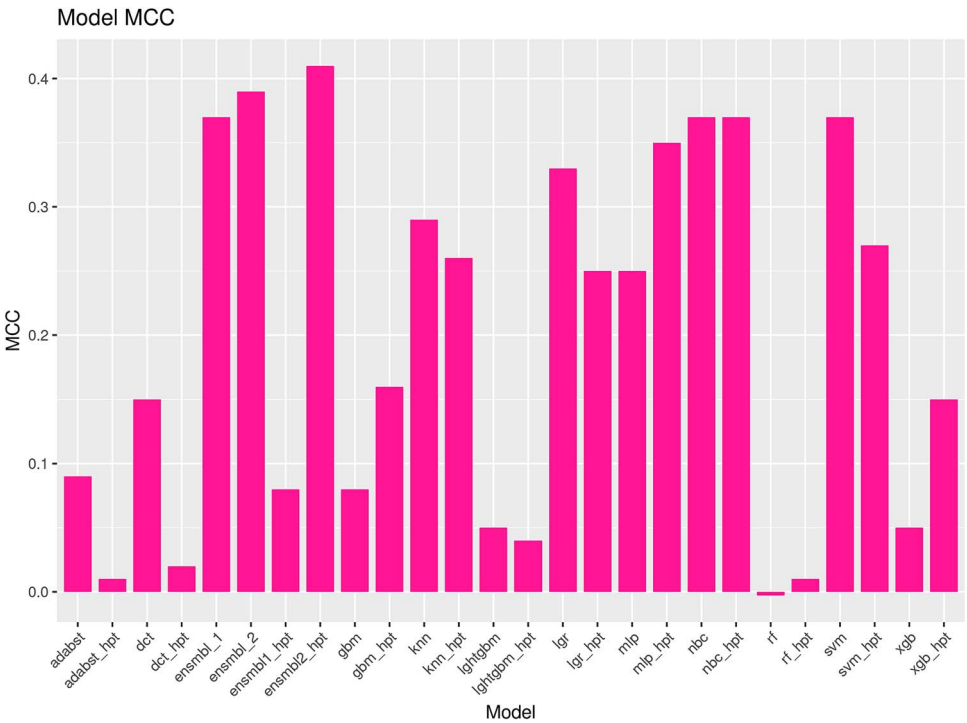

Figure 6. MCC evaluation of different models before hyperparamter tuning lgr (logistic regression), rf (random forest), nbc (naive Bayes classifier), xgboost (extreme gradient boosting), adaboost (adaptive boosting), dct (decision tree), lghtgbm (light gradient boosting machine), gbm (gradient boosting machine), knn (k-nearest neighbor), svm (support vector machine), mlp (multilayer perceptron), ensmbl1 (lgr + nbc + svm + mlp with soft voting), ensmbl2 (lgr + nbc + svm with soft voting). MCC evaluation of different models after hyperparamter tuning (hpt) lgr_hpt, rf_hpt, nbc_hpt, xgboost_hpt, adaboost_hpt, dct_hpt, lghtgbm_hpt, gbm_hpt, knn_hpt, svm_hpt, mlp_hpt, ensmbl1_hpt (lgr + nbc + svm + mlp with soft voting), ensmbl2_hpt (lgr + nbc + svm with soft voting).

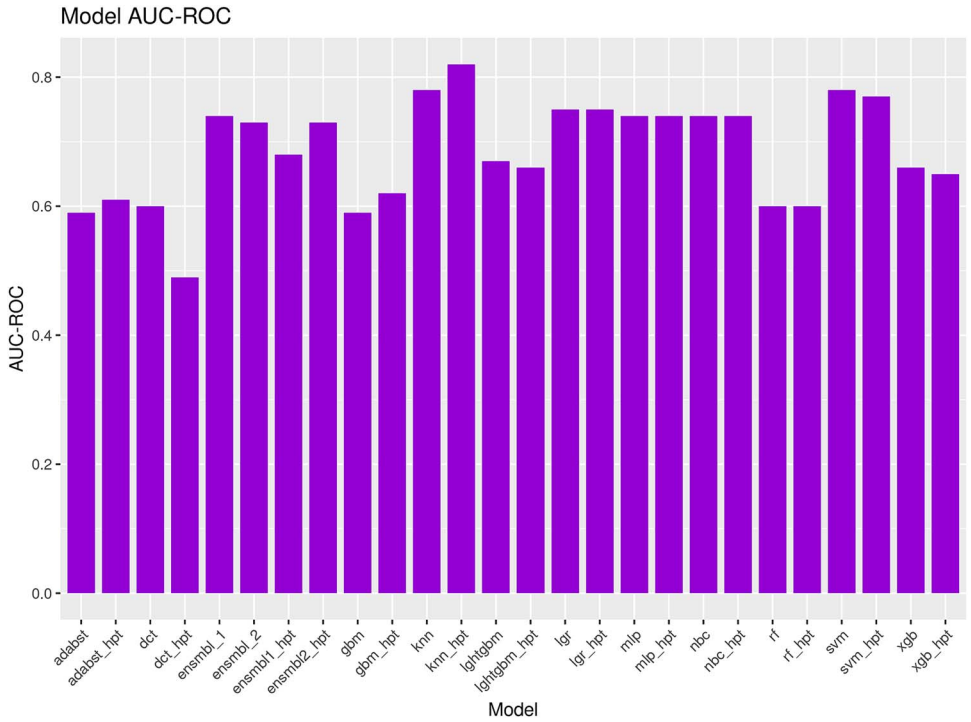

Figure 7. AUC–ROC evaluation of different models before hyperparamter tuning lgr (logistic regression), rf (random forest), nbc (naive Bayes classifier), xgboost (extreme gradient boosting), adaboost (adaptive boosting), dct (decision tree), lghtgbm (light gradient boosting machine), gbm (gradient boosting machine), knn (k-nearest neighbor), svm (support vector machine), mlp (multilayer perceptron), ensmbl1 (lgr + nbc + svm + mlp with soft voting), ensmbl2 (lgr + nbc + svm with soft voting). AUC–ROC evaluation of different models after hyperparamter tuning (hpt)- lgr_hpt, rf_hpt, nbc_hpt, xgboost_hpt, adaboost_hpt, dct_hpt, lghtgbm_hpt, gbm_hpt, knn_hpt, svm_hpt, mlp_hpt, ensmbl1_hpt (lgr + nbc + svm + mlp with soft voting), ensmbl2_hpt (lgr + nbc + svm with soft voting).

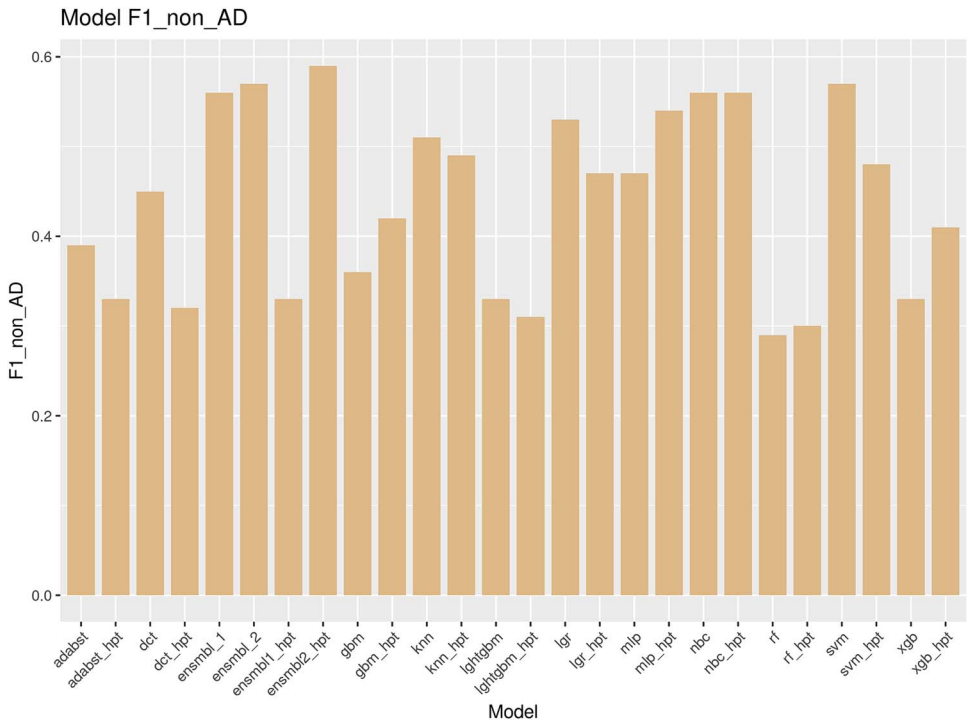

Figure 8. F1 score evaluation (non-AD samples) of different models before hyperparamter tuning lgr (logistic regression), rf (random forest), nbc (naive Bayes classifier), xgboost (extreme gradient boosting), adaboost (adaptive boosting), dct (decision tree), lghtgbm (light gradient boosting machine), gbm (gradient boosting machine), knn (k-nearest neighbor), svm (support vector machine), mlp (multilayer perceptron), ensmbl1 (lgr + nbc + svm + mlp with soft voting), ensmbl2 (lgr + nbc + svm with soft voting). F1 score evaluation (non-AD samples) of different models after hyperparamter tuning (hpt) lgr_hpt, rf_hpt, nbc_hpt, xgboost_hpt, adaboost_hpt, dct_hpt, lghtgbm_hpt, gbm_hpt, knn_hpt, svm_hpt, mlp_hpt, ensmbl1_hpt (lgr + nbc + svm + mlp with soft voting), ensmbl2_hpt (lgr + nbc + svm with soft voting).

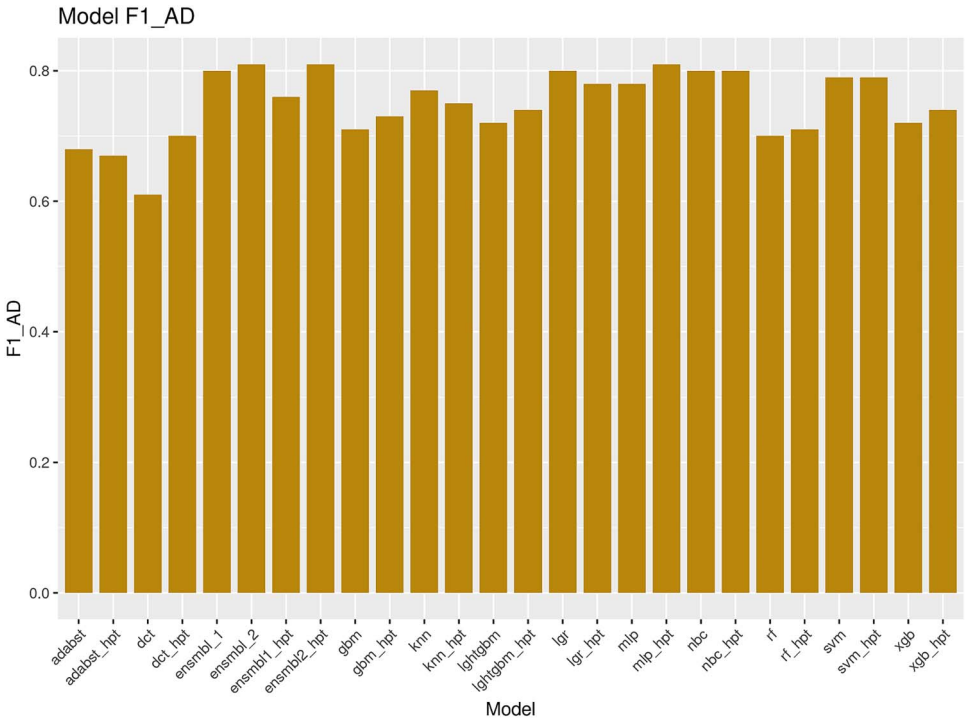

Figure 9. F1 score evaluation (AD samples) of different models before hyperparamter tuning lgr (logistic regression), rf (random forest), nbc (naive Bayes classifier), xgboost (extreme gradient boosting), adaboost (adaptive boosting), dct (decision tree), lghtgbm (light gradient boosting machine), gbm (gradient boosting machine), knn (k-nearest neighbor), svm (support vector machine), mlp (multilayer perceptron), ensmbl1 (lgr + nbc + svm + mlp with soft voting), ensmbl2 (lgr + nbc + svm with soft voting). F1 score evaluation (AD samples) of different models after hyperparamter tuning (hpt) lgr_hpt, rf_hpt, nbc_hpt, xgboost_hpt, adaboost_hpt, dct_hpt, lghtgbm_hpt, gbm_hpt, knn_hpt, svm_hpt, mlp_hpt, ensmbl1_hpt (lgr + nbc + svm + mlp with soft voting), ensmbl2_hpt (lgr + nbc + svm with soft voting).

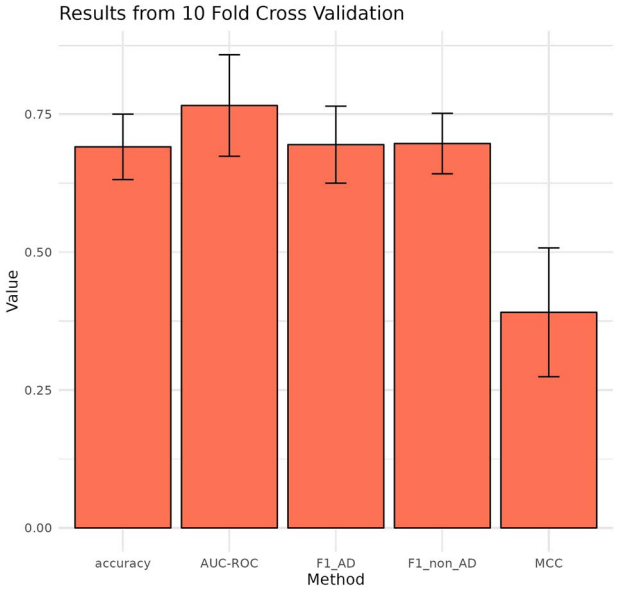

Figure 10. Performance evaluation of the selected model after 10-fold cross-validation with standard deviations. Accuracy (0.691 ± 0.059), MCC (0.391 ± 0.117), AUC–ROC (0.766 ± 0.092), F1_AD (0.695 ± 0.070), F1_non_AD (0.697 ± 0.055).

problems more accurately especially in the biological domain [41–44].

ML algorithms formed the basis of the predictions made in this investigation. The flexibility and power of ML was characterized by the use of a wide variety of algorithms to recognize patterns from RNA-seq data. Following the example of other published studies that systematically experimented with hyperparameters led to improved model performance in our study [45, 46].

A sign of the intricacy of AD categorization is the discovery of trade-offs across various measures (accuracy, precision, recall, F1 score, MCC, and AUC–ROC) used to evaluate the model performance. This is a common practice followed by a number of earlier studies that have emphasized the necessity to use multiple criteria to objectively evaluate classification models instead of relying on a single one [47–49].

It is to be noted that the most promising result of this study is the establishment of a five-gene signature that holds true across all ML models. This signature has the potential to be integrated into a biomarker panel for AD diagnosis. There is a history of such gene signature-based novel diagnostic biomarkers discovery using integrated ML and transcriptomic investigations, for example in the cases of AD [50], breast cancer [47], coronavirus disease-2019 [50], psoriasis [51], tuberculosis [52], and so on.

The set of five genes (*CNKSR1*, *EPHA2*, *CLSPN*, *OLFML3*, and *TARBP1*) identified through our investigation demand closer attention in terms of their relationship with AD. According to the UniProt database, *CNKSR1*, *EPHA2*, *CLSPN*, *OLFML3*, and *TARBP1* encode Connector enhancer of kinase suppressor of ras 1, Ephrin type-A receptor 2, Claspin, olfactomedin-like protein, and probable methyltransferase TARBP1, respectively [53]. Studies have suggested the role of *CNKSR1* in brain development [54], *EPHA2* in axon guidance [55], and *CLSPN* in cell homeostasis [56]. *OLFML3* has been recognized as a microglia-specific gene whose loss of expression disrupts microglia-associated biological functions [57]. Previously, *OLFML3*, *EPHA2*, and *TARBP1* were found to be associated with AD [58–60]. An *in vitro* study showed

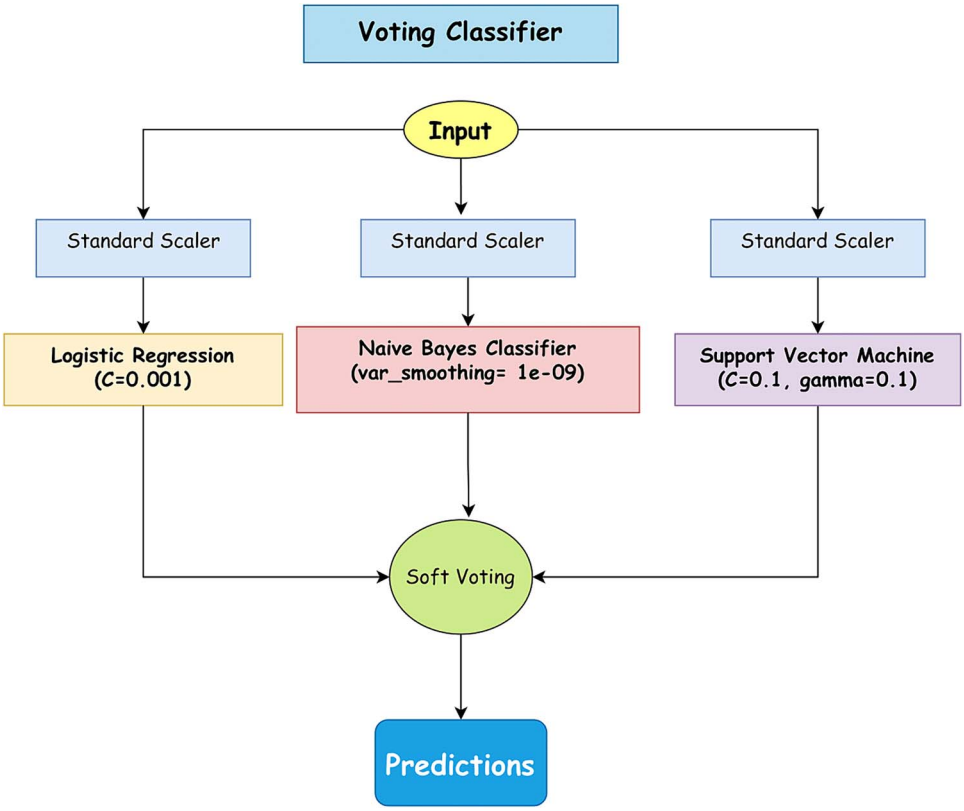

Figure 11. Schematic representation of AITeQ. Following scaling, the input passes through logistic regression, naive Bayes classifier, and SVM with well-defined hyperparameters. The three predictions are then subjected to a soft voting mechanism that makes the final prediction.

that *EPHA2* enhances proinflammatory cytokine release in microglia cells [59]. Olfactomedin-like protein was enriched in amyloid plaque proteome in early onset AD [58]. *TARBP1* was also differentially expressed in other gene expression-based studies [60].

The practical value of this study lies in the discovery of a robust set of predictors that can accurately differentiate between AD patients and healthy people. This is because ML has the ability to accurately identify small alterations in gene expression that might have been unnoticed by conventional analytic techniques. Gene expression signatures offer a more extensive depiction of cellular activity in comparison to individual biomarker testing. The limitations of using biomarkers include the fact that high levels of amyloid are already present in some people who show no symptoms of AD, variability of biomarker profiles over the course of the disease, heterogeneous progression of AD, low levels of biomarkers in the blood, and so on [61]. On the other hand, positron emission tomography (PET) scans, although useful for evaluating brain activity, it provide a less comprehensive image when it comes to AD. PET scans also entail a certain degree of radiation exposure, which might not be suitable for all individuals, especially pregnant women or young children [62]. In this regard, gene signature-based predictions offer accessibility to more people. There is also a potential for misdiagnosis by PET scans as suggested by one study [63]. Hence, the usage of ML to analyze gene signatures as proposed in this study has great potential in enabling safer and more precise detection of AD.

While translating the findings from this study, there are a number of caveats to keep in mind despite the encouraging results. It is a retrospective study that relies on a limited number of datasets. As a result, it has the potential to introduce certain biases that might impact how well the conclusions generalize to new data. Therefore, it is imperative to validate the findings with a larger number of samples collected from a variety of demographics. Another challenge is the variability in RNA-seq data due to biological and technical factors, such as batch effects, sequencing depth, and normalization methods. Batch effects can lead to spurious correlations between genes and disease outcomes, while sequencing depth and normalization methods can affect the accuracy and reproducibility of gene expression measurements [64]. ML algorithms can be sensitive to these factors, and appropriate data preprocessing and normalization methods are necessary to ensure accurate classification results [65]. Apart from all these, the multifaceted nature of neurodegenerative disorders, including but not limited to non-coding RNA-mediated regulations, protein–protein interaction networks, and epigenetic alterations, calls for an approach that goes beyond just focusing on gene expression.

Conclusion

Results from the current study opened up several promising new lines of inquiry. The promise of ML in understanding the complex nature of AD has been demonstrated by its application on disease prediction from RNA-seq data. The importance of a possible biomarker panel for accurate diagnosis of AD is highlighted by the discovery of a consistent five-gene signature. It is crucial to further investigate the functional role played by the identified five-gene signature with respect to AD etiology. The diagnostic potential of the gene signature should be validated in subsequent studies involving a variety of populations through longitudinal investigations.

Key Points

- A set of five genes (*CNKSR1*, *EPHA2*, *CLSPN*, *OLFML3*, and *TARBP1*) were identified following differential gene expression and feature importance analysis.
- Twelve diverse ML algorithms were trained and tested using the gene expression patterns of the identified five genes. The ensemble model consisting of logistic regression, naive Bayes classifier, and SVM with customized hyperparameters was found to be the best-performing model for differentiating AD samples from control.
- AITeQ, a user-friendly, reliable, and accurate ML framework for AD prediction was developed based on the five-gene signature.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Funding

This research work did not receive any funding.

Data availability

The data generated in this study are included within the manuscript and the supplementary files.

Code availability

Code for using the AITeQ model is available at: https://github.com/ishtiaque-ahammad/AITeQ.

Author contributions

Ishtiaque Ahammad, Anika Bushra Lamisa, Arittra Bhattacharjee, and Md. Shamsul Arefin trained and evaluated the ML models and developed the AITeQ framework. Anika Bushra Lamisa and Tabassum Binte Jamal conducted the differential gene expression analysis and data preprocessing for the ML models. Ishtiaque Ahammad and Anika Bushra Lamisa wrote the original draft of the manuscript. Zeshan Mahmud Chowdhury, Mohammad Uzzal Hossain, and Keshob Chandra Das reviewed and edited the manuscript. Md. Salimullah and Chaman Ara Keya supervised the research project.

References

1. Twine NA, Janitz K, Wilkins MR. *et al.* Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. *PloS One* 2011;**6**:e16266. https://doi.org/10.1371/journal.pone.0016266
2. Vadapalli S, Abdelhalim H, Zeeshan S. *et al.* Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. *Brief Bioinform* 2022;**23**:bbac191. https://doi.org/10.1093/bib/bbac191
3. Wenric S, Shemirani R. Using supervised learning methods for gene selection in RNA-seq case-control studies. *Front Genet* 2018;**9**:297. https://doi.org/10.3389/fgene.2018.00297
4. Choi SH, Labadorf AT, Myers RH. *et al.* Evaluation of logistic regression models and effect of covariates for case–control study in RNA-seq analysis. *BMC Bioinformatics* 2017;**18**:91. https://doi.org/10.1186/s12859-017-1498-y
5. Zhang F, Petersen M, Johnson L. *et al.* Recursive support vector machine biomarker selection for Alzheimer's disease. *J Alzheimers Dis* 2021;**79**:1691–700. https://doi.org/10.3233/JAD-201254
6. Kumar A, Singh TR. A new decision tree to solve the puzzle of Alzheimer's disease pathogenesis through standard diagnosis scoring system. *Interdiscip Sci Comput Life Sci* 2017;**9**:107–15. https://doi.org/10.1007/s12539-016-0144-0
7. Velazquez M, Lee Y, Alzheimer's Disease Neuroimaging Initiative. Random forest model for feature-based Alzheimer's disease conversion prediction from early mild cognitive impairment subjects. *PloS One* 2021;**16**:e0244773. https://doi.org/10.1371/journal.pone.0244773
8. Kruthika KR, Rajeswari MHD. *et al.* Multistage classifier-based approach for Alzheimer's disease prediction and retrieval. *Inform Med Unlocked* 2019;**14**:34–42. https://doi.org/10.1016/j.imu.2018.12.003
9. Ahmed H, Soliman H, Elmogy M. Early detection of Alzheimer's disease using single nucleotide polymorphisms analysis based on gradient boosting tree. *Comput Biol Med* 2022;**146**:105622. https://doi.org/10.1016/j.compbiomed.2022.105622
10. Li J, Zhang Y, Lu T. *et al.* Identification of diagnostic genes for both Alzheimer's disease and metabolic syndrome by the machine learning algorithm. *Front Immunol* 2022;**13**:1037318. https://doi.org/10.3389/fimmu.2022.1037318
11. Yuen SC, Liang X, Zhu H. *et al.* Prediction of differentially expressed microRNAs in blood as potential biomarkers for Alzheimer's disease by meta-analysis and adaptive boosting ensemble learning. *Alzheimers Res Ther* 2021;**13**:126. https://doi.org/10.1186/s13195-021-00862-z
12. Ludwig N, Fehlmann T, Kern F. *et al.* Machine learning to detect Alzheimer's disease from circulating non-coding RNAs. *Genomics Proteomics Bioinformatics* 2019;**17**:430–40. https://doi.org/10.1016/j.gpb.2019.09.004
13. Bevilacqua R, Barbarossa F, Fantechi L. *et al.* Radiomics and artificial intelligence for the diagnosis and monitoring of Alzheimer's disease: a systematic review of studies in the field. *J Clin Med* 2023;**12**:5432. https://doi.org/10.3390/jcm12165432
14. Feng Q, Ding Z. MRI radiomics classification and prediction in Alzheimer's disease and mild cognitive impairment: a review. *Curr Alzheimer Res* 2020;**17**:297–309. https://doi.org/10.2174/1567205017666200303105016
15. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;**30**:207–10. https://doi.org/10.1093/nar/30.1.207
16. Babraham Bioinformatics. *FastQC A Quality Control Tool for High Throughput Sequence Data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (23 May 2024, date last accessed).
17. Guo J, Gao J, Liu Z. HISAT2 parallelization method based on spark cluster. *J Phys Conf Ser* 2022;**2179**:012038. https://doi.org/10.1088/1742-6596/2179/1/012038
18. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30. https://doi.org/10.1093/bioinformatics/btt656
19. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. https://doi.org/10.1186/s13059-014-0550-8

20. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* 1995;**57**:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
21. Ritchie ME, Phipson B, Wu D. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47–7. https://doi.org/10.1093/nar/gkv007
22. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics* 2013;**14**:106. https://doi.org/10.1186/1471-2105-14-106
23. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32. https://doi.org/10.1023/A:1010933404324
24. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001;**29**:1189–1232. https://doi.org/10.1214/aos/1013203451
25. Zeng X, Chen Y-W, Tao C. Feature selection using recursive feature elimination for handwritten digit recognition. In: *2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. 2009, p. 1205–8. Kyoto: IEEE. https://doi.org/10.1109/IIH-MSP.2009.145
26. Muthukrishnan R, Rohini R. LASSO: a feature selection technique in predictive modeling for machine learning. In: *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*. 2016, p. 18–20. Coimbatore, India: IEEE. https://doi.org/10.1109/ICACA.2016.7887916
27. Raju VNG, Lakshmi KP, Jain VM. *et al.* Study the influence of normalization/transformation process on the accuracy of supervised classification. In: *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. 2020, p. 729–35. Tirunelveli, India: IEEE. https://doi.org/10.1109/ICSSIT48917.2020.9214160
28. Cox DR. The regression analysis of binary sequences. *J R Stat Soc Ser B Methodol* 1958;**20**:215–32. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x
29. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97. https://doi.org/10.1007/BF00994018
30. Quinlan JR. Induction of decision trees. *Mach Learn* 1986;**1**:81–106. https://doi.org/10.1007/BF00116251
31. Zhang H, Su J. Naive Bayesian classifiers for ranking. In: *Machine Learning: ECML 2004*, vol. **3201**, J-F Boulicaut, F Esposito, F Giannotti, D Pedreschi (eds). *Lecture Notes in Computer Science*, Vol. 3201. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, 501–12. https://doi.org/10.1007/978-3-540-30115-8_46
32. Mucherino A, Papajorgji PJ, Pardalos PM. K-nearest neighbor classification. In: *Data Mining in Agriculture*, Vol. **34**. Springer Optimization and Its Applications, New York, NY: Springer New York, 2009, 83–106. https://doi.org/10.1007/978-0-387-88615-2_4
33. Li R, Liao B, Wang B. *et al.* Identification of tumor tissue of origin with RNA-seq data and using gradient boosting strategy. *Biomed Res Int* 2021;**2021**:1–14. https://doi.org/10.1155/2021/6653793
34. Cao Y, Miao Q-G, Liu J-C. *et al.* Advance and prospects of AdaBoost algorithm. *Acta Autom Sin* 2013;**39**:745–58. https://doi.org/10.1016/S1874-1029(13)60052-X
35. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. p. 785–94. San Francisco California USA: ACM, 2016. https://doi.org/10.1145/2939672.2939785
36. Ke G, Meng Q, Finley T. *et al.* LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. p. 3149–57. Red Hook, NY, USA: Curran Associates Inc., 2017.
37. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 1958;**65**:386–408. https://doi.org/10.1037/h0042519
38. LaValle SM, Branicky MS, Lindemann SR. On the relationship between classical grid search and probabilistic roadmaps. *Int J Robot Res* 2004;**23**:673–92. https://doi.org/10.1177/0278364904045481
39. Pedregosa F, Varoquaux G, Gramfort A. *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2012;**12**:2825–2830.
40. Carneiro T, Medeiros Da Nobrega RV, Nepomuceno T. *et al.* Performance analysis of Google colaboratory as a tool for accelerating deep learning applications. IEEE. *Access* 2018;**6**:61677–85. https://doi.org/10.1109/ACCESS.2018.2874767
41. Lokeswari YV, Jacob SG. Prediction of child tumours from microarray gene expression data through parallel gene selection and classification on spark. In: *Computational Intelligence in Data Mining*, Vol. **556**, HS Behera, DP Mohapatra (eds). Singapore: Springer Singapore, 2017, 651–61. https://doi.org/10.1007/978-981-10-3874-7_62
42. Alghunaim S, Al-Baity HH. On the scalability of machine-learning algorithms for breast cancer prediction in big data context. *IEEE Access* 2019;**7**:91535–46. https://doi.org/10.1109/ACCESS.2019.2927080
43. Matamala N, Vargas MT, González-Cámpora R. *et al.* Tumor microRNA expression profiling identifies circulating microRNAs for early breast cancer detection. *Clin Chem* 2015;**61**:1098–106. https://doi.org/10.1373/clinchem.2015.238691
44. Rana P, Thai P, Dinh T. *et al.* Relevant and non-redundant feature selection for cancer classification and subtype detection. *Cancer* 2021;**13**:4297. https://doi.org/10.3390/cancers13174297
45. Le H, Peng B, Uy J. *et al.* Machine learning for cell type classification from single nucleus RNA sequencing data. *PloS One* 2022;**17**:e0275070. https://doi.org/10.1371/journal.pone.0275070
46. Jin T, Nguyen ND, Talos F. *et al.* ECMarker: interpretable machine learning model identifies gene expression biomarkers predicting clinical outcomes and reveals molecular mechanisms of human disease in early stages. *Bioinformatics* 2021;**37**:1115–24. https://doi.org/10.1093/bioinformatics/btaa935
47. Mirza Z, Ansari MS, Iqbal MS. *et al.* Identification of novel diagnostic and prognostic gene signature biomarkers for breast cancer using artificial intelligence and machine learning assisted transcriptomics analysis. *Cancer* 2023;**15**:3237. https://doi.org/10.3390/cancers15123237
48. Dessie EY, Gautam Y, Ding L. *et al.* Development and validation of asthma risk prediction models using co-expression gene modules and machine learning methods. *Sci Rep* 2023;**13**:11279. https://doi.org/10.1038/s41598-023-35866-2
49. Zhang X, Jonassen I, Goksøyr A. Machine learning approaches for biomarker discovery using gene expression data. In: Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of Sao Paulo, Sao Paulo, Brazil, HI Nakaya (eds). *Bioinformatics*. Exon Publications, 2021, 53–64. https://doi.org/10.36255/exonpublications.bioinformatics.2021.ch4
50. Lai G, Liu H, Deng J. *et al.* A novel 3-gene signature for identifying COVID-19 patients based on bioinformatics and machine learning. *Genes* 2022;**13**:1602. https://doi.org/10.3390/genes13091602
51. Le NQK, Do DT, Nguyen T-T-D. *et al.* Identification of gene expression signatures for psoriasis classification using machine learning techniques. *Med Omics* 2021;**1**:100001. https://doi.org/10.1016/j.meomic.2020.100001
52. DiNardo AR, Gandhi T, Heyckendorf J. *et al.* Gene expression signatures identify biologically and clinically distinct

tuberculosis endotypes. *Eur Respir J* 2022;**60**:2102263. https://doi.org/10.1183/13993003.02263-2021
53. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* Jan. 2023;**51**:D523–31. https://doi.org/10.1093/nar/gkac1052
54. Kazeminasab S, Taskiran II, Fattahi Z. *et al.* CNKSR1 gene defect can cause syndromic autosomal recessive intellectual disability. *Am J Med Genet Part B Neuropsychiatr Genet* 2018;**177**:691–9. https://doi.org/10.1002/ajmg.b.32648
55. Imondi R, Wideman C, Kaprielian Z. Complementary expression of transmembrane ephrins and their receptors in the mouse spinal cord: a possible role in constraining the orientation of longitudinally projecting axons. *Development* 2000;**127**:1397–410. https://doi.org/10.1242/dev.127.7.1397
56. Azenha D, Hernandez-Perez S, Martin Y. *et al.* Implications of CLSPN variants in cellular function and susceptibility to cancer. *Cancer* 2020;**12**. https://doi.org/10.3390/cancers12092396
57. Butovsky O, Jedrychowski MP, Moore CS. *et al.* Identification of a unique TGF-β dependent molecular and functional signature in microglia. *Nat Neurosci* 2014;**17**:131–43. https://doi.org/10.1038/nn.3599
58. Drummond E, Kavanagh T, Pires G. *et al.* The amyloid plaque proteome in early onset Alzheimer's disease and down syndrome. *Acta Neuropathol Commun* 2022;**10**:53. https://doi.org/10.1186/s40478-022-01356-1
59. Ma X, Zhang Y, Gou D. *et al.* Metabolic reprogramming of microglia enhances proinflammatory cytokine release through EphA2/p38 MAPK pathway in Alzheimer's disease. *J Alzheimers Dis* 2022;**88**:771–85. https://doi.org/10.3233/JAD-220227
60. Gns HS, Rajalekshmi SG, Burri RR. Revelation of pivotal genes pertinent to Alzheimer's pathogenesis: a methodical evaluation of 32 GEO datasets. *J Mol Neurosci* 2022;**72**:303–22. https://doi.org/10.1007/s12031-021-01919-2
61. Omar SH, Preddy J. Advantages and pitfalls in fluid biomarkers for diagnosis of Alzheimer's disease. *J Pers Med* 2020;**10**:63. https://doi.org/10.3390/jpm10030063
62. Bao W, Xie F, Zuo C. *et al.* PET neuroimaging of Alzheimer's disease: radiotracers and their utility in clinical research. *Front Aging Neurosci* 2021;**13**. https://doi.org/10.3389/fnagi.2021.624330
63. Shipley SM, Frederick MC, Filley CM. *et al.* Potential for misdiagnosis in community-acquired PET scans for dementia. *Neurol Clin Pract* 2013;**3**:305–12. https://doi.org/10.1212/CPJ.0b013e318296f2df
64. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* 2019;**20**:296. https://doi.org/10.1186/s13059-019-1874-1
65. Zhang Y, Patil P, Johnson WE. *et al.* Robustifying genomic classifiers to batch effects via ensemble learning. *Bioinformatics* 2021;**37**:1521–7. https://doi.org/10.1093/bioinformatics/btaa986