ORIGINAL ARTICLE
eISSN 2005-7571
Korean J Biol Psychiatry 2023;30(1):24-30
https://doi.org/10.22857/kjbp.2023.30.1.004

Identification of Combined Biomarker for Predicting Alzheimer's Disease Using Machine Learning

Ki-Yeol Kim, PhD

Oral Cancer Research Institute,Yonsei University College of Dentistry, Yonsei University, Seoul, Korea

Objectives Alzheimer's disease (AD) is the most common form of dementia in older adults, damaging the brain and resulting in impaired memory, thinking, and behavior. The identification of differentially expressed genes and related pathways among affected brain regions can provide more information on the mechanisms of AD. The aim of our study was to identify differentially expressed genes associated with AD and combined biomarkers among them to improve AD risk prediction accuracy.

Methods Machine learning methods were used to compare the performance of the identified combined biomarkers. In this study, three publicly available gene expression datasets from the hippocampal brain region were used.

Results We detected 31 significant common genes from two different microarray datasets using the limma package. Some of them belonged to 11 biological pathways. Combined biomarkers were identified in two microarray datasets and were evaluated in a different dataset. The performance of the predictive models using the combined biomarkers was superior to those of models using a single gene. When two genes were combined, the most predictive gene set in the evaluation dataset was *ATR* and *PRKCB* when linear discriminant analysis was applied.

Conclusions Combined biomarkers showed good performance in predicting the risk of AD. The constructed predictive nomogram using combined biomarkers could easily be used by clinicians to identify high-risk individuals so that more efficient trials could be designed to reduce the incidence of AD.

Keywords Alzheimer's disease; Risk prediction; Gene expression; Combined biomarker; Machine learning.

Received: December 11, 2022 / **Revised:** January 12, 2023 / **Accepted:** January 30, 2023
Address for correspondence: Ki-Yeol Kim, PhD
Oral Cancer Research Institute, Yonsei University College of Dentistry, Yonsei University, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea
Tel: +82-2-2228-3043, **Fax:** +82-2-392-2959, **E-mail:** kky1004@yuhs.ac

Introduction

Alzheimer's disease (AD) is the most common cause of dementia in older adults with the loss of cognitive function and memory.[1] The common symptoms of AD are difficulties in remembering recent events, thinking and reasoning, speaking and writing, making judgment and decisions, planning and performing familiar tasks, and changes in personality and behavior.[2]

AD is not a part of normal aging, but increasing age is the strongest risk factor for AD.[3] Three in 10 people over the age of 85 years and one in every eight people over 65 years are estimated to develop AD.[4] Family history and genetics, mild cognitive impairment (MCI), past head trauma, lifestyle, heart health, lifelong learning, and social engagement are other risk factors of AD.[5] The risk of developing AD is higher if a first-degree relative (parent or siblings) has the disease.[6] The genetic mechanism of AD among families remains unexplained.[7] People with MCI have a higher chance of developing AD but is not a certainty and can be prevented by developing a healthy lifestyle.[8] Some studies showed that the risk factors of heart disease may also increase the risk of developing AD.[9] The diagnosis of AD is usually based on the patient's medical history, mental status testing, and physical testing, even though several histopathological markers, such as extracellular β-amyloid plaques and neurofibrillary tangles within neurons, can determine AD presence.[10]

Artificial intelligence (AI) is a method of simulating human intelligence processes using machines, especially computer systems. It can be used to analyze and improve the predictive performance of models in various research areas. Machine learning (ML), a major branch of AI, has been widely used.[11-15]

In this study, we investigated significant gene sets related to

AD and identified combined biomarkers from selected significant gene sets. We also performed functional annotations for AD-related genes. We aimed to provide a systematic approach to discovering new therapeutic targets for the treatment of AD. A risk-predictive nomogram was constructed for practical usage. This nomogram could be used as an objective guideline to assess high risk for AD. With the identification of high-risk individuals, more efficient therapeutic trials can be designed to reduce the incidence of AD.

Methods

Data preparation

Three publicly available gene expression datasets (GSE5281, GSE1297, and GSE48350) were used in this study. These datasets are accessible from a public microarray database (gene expression omnibus [GEO]). GSE5281 and GSE48350 consist of six brain regions and four brain regions (a total of 54675 probes), respectively. These datasets were conducted in the same GPL570 platform, Affy HG-U133 plus 2.0 (http://www.affymetrix.com/analysis/index.affx). GSE1297 includes hippocampal (HIP) gene expression data on nine controls and 22 AD patients of varying severity, incipient, and moderate and severe (a total of 22283 probes). This dataset was conducted in the GPL96 platform, Affy HG-U133A. In this study, we considered the HIP region in GSE5271 and GSE48350 and the severe stage in GSE1297. The datasets are summarized in Table 1.

Gene ontology

Gene ontology (GO)[16] is a structured, defined, and controlled vocabulary for large-scale gene annotation. The Database for Annotation, Visualization, and Integrated Discovery (DAVID) was developed as a comprehensive functional annotation tool for relating functional terms to gene lists using a clustering algorithm.[17] To analyze differentially expressed genes at the functional level, we performed pathway enrichment analysis using the online DAVID tool (https://david.ncifcrf.gov/).

GO hierarchy contains three subontologies: biological processes, cellular components, and molecular functions. The p-value is a modified Fisher's exact test p-value. Because DAVID examines thousands of gene sets, it is necessary to test multiple hypotheses. DAVID provides a Benjamini-Hochberg false discovery rate (FDR)-adjusted p-value, with a smaller p-value indicating enrichment. We used the p-value and Benjamini-Hochberg FDR to determine the significance of the enrichment of the terms for each annotation.

ML methods

In recent decades, rapid advancements in computational algorithms and the increased availability of big data have enabled AI, one of the most exciting technologies in our everyday lives, to analyze and improve the predictive performance of models in various research areas. Specifically, ML, a major branch of AI, has been used widely. The ML algorithms used in this study are as follows.[18]

Linear discriminant analysis

Linear discriminant analysis (LDA) is a generalization of Fisher's linear discriminant. It is a method used in statistics and other fields to find a linear combination of features that can characterize or separate two or more classes of objects or events. The resulting combination might be used as a linear classifier or for dimensionality reduction before classification. LDA is a kind of dimension-reduction technique commonly applied to supervised classification problems. It is used to model differences between groups, i.e., separating two or more groups from each other.[19]

k-Nearest neighbors algorithm

The k-nearest neighbors (KNN) algorithm is one of the simplest techniques used in ML.[20] It is used for both classification and regression and is preferred by many in the industry because of its ease of use and low calculation time. The KNN algorithm works by finding the distance between data points. The most common way to find this distance is to use the Euclidean distance. KNN computes the distance between each data point and the test data. It then finds the probability of these points being similar to the test data and classifies the data points based on which of them share the highest probability.[21]

Table 1. Summarization of datasets used in the study

	GSE5281		GSE1297		GSE48350	
Platform	GPL570, Affy HG-U133 plus 2.0		GPL96, Affy HG-U133A		GPL570, Affy HG-U133 plus 2.0	
Number of probes	54675 probes		22283 probes		54675 probes	
Group (samples)	Normal (13)	AD (10)	Normal (9)	Severe AD (7)	Normal (23)	AD (18)
Age (yr, mean±SD)	79.6±9.4	77.8±5.7	85.3±2.7	84±4.0	83.7±9.0	84.2±6.8
Gender, female/male	3/10	4/6	2/7	5/2	11/12	9/9

AD, Alzheimer's disease; SD, standard deviation

Support vector machine

A support vector machine (SVM) is a supervised learning model with associated learning algorithms that can analyze data for classification and regression analysis.[22] The SVM algorithm is a popular ML tool that offers solutions to both classification and regression problems. It was developed at AT&T Bell Laboratories by Vapnik and colleagues.[22] The objective of the SVM algorithm is to find a hyperplane in the N-dimensional space (where N is the number of features) that distinctly classifies the data points. Support vectors are data points that are closer to the hyperplane and can influence the position and orientation of the hyperplane. Using these support vectors, the margin of the classifier is maximized. Deleting the support vectors will change the position of the hyperplane.

Random forest

Random forest (RF) is an ensemble learning method for classification, regression, and other tasks. It operates by constructing a multitude of decision trees at the training time, and outputting the class that is the mode of the classes (classification) or the mean/average prediction (regression) of individual trees.[23] RF algorithms can be used to solve both regression and classification problems, making it a diverse model that is widely used by engineers.

All ML models were implemented using R programming language, version 4.1.3 (R Foundation for Statistical Computing, Vienna, Austria),[1] including GEOquery and limma packages for downloading the GEO datasets and identifying significant genes. The nomogram for predicting AD risk was created based on the significant genes selected.

In this study, we compared the accuracies of ML algorithms for predicting AD risk using the identified combined biomarkers. The study process is shown in Fig. 1.

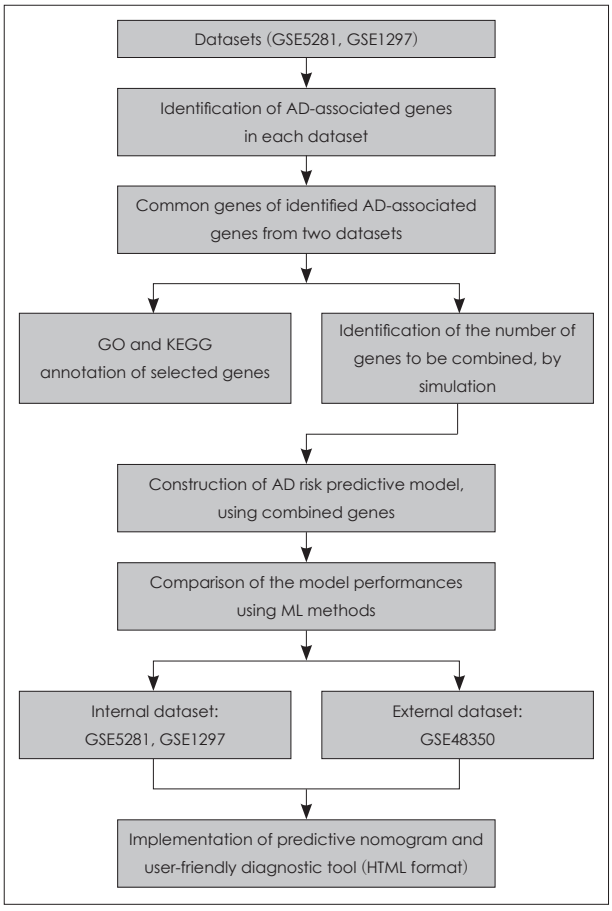

Fig. 1. Study design. The data for duplicated genes in each gene expression dataset were averaged. AD-associated genes were identified in each dataset and the common genes were obtained from these datasets. AD, Alzheimer's disease; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; ML, machine learning; HTML, Hypertext Markup Language format.

Results

Expression patterns for identifying differentially expressed genes

We detected 31 differently expressed common genes between the normal and AD groups using the limma package. Fig. 2 shows the expression patterns of the 31 common genes in the three different datasets (Fig. 2).

The identified genes showed highly divergent expression patterns between the normal and AD groups in the GSE5281 and GSE1297 datasets but not in the GSE48350 dataset (Fig. 2C). The upregulated genes in the AD group included *COX7C*, *DOCK3*, *CDK5*, *SLC25A12*, *PLK2*, *IQSEC1*, *NDUFB8*, *CAPRIN2*, *WDFY3*, and *PKP4*. On the other hands, *RXF2*, *TTN*, *RHOQ*, *SFTPB*, and *KIAA0485* were downregulated in the AD groups (Fig. 2A and B).

GO and pathway analysis

We next performed GO term annotation and pathway enrichment analysis of the differentially expressed genes at the functional level using DAVID tool. The results are summarized in Table 2.

The AD-related genes were *APC*, *NDUFB8*, *CDK5*, *COX7C*, and *ITPR1*. The Huntington's disease-related pathway was also included, although the FDR-adjusted p-value was not significant. The pathway significance may have been underestimated by selecting a small number of common genes (31 common genes were used for annotation). The most significant enriched GO term was "serine/threonine-protein kinase" (Table 2).

Identification of combined predictive markers of AD risk

To select the optimal number of combined biomarkers for risk

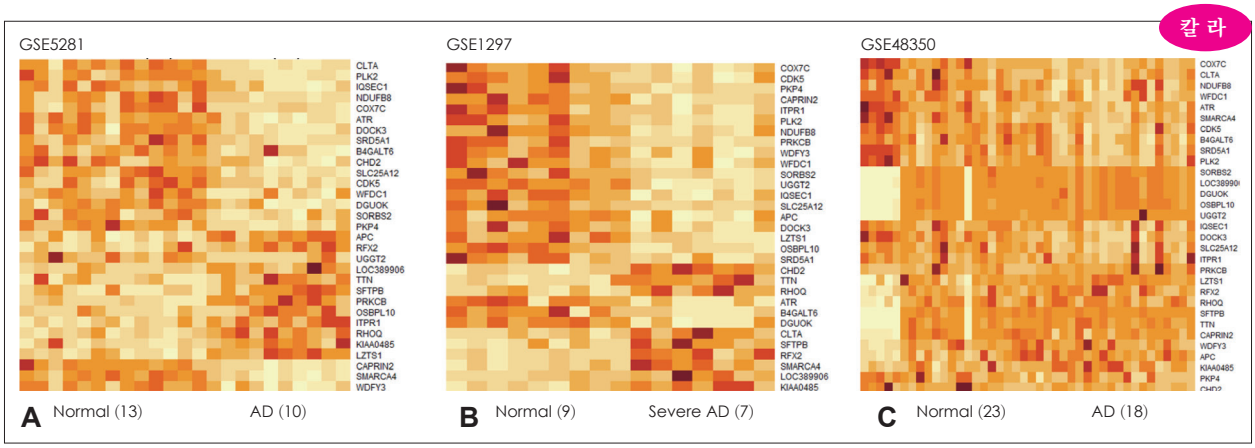

Fig. 2. Expression patterns of 31 genes common in the two different datasets. A: Gene expression pattern in GSE5281 between the normal and severe AD groups. B: Gene expression pattern in GSE1297 between the normal and severe AD groups. C: Gene expression pattern in GSE48350 between the normal and AD groups. The brighter the color, the more upregulated the gene. The dark orange color indicates downregulation. AD, Alzheimer's disease.

Table 2. Summary of GO terms identified using the DAVID annotation database

Category	Term	Count	p-value*	Benjamini†
KEGG_PATHWAY	Pathways of neurodegeneration-multiple diseases	6	1.7E-3	2.1E-1
KEGG_PATHWAY	Alzheimer's disease	5	5.6E-3	3.6E-1
KEGG_PATHWAY	Huntington's disease	4	2.0E-2	8.1E-1
KEGG_PATHWAY	Retrograde endocannabinoid signaling	3	3.3E-2	9.1E-1
UP_KEYWORDS	Serine/threonine-protein kinase	5	2.2E-3	4.0E-2
GOTERM_BP_DIRECT	Peptidyl-serine phosphorylation	4	2.3E-3	5,1E-1
GOTERM_BP_DIRECT	Regulation of synaptic plasticity	3	2.8E-3	5,1E-1
GOTERM_CC_DIRECT	Postsynaptic density	5	4.5E-4	6.1E-2
GOTERM_CC_DIRECT	Perinuclear region of the cytoplasm	5	1.9E-2	9.4E-1
GOTERM_MF_DIRECT	Protein serine/threonine-protein kinase activity	5	2.6E-3	3.3E-1
GOTERM_MF_DIRECT	ATP binding	8	6.3E-3	4.0E-1

*p-value, modified Fisher's exact test p-value; †Benjamini, Benjamini-Hochberg false discovery rate adjusted p-value. GO, gene ontology; DAVID, Database for Annotation, Visualization, and Integrated Discovery; KEGG, Kyoto Encyclopedia of Genes and Genomes; BP, biological processe; CC, cellular component; MF, molecular function

prediction, we tested random sets of 1–5 genes, with 200–500 replicates and evaluated the association between the number of genes and the predictive model accuracy (Fig. 3).

Predictive accuracy indicates the probability of the concordance between the predicted and observed responses (logistic regression). The accuracy increased with increasing numbers of combined genes. We focused on identifying a predictive model based on the least number of genes. Therefore, we selected a two-gene set for further analysis. Table 3 summarizes the five combined sets of two genes resulting from the simulations shown in Fig. 3.

Performance comparison of AD risk predictive models

We then compared the performance of the risk-predictive models using different ML algorithms. For this experiment, the dataset was randomly split into training (70% of data) and testing (30% of data) datasets. The random dataset spilt was processed repeatedly 100 times, and the model performance was summarized according to the mean values and standard deviations calculated for all processing cycles (Table 4).

We compared the performance of two model types, one predicting the risk probability of AD based on a single gene and the other predicting the risk based on combined biomarkers. The performance of the models based on combined biomarkers was superior to that of the models based on single genes (Table 4). The predictive accuracies ranged from 0.814 to 1.000 for the training dataset and from 0.708 to 0.803 for the testing dataset for single gene models. For models based on combined biomarkers, RF was the best-performing model with the training dataset (accuracy = 1.0). Performance with the test datasets tended to depend on the combined biomarkers used.

When we tested the models in GSE1297 and GSE48350, the per-

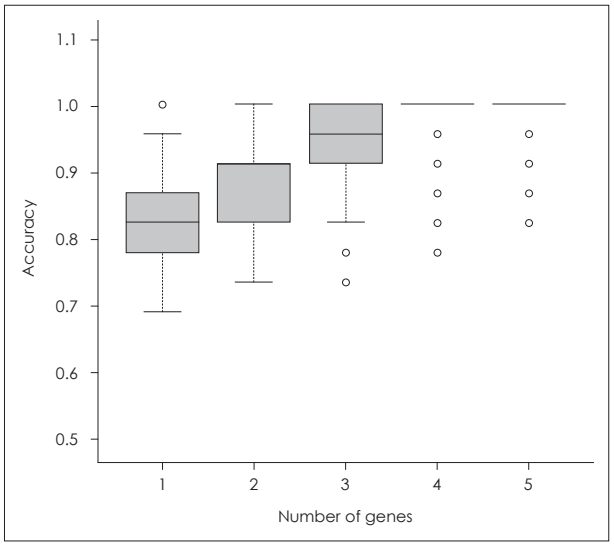

Fig. 3. Comparison of predictive accuracies using a combination of different numbers of genes among the 31 common significant genes. The vertical and horizontal axes represent the predictive accuracy and number of genes combined, respectively.

Table 3. Summary of five combined gene sets

Gene set	Gene symbol	Description
1	*CHD2*	Chromodomain helicase DNA-binding protein 2
	DOCK3	Dedicator of cytokinesis 3
2	*CAPRIN2*	Caprin family member 2
	COX7C	Cytochrome c oxidase subunit 7C
3	*LZTS1*	Leucine zipper tumor suppressor 1
	SORBS2	Sorbin and SH3 domain containing 2
4	*ITPR1*	ITPR1 antisense RNA 1 (head to head)
	IQSEC1	IQ motif and Sec7 domain 1
5	*ATR*	ATR serine/threonine kinase
	PRKCB	PDS5 cohesin-associated factor B

formance in GSE1297 was much better than in GSE48350. This can be interpreted as a characteristic of the dataset as GSE1297 consists of normal and severe AD (Table 1). Considering the accuracy of the three datasets, we constructed a predictive nomogram by combining *LZTS1* and *SORBS2* genes.

Nomogram

A nomogram was constructed to predict the AD risk probability using *ATR* and *PRKCB* (Fig. 4A) as the best combination of two genes for predicting the risk of AD (Table 4). The predictive accuracy was 80.9% in the GSE48350 dataset using the LDA model.

ATR and *PRKCB* are genes upregulated and downregulated in AD, respectively. The risk probability for AD increased when the total nomogram points decreased. If the total points were 60 for a patient, the AD risk probability was about 50% (Fig. 4A). For practical usage of the nomogram, we constructed a nomogram in Hypertext Markup Language format and populated it with calculated total scores and probabilities (Fig. 4B). Fig. 4B shows that the AD risk probability of a patient with expression values of *ATR* and *PRKCB* of 0.6 and 0.4, respectively, was 32.5%. Additionally, the total points calculated by the nomogram could be used for stratifying patients according to AD risk probability.

Discussion

This study identified 31 significant genes for AD risk prediction. Among them, *Lzts1* was reported to control both neuronal delamination and outer radial glial-like cell generation during mammalian cerebral development, suggesting a role in neuronal development.[24)25)]

The neurodevelopmental spectrum was seen with *CHD2* variants.[26)] *DOCK3*-related neurodevelopmental syndrome was reported in a boy with developmental delay and hypotonia.[27)] Genetic variants in *SLC9A9* were associated with measures of attention-deficit/hyperactivity disorder symptoms in families.[28)]

Kirtay et al.[29)] identified a physiological function of *ATR* beyond its DNA damage response role, in regulating neuronal activity. Gerschütz et al.[30)] reported that *PRKCB* and *MAPK1* were increased in the late AD stages. *MAPK1* and *PRKCB* levels were low in the brainstem and cerebellum. The authors proposed that alterations in the expression of these two genes occurred early in the pathogenesis of AD in a region-specific manner. Antonell et al.[31)] and Zhou et al.[32)] also reported that the low expression of *PRKCB* was a potential causative factor of AD. This study also confirmed that *PRKCB* was downregulated in the AD group (Fig. 2A and B).

While ML can easily identify dataset trends and patterns, it requires massive datasets for training. Due to the small sample size, the limitation of this study was that the dataset could not represent the entire population of patients with AD. A model trained on a random sample of a dataset might have poor generalizability and perform poorly outside of that sample. Indeed, the use of larger training and test sets resulted in more accurate and reliable predictions.[33)]

The implemented predictive model was presented in the form of a diagram, referred to as a nomogram. A nomogram is a graphical representation of a statistical model. It provides the probability of a particular clinical outcome. The nomogram introduced in this study can serve as an objective guideline to assess high risk for AD. With the identification of high-risk individuals, more efficient trials can be designed to reduce the incidence of AD. The

Table 4. Comparison of the predictive accuracy of the models with the training, testing datasets, and two independent datasets

	GSE5281 (13 normal, 10 AD)								GSE1297				GSE48350			
	Training data (100 times)				Test data				(9 normal, 7 severe AD)				(23 normal, 18 AD)			
Single genes (31 common significant genes)																
	LDA	KNN	SVM	RF	LDA	KNN	SVM	RF	LDA	KNN	SVM	RF	LDA	KNN	SVM	RF
	0.826	0.829	0.814	1.000	0.803	0.790	0.708	0.740	0.832	0.860	0.868	1.000	0.678	0.730	0.729	1.000
	(0.089)	(0.086)	(0.133)	(0.000)	(0.161)	(0.166)	(0.180)	(0.182)	(0.090)	(0.086)	(0.107)	(0.000)	(0.079)	(0.055)	(0.083)	(0.000)
Combined gene sets																
1 *CHD2*	0.882	0.968	0.999	1.000	0.911	0.953	0.996	0.990	1.000	0.875	1.000	1.000	0.761	0.761	0.809	1.000
DOCK3	(0.059)	(0.051)	(0.005)	(0.000)	(0.096)	(0.082)	(0.023)	(0.057)								
2 *CAPRIN2*	0.957	0.845	0.996	1.000	0.963	0.848	0.898	1.000	0.956	0.956	1.000	1.000	0.738	0.833	0.857	1.000
COX7C	(0.026)	(0.085)	(0.014)	(0.000)	(0.069)	(0.140)	(0.139)	(0.000)								
3 *LZTS1*	0.951	0.900	0.988	1.000	0.913	0.883	0.876	1.000	1.000	1.000	1.000	1.000	0.738	0.833	0.881	1.000
SORBS2	(0.033)	(0.065)	(0.025)	(0.000)	(0.109)	(0.104)	(0.110)	(0.143)								
4 *ITPR1*	0.917	0.870	0.934	1.000	0.861	0.783	0.775	1.000	0.937	1.000	0.937	1.000	0.714	0.762	0.810	1.000
IQSEC1	(0.036)	(0.070)	(0.064)	(0.000)	(0.120)	(0.163)	(0.161)	(0.187)								
5 *ATR*	0.985	0.968	0.992	1.000	0.883	0.923	0.838	0.800	0.812	0.875	0.937	1.000	0.809	0.7857	0.833	1.000
PRKCB	(0.025)	(0.029)	(0.042)	(0.000)	(0.096)	(0.083)	(0.139)	(0.175)								

Machine learning methods are indicated, and the values are presented as the mean (standard deviation) calculated from 100 reiterations. AD, Alzheimer's disease; LDA, linear discriminant analysis; KNN, k-nearest neighbors; SVM, support vector machine; RF, random forest

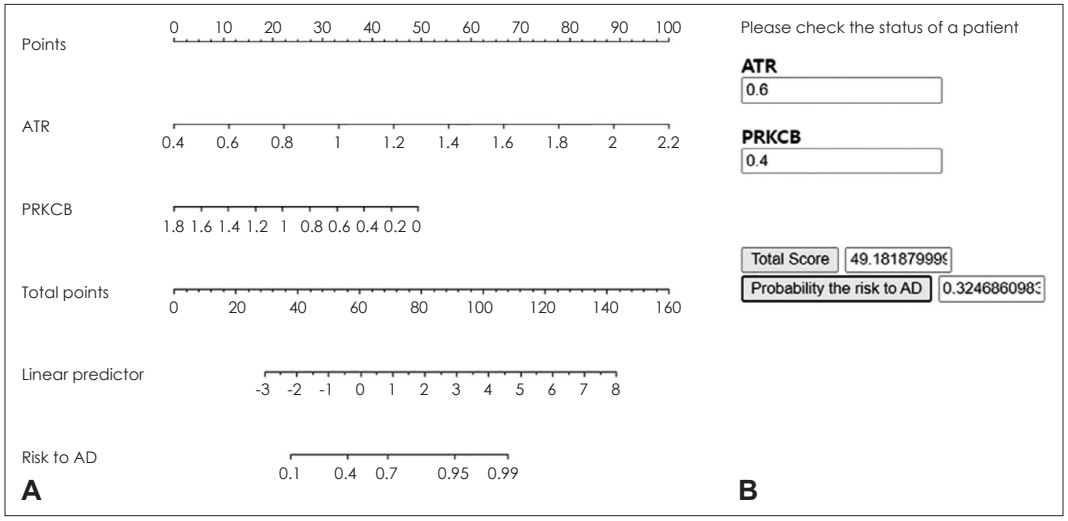

Fig. 4. Nomogram for predicting the probability of AD risk. A: The probability of AD risk for each patient could be identified using the nomogram. B: Practical usage of the nomogram is available in a Hypertext Markup Language format. AD, Alzheimer's disease.

constructed nomogram could be used as a test version. A predictive model using clinical variables and specific gene expression in large datasets of aging populations is needed in the future.

Acknowledgments

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A1A01048175).

Conflicts of interest

The author has no financial conflicts of interest.

ORCID iD

Ki-Yeol Kim https://orcid.org/0000-0001-5357-1067

REFERENCES

1) **Puthiyedth N, Riveros C, Berretta R, Moscato P.** Identification of differentially expressed genes through integrated study of Alzheimer's disease affected brain regions. PLoS One 2016;11:e0152342.
2) **Bradbury R, Brodney MA.** Alzheimer's disease. Topics in medicinal chemistry. Heidelberg: Springer Berlin Heidelberg;2007.
3) **Zhou S, Ma G, Luo H, Shan S, Xiong J, Cheng G.** Identification of 5 potential predictive biomarkers for Alzheimer's disease by integrating the unified test for molecular signatures and weighted gene coexpression network analysis. J Gerontol A Biol Sci Med Sci 2023;

78:653-658.
4) **Alzheimer's Association.** 2022 Alzheimer's disease facts and figures. Alzheimers Dement 2022;18:700-789.
5) **Baumgart M, Snyder HM, Carrillo MC, Fazio S, Kim H, Johns H.** Summary of the evidence on modifiable risk factors for cognitive decline and dementia: a population-based perspective. Alzheimers Dement 2015;11:718-726.
6) **Golde TE.** Alzheimer's disease - the journey of a healthy brain into organ failure. Mol Neurodegener 2022;17:18.
7) **Tanzi RE.** The genetics of Alzheimer disease. Cold Spring Harb Perspect Med 2012;2:a006296.
8) **Roberts R, Knopman DS.** Classification and epidemiology of MCI. Clin Geriatr Med 2013;29:753-772.
9) **Pendlebury ST, Rothwell PM.** Prevalence, incidence, and factors associated with pre-stroke and post-stroke dementia: a systematic review and meta-analysis. Lancet Neurol 2009;8:1006-1018.
10) **Serrano-Pozo A, Frosch MP, Masliah E, Hyman BT.** Neuropathological alterations in Alzheimer disease. Cold Spring Harb Perspect Med 2011;1:a006189.
11) **Lv M, Cui C, Chen P, Li Z.** Identification of osteoporosis markers through bioinformatic functional analysis of serum proteome. Medicine (Baltimore) 2020;99:e22172.
12) **Ralston SH.** Genetics of osteoporosis. Proc Nutr Soc 2007;66:158-165.
13) **Kim SK, Yoo TK, Oh E, Kim DW.** Osteoporosis risk prediction using machine learning and conventional methods. Annu Int Conf IEEE Eng Med Biol Soc 2013;2013:188-191.
14) **Shim JG, Kim DW, Ryu KH, Cho EA, Ahn JH, Kim JI, et al.** Application of machine learning approaches for osteoporosis risk prediction in postmenopausal women. Arch Osteoporos 2020;15:169.
15) **Zheng Z, Zhang X, Oh BK, Kim KY.** Identification of combined biomarkers for predicting the risk of osteoporosis using machine learning. Aging (Albany NY) 2022;14:4270-4280.
16) **Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al.** Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25-29.
17) **Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, et al.** The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 2007;8:R183.
18) **Zhang X, Jang MI, Zheng Z, Gao A, Lin Z, Kim KY.** Prediction of chemosensitivity in multiple primary cancer patients using machine learning. Anticancer Res 2021;41:2419-2429.
19) **McLachlan GJ.** Discriminant analysis and statistical pattern recognition. Hoboken, NJ: Wiley-Interscience;2004.
20) **Liu YZ, Dvornyk V, Lu Y, Shen H, Lappe JM, Recker RR, et al.** A novel pathophysiological mechanism for osteoporosis suggested by an in vivo gene expression study of circulating monocytes. J Biol Chem 2005;280:29011-29016.
21) **Altman NS.** An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 1992;46:175-185.
22) **Vapnik V.** The nature of statistical learning theory, 2nd ed. New York: Springer;2000.
23) **Breiman L.** Random forests. Mach Learn 2001;45:5-32.
24) **Kawaue T, Shitamukai A, Nagasaka A, Tsunekawa Y, Shinoda T, Saito K, et al.** Lzts1 controls both neuronal delamination and outer radial glial-like cell generation during mammalian cerebral development. Nat Commun 2019;10:2780.
25) **Kropp M, Wilson SI.** The expression profile of the tumor suppressor gene Lzts1 suggests a role in neuronal development. Dev Dyn 2012;241:984-994.
26) **Willison AG, Thomas RH.** The neurodevelopmental spectrum seen with CHD2 variants. Pediatr Investig 2022;6:147-148.
27) **Iwata-Otsubo A, Ritter AL, Weckselbatt B, Ryan NR, Burgess D, Conlin LK, et al.** DOCK3-related neurodevelopmental syndrome: biallelic intragenic deletion of DOCK3 in a boy with developmental delay and hypotonia. Am J Med Genet A 2018;176:241-245.
28) **Markunas CA, Quinn KS, Collins AL, Garrett ME, Lachiewicz AM, Sommer JL, et al.** Genetic variants in SLC9A9 are associated with measures of attention-deficit/hyperactivity disorder symptoms in families. Psychiatr Genet 2010;20:73-81.
29) **Kirtay M, Sell J, Marx C, Haselmann H, Ceanga M, Zhou ZW, et al.** ATR regulates neuronal activity by modulating presynaptic firing. Nat Commun 2021;12:4067.
30) **Gerschütz A, Heinsen H, Grünblatt E, Wagner AK, Bartl J, Meissner C, et al.** Neuron-specific alterations in signal transduction pathways associated with Alzheimer's disease. J Alzheimers Dis 2014;40:135-142.
31) **Antonell A, Lladó A, Sánchez-Valle R, Sanfeliu C, Casserras T, Rami L, et al.** Altered blood gene expression of tumor-related genes (PRKCB, BECN1, and CDKN2A) in Alzheimer's disease. Mol Neurobiol 2016;53:5902-5911.
32) **Zhou Z, Chen F, Zhong S, Zhou Y, Zhang R, Kang K, et al.** Molecular identification of protein kinase C beta in Alzheimer's disease. Aging (Albany NY) 2020;12:21798-21808.
33) **Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH.** Predicting sample size required for classification performance. BMC Med Inform Decis Mak 2012;12:8.