science, engineering and health studies

Research Article

Science, Engineering and Health Studies
https://lio1.tci-thaijo.org/index.php/sehs
ISSN (Online): 2630-0087

Forward selection models for classifying mild cognitive impairment and Alzheimer's disease based on single nucleotide polymorphisms

Thitipong Kawichai[1], Pakawat Ingkanisorn[2], and Phuphiphat Jaikaew[2]*

[1] Department of Mathematics and Computer Science, Academic Division, Chulachomklao Royal Military Academy, Nakhon Nayok 26001, Thailand
[2] Department of Biotechnology, Faculty of Science and Technology, Thammasat University, Pathum Thani 12120, Thailand

***Corresponding author**:
Phuphiphat Jaikaew
pupipat@tu.ac.th

***Received**: 12 December 2023*
***Revised**: 4 April 2024*
***Accepted**: 8 April 2024*
***Published**: 14 December 2024*

***Citation**:
Kawichai, T., Ingkanisorn, P., and Jaikaew, P. (2024). Forward selection models for classifying mild cognitive impairment and Alzheimer's disease based on single nucleotide polymorphisms. Science, Engineering and Health Studies, 18, 24050018.

ABSTRACT

Early detection of Alzheimer's disease (AD) is crucial for patients to begin treatment early to slow the disease's progression. While mild cognitive impairment (MCI) is considered an early translational stage of AD, clinically diagnosing MCI is difficult due to its inconsistent symptoms and the lack of standardized diagnostic tests. In this work, we proposed forward selection models to classify patients with AD, patients with MCI and healthy controls (HCs) based on single nucleotide polymorphisms (SNPs). In the proposed method, the initial SNP data were prescreened via genome-wide association studies with a suggestive significance threshold. Then, the qualified SNPs were reselected using the forward SNP selection algorithm to create classification models. Consequently, the forward selection models significantly outperformed the preselection models, those based on all prescreened SNPs, with an area under the precision-recall curve (AUPRC) value of 0.93 in the AD-HC classification, an AUPRC value of 0.94 in the MCI-HC classification, and an AUPRC value of 0.81 in the AD-MCI classification. Moreover, the proposed method could identify AD-associated and MCI-associated SNPs, which would support the clinical diagnosis of AD and MCI in the future.

Keywords: Alzheimer's disease; classification; forward selection; mild cognitive impairment; single nucleotide polymorphisms

1. INTRODUCTION

In the era of globally aging societies, one serious health issue affecting many older adults is dementia, a neurological condition that can affect their daily lives, including impaired thinking, memory, and decision-making. Alzheimer's disease (AD) is the most common type of dementia found in 60–80% of all patients with dementia (Baek et al., 2022). It has been estimated that there are currently about 50 million patients with AD worldwide, and this number is expected to double every five years (Breijyeh and Karaman, 2020). Currently, there is no treatment to stop or reverse the development of AD. Therefore, early and accurate diagnosis of AD is critical, enabling affected individuals to start treatment to slow the disease's progression (Rasmussen and Langerman, 2019).

Mild cognitive impairment (MCI), a condition of cognitive decline, is considered a translational and early detectable stage of AD (Schmidt-Morgenroth et al., 2023). However, clinically diagnosing MCI is sometimes

challenging due to its inconsistent symptoms and the lack of standardized diagnostic tests (Chen et al., 2021a). Numerous studies have identified many genetic markers, mainly single nucleotide polymorphisms (SNPs), associated with AD and MCI to support their early detection (Kim et al., 2014). SNPs are common single-base-pair variations found throughout the human genome. A standard and often used method to identify AD- and MCI-associated SNPs is a genome-wide association study (GWAS), which statistically compares the genomes of many individuals. Typically, a GWAS requires a large sample size to provide the analysis with an adequate statistical power and uses a very stringent *p*-value threshold to reduce false positives (Chen et al., 2021b). These requirements sometimes limit the findings of AD- and MCI-associated SNPs and the accurate identification of AD and MCI based on SNP data.

Numerous machine learning (ML) algorithms have recently been introduced to support the early and accurate detection of AD and MCI. Ahmed et al. (2020) used ML models to identify AD based on AD-associated SNPs. In 2022, a new three-step method based on convolutional neural networks (CNNs) was proposed to identify AD-associated SNPs and perform AD classification based on them (Jo et al., 2022). The most recent method uses CNNs with deep transfer learning and support vector machine (SVM) models to classify AD based on GWAS data (Alatrany et al., 2023). However, most existing methods have been developed to classify AD but not MCI.

Forward selection (FS) is an ML technique that iteratively adds one feature at a time to increase model performance. Due to its easy applicability and efficiency, this technique has been widely used in various fields (Jeong et al., 2022; Salcedo-Sanz et al., 2018; Tangmanussukum et al., 2022). In GWAS, factors such as a stringent threshold and the small number of available samples can limit the identification of disease-associated SNPs. Therefore, using the FS technique to improve the identification of AD- and MCI-associated SNPs based on GWAS is promising.

In this work, we proposed an ML-based method for classifying AD and MCI based on SNP data collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) repository. An illustrated overview of this work is shown in Figure 1. The initial SNP data of patients with AD, patients with MCI, and healthy controls (HCs) were combined to form three different data sets: patients with AD and HCs (AD-HC), patients with MCI and HCs (MCI-HC), and patients with AD and MCI (AD-MCI). Each SNP data set was prescreened via a GWAS with a reduced significance threshold to increase the number of identified SNPs. Then, a forward SNP selection algorithm was used to reselect the prescreened SNPs to optimize AD, MCI, and HC classification. In this step, three binary classification models were constructed: AD-HC, MCI-HC, and AD-MCI. The SNP sets used for the classification could be considered potential SNPs associated with AD and MCI. Then, the SNPs chosen by the forward SNP selection algorithm were validated by searching for supporting evidence in the GWAS Catalog (Sollis et al., 2023).

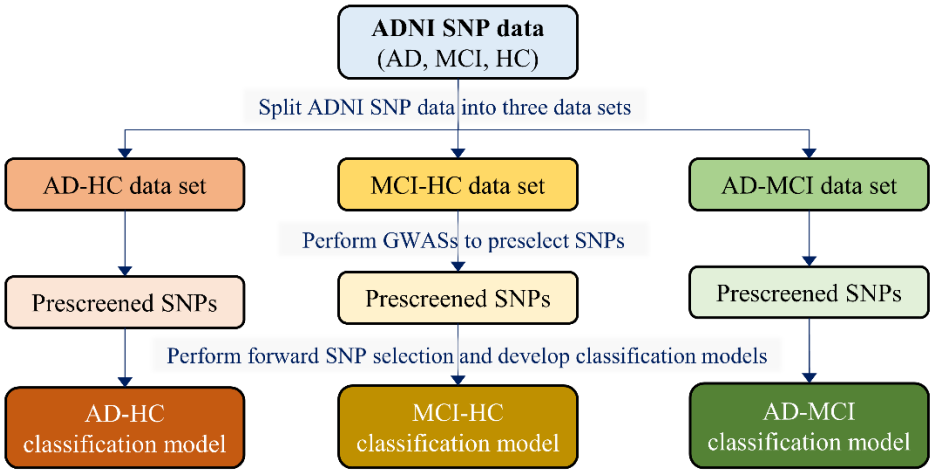

Figure 1. Framework used to construct SNP-based classification models for AD and MCI prediction
Note: The main procedures include (1) splitting the ADNI data into three data sets (AD-HC, MCI-HC, and AD-MCI), (2) prescreening SNPs via GWAS, and (3) forward selecting the prescreened SNPs and creating classification models.

2. MATERIALS AND METHODS

2.1 Data sets

The SNP data of 808 individuals were downloaded from the ADNI repository (https://adni.loni.usc.edu), comprising 232 patients with AD, 329 patients with MCI, and 247 HCs. These data comprised SNPs on the 22 autosomal chromosomes and the X chromosome. The ADNI SNP data were organized into three data sets: patients with AD and HCs (AD-HC), patients with MCI and HCs (MCI-HC), and patients with AD and MCI (AD-MCI).

2.2 Preselection of SNPs and data encoding

In this step, the downloaded SNPs in each data set were first preliminarily screened via GWAS. Then, the data of prescreened SNPs were encoded and used to develop FS models. Before conducting the GWAS, all SNP data underwent quality control using PLINK (version 1.07) (Purcell et al., 2007), considering a SNP call rate of 90%, sample call rate of 95%, minor allele frequency of 1%, and Hardy-Weinberg *p*-value of 0.00001. The GWAS used Fisher's exact tests to identify SNPs potentially associated with AD and MCI in the AD-HC, MCI-HC, and AD-MCI data

sets. A relaxed *p*-value threshold of 1×10^{-5}, known as the suggestive significance threshold (Hammond et al., 2021), was used instead of the commonly used conservative *p*-value threshold of 5×10^{-8} to increase the number of identified SNPs.

After obtaining the prescreened SNPs in the AD-HC, MCI-HC, and AD-MCI data sets, their genotype data were encoded into binary values for each individual (Figure 2). The alleles observed at each SNP (e.g., A and a) were used to generate possible genotypes, which served as separate features (i.e., AA, Aa, and aa). At each SNP, the feature representing an individual's genotype was set to one, with the other features set to zero. The values of the features of all prescreened SNPs were concatenated, and the entire row formed the binary SNP features of an individual. In the AD-HC classification, the patients with AD were labeled as one, and the HCs were labeled as zero. In the MCI-HC classification, the patients with MCI were labeled as one, and the HCs were labeled as zero. In the AD-MCI classification, the patients with AD were labeled as one, and the patients with MCI were labeled as zero.

Sample	SNP1	SNP2	Class
1	AA	Aa	AD
2	Aa	aa	HC
3	aa	AA	AD

Genotype data encoding

	SNP1			SNP2			
Sample	AA	Aa	aa	AA	Aa	aa	Class
1	1	0	0	0	1	0	1
2	0	1	0	0	0	1	0
3	0	0	1	1	0	0	1

Figure 2. Example of genotype data encoding
Note: The upper table contains the genotype data of two SNPs (SNP1 and SNP2) and the class labels of three individuals. The lower table shows the genotype data encoding of those three samples, with their genotype at each SNP set to one and the other genotypes set to zero.

2.3 Development of FS models

Initially, six ML models, including naïve Bayes (NB) (Devroye et al., 1996), k-nearest neighbors (KNN) (Cover and Hart, 1967), support vector machine (SVM) (Cortes and Vapnik, 1995), multi-layer perceptron (MLP) (Haykin, 2009), random forest (RF) (Ho, 1995), and extreme gradient boosting (XGB) (Chen and Guestrin, 2016), were compared in the AD-HC, MCI-HC, and AD-MCI classifications based on the encoded data of all prescreened SNPs. The best-performing ML algorithm was then chosen to develop the FS models. The FS models were independently constructed for each classification case (AD-HC, MCI-HC, and AD-MCI). For the MLP, the simple architecture with a single hidden layer was used due to the limited existing training data. The pseudocode of the forward SNP selection algorithm is shown in Table 1, where $|\cdot|$ represents the number of elements in a set or list, and G_S is a genotype matrix of binary values corresponding to the list of prescreened SNPs $S = \{s_1, s_2, \ldots, s_k\}$, where k is the total number of SNPs in S. We implemented the proposed algorithm using the Python programming language (version 3.9.7) and related packages, including numpy (version 1.24.3), scikit-learn (version 1.2.1), and xgboost (version 1.6.2).

Table 1. The forward SNP selection algorithm

Algorithm: Forward SNP selection

Input: List of all prescreened SNPs $S = \{s_1, s_2, \ldots, s_k\}$ and genotype matrix G_S corresponding to the SNPs in S

Output: List of the forward selected SNPs F

1 initialize $F = \varnothing$ and *isImproved* = Yes
2 **while** $|F| < k$ and *isImproved* == Yes **do**
3 **for each** $s \in S$ **do**
4 evaluate the performance p_i of a model trained on genotype matrix $G_{F \cup \{s\}}$, where $i = 1, 2, \ldots, |S|$
5 **end for**
6 $imax = \underset{i}{\operatorname{argmax}}\, p_i$
7 **if** $p_{imax} > p_{latest}$ **then**
8 Set $F = F \cup \{s_{imax}\}$, $S = S - \{s_{imax}\}$, and $p_{latest} = p_{imax}$
9 *isImproved* = Yes
10 **else**
11 *isImproved* = No
12 **end if**
13 **end while**
14 **return** F

The forward SNP selection algorithm aims to identify an approximately optimal subset (F) of a given SNP set (S) via FS to create the best-performing classification model. As shown in Table 1, the list of forward-selected SNPs (F) is initialized as an empty set. In order to choose which SNP will be moved into F, the genotype data of SNPs from S is experimentally used to train a model with the genotype data of SNPs in F, and the classification performance (p_i) of the trained model is evaluated, where $i = 1, 2, \ldots, |S|$. A SNP that gives the best classification performance (p_{imax}) will be moved into F if the performance of the model trained on $G_{F \cup \{s_{imax}\}}$ is better than that of the latest iteration (p_{latest}). The algorithm stops the iteration when F contains all prescreened SNPs or the performance of the models trained on $G_{F \cup \{s\}}$, for every s in S, is not greater than that of the latest model.

2.4 Evaluation of model performance

We evaluated the performance of a classification model using ten-fold nested cross-validation. This validation technique preserved 10% of the data for testing model performance in each iteration, and this process was repeated ten times with different testing folds. Only the training data set (90% of the data) was used to tune the hyperparameters of ML models. A grid search with five-fold cross-validation was performed on a training data set to specify the generalized values of the hyperparameters. The considered values of the hyperparameters are listed in Table 2.

Table 2. The considered hyperparameter values of the ML models

Model	Hyperparameter	Considered values
NB	-	-
KNN	*n_neighbors*	[5, 9, 13, 17, 21]
RF	*n_estimators*	[128, 256, 512, 1024]
MLP	*hidden_layer_sizes*	[20, 40, 60, 100]
SVM	*C*	[0.01, 0.1, 1, 10]
XGB	*n_estimators*	[128, 256, 512, 1024]

This work used standard evaluation metrics, including precision (PRE), recall (REC), accuracy (ACC), Matthew's correlation coefficients (MCC), and F1-score (F_1). The formulas of the performance metrics are shown in Equations (1) – (4), where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively. We also calculated the area under a receiver operating characteristic curve (AUROC) and the area under a precision-recall curve (AUPRC) to comprehensively evaluate model performance. The values of all evaluation metrics, except MCC, were between 0 and 1. The higher the value of an evaluation metric, the better the model performance.

$$\text{PRE} = \frac{TP}{TP+FP}, \text{REC} = \frac{TP}{TP+FN} \quad (1)$$

$$\text{ACC} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (3)$$

$$F_1 = \frac{2 \times \text{PRE} \times \text{REC}}{\text{PRE}+\text{REC}} \quad (4)$$

3. RESULTS AND DISCUSSION

3.1 Preselection of SNPs

The Manhattan plots for the AD-HC, MCI-HC, and AD-MCI GWAS to preselect SNPs are shown in Figure 3, where the top horizontal line in red represents the gold standard threshold (p-value = 5×10^{-8}), and the bottom horizontal line in blue represents the suggestive threshold (p-value = 1×10^{-5}). With the gold standard threshold, 25 SNPs on chromosome 19 were identified in the AD-HC case. One SNP on chromosome 11 was identified in the MCI-HC case, and five SNPs on chromosome 19 were identified in the AD-MCI case. To increase the number of prescreened SNPs, those below the suggestive threshold (p-value < 1×10^{-5}) were selected and used to develop the classification models: 138 for the AD-HC case, 109 for the MCI-HC case, and 78 for the AD-MCI case.

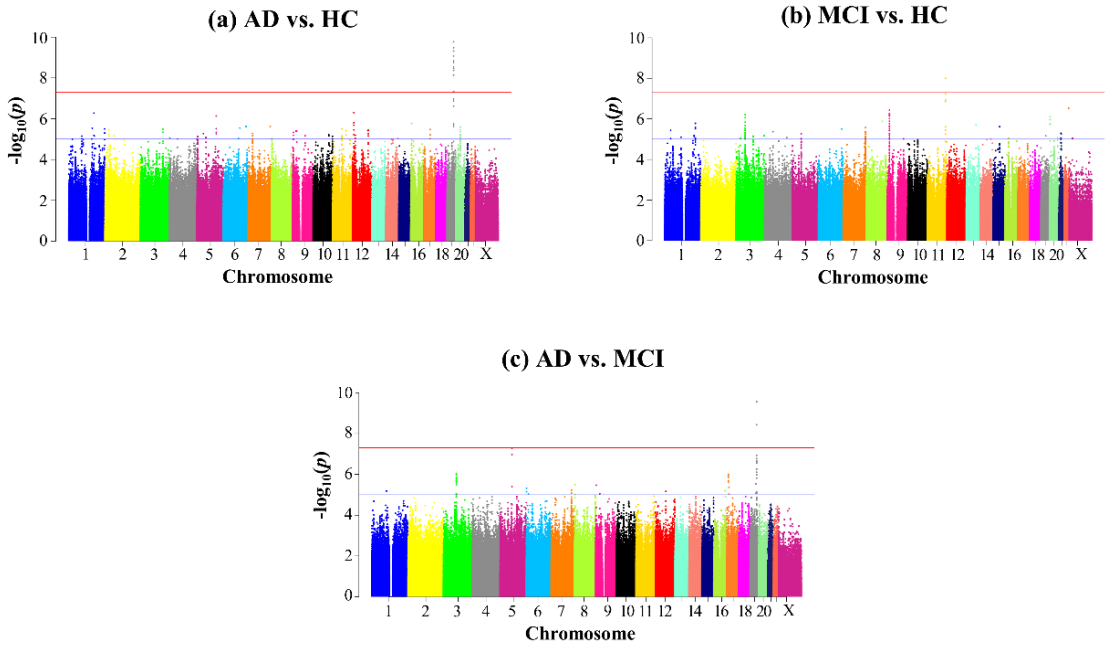

Figure 3. The Manhattan plots for the GWAS of (a) AD-HC, (b) MCI-HC, and (c) AD-MCI cases

3.2 Performance comparison of the ML models

The mean AUROC values of six ML models were compared in Table 3 to identify the best one to be applied in the forward SNP selection. For each classification case, models were created based on the genotype data of all preselected SNPs. We compared the mean AUROC values of two models using paired t-tests, particularly in cases of two approximately equal values, and then specified ranks. An average rank for all classification cases was computed for each model. The lower the average rank, the better the model performs over all classification cases.

According to Table 3, the SVM achieved the lowest average rank with mean AUROC values of 0.915, 0.892, and 0.820 in the AD-HC, MCI-HC, and AD-MCI classification cases, respectively. While the MLP achieved the highest AUROC values for the AD-HC and MCI-HC classification cases, it did not perform well in the AD-MCI classification case. Therefore, we used the SVM for forward SNP selection to classify AD, MCI, and HC samples.

Table 3. Performance comparison of the six ML models

Model	AD-HC		MCI-HC		AD-MCI		Average rank
	AUROC	Rank	AUROC	Rank	AUROC	Rank	
SVM	0.915	1	0.892	2	**0.820**	1	1.33
MLP	**0.916**	1	**0.896**	1	0.784	4	2.00
XGB	0.899	3	0.881	3	0.804	2	2.67
RF	0.869	4	0.848	4	0.792	3	3.67
KNN	0.821	6	0.784	5	0.755	5	5.33
NB	0.852	5	0.784	5	0.748	6	5.33

Note: The largest AUROC value for each case is shown in bold.

3.3 Improved performance of FS models

We presumed that the genotype data of some preselected SNPs may not be useful for AD-HC, MCI-HC, and AD-MCI classification. Therefore, we used an SVM model for forward SNP selection to identify sets of important SNPs for each classification case. Consequently, 50 of the 138 SNPs, 36 of the 109 SNPs, and 34 of the 78 SNPs were reselected and used in the AD-HC, MCI-HC, and AD-MCI classifications, respectively. Based on ten replicates, the average performance values of the SVM models with the preselection (PS) and FS methods are shown in Table 4.

In the AD-HC classification case, the FS method had significantly higher AUPRC (p-value = 0.041) and AUROC (p-value = 0.039) values than the PS method although their other performance values did not differ significantly. In the MCI-HC classification case, the FS method had significantly higher values than the PS method for all performance metrics, except REC (p-value = 0.694). Similarly, in the AD-MCI classification case, the FS method had significantly higher values than the PS method for all performance metrics except REC.

Table 4 shows that forward SNP selection significantly improved the performance of models trained on the preselected SNPs, implying that the preselected SNPs included some that did not contribute substantially to classifications, resulting in lower performance. Additionally, the forward SNP selection algorithm could efficiently select SNPs that were important for classification.

Table 4. Performance of the models with the PS and FS methods

Performance metrics	AD-HC			MCI-HC			AD-MCI		
	PS	FS	p-value	PS	FS	p-value	PS	FS	p-value
AUPRC	0.91	**0.93**	0.041*	0.92	**0.94**	6.3×10^{-4}**	0.75	**0.81**	4.1×10^{-5}**
AUROC	0.91	**0.93**	0.039*	0.89	**0.91**	2.4×10^{-3}**	0.82	**0.85**	1.5×10^{-5}**
PRE	0.88	0.88	0.925	0.84	**0.89**	0.021*	0.67	**0.74**	0.037*
REC	**0.88**	0.87	0.871	0.89	0.88	0.694	**0.88**	0.85	0.392
ACC	0.88	0.88	0.499	0.84	**0.87**	2.1×10^{-3}**	0.77	**0.81**	0.033*
MCC	0.76	0.76	0.419	0.67	**0.74**	2.0×10^{-3}**	0.57	**0.63**	0.014*
F_1	0.87	0.87	0.444	0.86	**0.88**	3.0×10^{-3}**	0.76	**0.78**	0.015*

Note: The highest performance values between the FS and PS methods for each classification case are shown in bold. Key: *, $p < 0.05$; **, $p < 0.01$.

3.4 Comparison of different SNP selection methods

Using forward SNP selection, we could identify the sets of SNPs important for the AD-HC, MCI-HC, and AD-MCI classification models. For each classification case, the number of SNPs obtained from three selection methods, including GWAS with the gold standard threshold (p-value < 5×10^{-8}), GWAS with the suggestive threshold (p-value < 1×10^{-5}), and forward SNP selection, is shown in Table 5. The complete lists of the SNPs are provided in the Supplementary Data File at https://github.com/thitipongk/ForwardSnpSelection.

Few SNPs were identified with the gold standard threshold (Table 5). Notably, in the MCI-HC classification case, only one SNP (rs2919475) was detected using GWAS with the stringent gold standard threshold. More SNPs were identified with the relaxed suggestive threshold: 138 for the AD-HC case, 109 for the MCI-HC case, and 78 for the AD-MCI case. However, many false positive SNPs could be selected for the classification models using this relaxed p-value threshold of 1×10^{-5}. By leveraging the classification models, the forward SNP selection algorithm prescreened SNPs obtained from GWAS with the suggestive threshold to identify those contributing substantially to the classifications. Consequently, 50 of the 138 SNPs, 36 of the 109 SNPs, and 34 of the 78 SNPs were identified as important for AD-HC, MCI-HC, and AD-MCI classification, respectively. Therefore, the proposed method could identify more important SNPs than the standard GWAS, potentially including novel SNPs associated with MCI and AD.

Interestingly, our analysis revealed that some SNPs identified by GWAS with the gold standard threshold overlapped those selected through FS. However, the final FS results excluded a subset of the gold standard SNPs. In the AD-HC classification case, 11 gold standard SNPs were not chosen using the proposed FS method (see the Supplementary Data File). The p-values of all excluded SNPs were above 10^{-11}, and most ranked at the bottom with p-values near the gold standard threshold. In the MCI-HC classification case, the gold standard SNP was also selected using the FS method. In the AD-MCI classification case, only one gold standard SNP, which had the highest p-value among the gold standard SNPs, was not selected by the proposed FS method. The forward SNP selection method identified SNPs based on their contributions to classification models. Therefore, it could be suggested that those removed SNPs might not contribute sufficiently to the

classification and could be suspected of being unassociated with AD and MCI. In addition, genotype data encoding could substantially increase the input data size, especially in cases initiated with many prescreened SNPs, because too many features can negatively affect the learning efficiency of ML models. Therefore, additional feature selection or extraction is required to enhance classification performance when many prescreened SNPs are obtained.

Table 5. The number of SNPs obtained with the different selection methods

Selection methods	Number of SNPs		
	AD-HC	MCI-HC	AD-MCI
GWAS with the gold standard threshold ($p < 5 \times 10^{-8}$)	25	1	5
GWAS with the suggestive threshold ($p < 1 \times 10^{-5}$)	138	109	78
Forward SNP selection	50	36	34

3.5 Identification and verification of AD- and MCI-associated SNPs

Based on forward SNP selection, 50 of the 138 SNPs were reselected for the AD-HC classification. These SNPs were considered genetic markers potentially associated with AD. Similarly, 34 of the 78 SNPs were reselected for the AD-MCI classification. 10 of which overlapped those reselected for the AD-HC classification (Figure 4). These SNPs are all in or near the genes apolipoprotein E (*APOE*) and apolipoprotein C1 (*APOC1*). Many studies have confirmed that polymorphisms in both *APOE* and *APOC1* underlie AD progression (Yamazaki et al., 2019; Zhou et al., 2014). According to the GWAS, the *p*-values of all overlapping SNPs were below the gold standard threshold of 5×10^{-8}. Moreover, five of the 10 SNPs (rs429358, rs769449, rs483082, rs56131196, and rs4420638) were already reported as AD biomarkers in the GWAS Catalog.

Among the other AD-associated SNPs identified by the AD-HC classification model, four (rs6857, rs283815, rs59007384, and rs75627662) were also reported in the GWAS Catalog. Interestingly, rs283815 and rs75627662 were not identified by regular GWAS with the gold standard threshold but were discovered by the proposed FS method. Additionally, two SNPs that were not reported as AD biomarkers in the GWAS Catalog were identified by both GWAS and the proposed FS method (rs438811 and rs5117).

In the AD-MCI classification case, three interesting SNPs were identified by the proposed FS method: rs283811, rs115881343, and rs66626994. While the GWAS method with the gold standard threshold did not discover rs283811, which was already reported as AD-associated SNP in the GWAS Catalog, the proposed FS method did discover this SNP, suggesting that it can discover rare SNPs associated with a trait that are missed by regular GWAS. Both rs115881343 and rs66626994 were not included among the SNPs associated with AD or MCI in the GWAS catalog. Nonetheless, rs115881343 was associated with age-inducing cognitive decline in the GWAS catalog, which may progress into MCI or AD. Moreover, rs66626994 was associated with high-density lipoprotein (HDL) cholesterol levels in the GWAS catalog. Since low HDL levels have been identified as an important risk factor for MCI (Cho et al., 2019), it can be suggested that rs66626994 is a potential MCI-associated SNP.

The MCI-associated SNPs identified by our FS method are summarized in Figure 5. Notably, while rs28669215 has not been identified by stringent GWAS, it was discovered by our FS method and is associated with impulsivity in the GWAS catalog. It has been recently revealed that impulsive behaviors are associated with less favorable prognoses in patients with MCI (Bidzan et al., 2023). Moreover, three SNPs were only discovered by the proposed FS method (rs4323397, rs10224365, and rs10238169), which were all located in introns of the gene encoding dipeptidyl-peptidase 6 (*DPP6*). Despite the lack of supporting reports in the GWAS Catalog, *DPP6* has been identified as a novel gene associated with dementia (Cacace et al., 2019). Lastly, rs2919475 was identified by both the FS method and regular GWAS. Despite the lack of supporting evidence, the relationship of rs2919475 with MCI warrants further investigation.

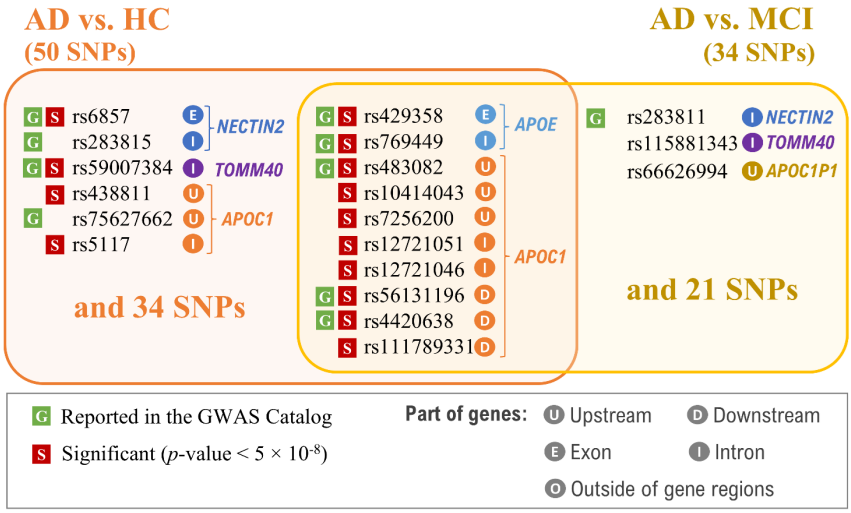

Figure 4. The SNPs identified in the AD-HC and AD-MCI classification cases

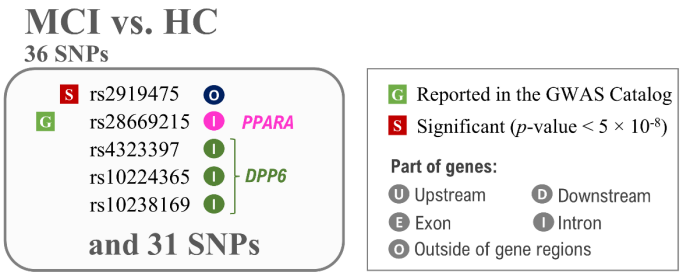

Figure 5. The SNPs identified in the MCI-HC classification case

4. CONCLUSION

This work proposed ML models with forward SNP selection to detect AD and MCI based on SNP data. They significantly enhanced classification performance compared to models based only on SNPs identified via GWAS. Moreover, forward SNP selection could efficiently screen and identify SNPs important for classification, which could be considered AD- and MCI-associated SNPs. Future studies should include other types of genetic variants (e.g., copy number variations) and data (e.g., electronic health records) in the classification models to enhance classification performance.

ACKNOWLEDGMENT

Data collection and sharing for this work was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD (Department of Defense) ADNI under Grant W81XWH-12-2-0012. ADNI was funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai, Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research and Development, LLC.; Johnson and Johnson Pharmaceutical Research and Development LLC.; Lumosity; Lundbeck; Merck and Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer, Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California.

REFERENCES

Ahmed, H., Soliman, H., and Elmogy, M. (2020). Early detection of Alzheimer's disease based on single nucleotide polymorphisms (SNPs) analysis and machine learning techniques. In *Proceedings of the 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, pp. 1–6. Sakheer, Bahrain.

Alatrany, A. S., Khan, W., Hussain, A. J., Mustafina, J., and Al-Jumeily, D. (2023). Transfer learning for classification of Alzheimer's disease based on genome wide data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(5), 2700–2711.

Baek, M. S., Kim, H.-K., Han, K., Kwon, H.-S., Na, H. K., Lyoo, C. H., and Cho, H. (2022). Annual trends in the incidence and prevalence of Alzheimer's disease in South Korea: A nationwide cohort study. *Frontiers in Neurology*, 13, 883549.

Bidzan, L., Grabowski, J., Przybylak, M., and Ali, S. (2023). Aggressive behavior and prognosis in patients with mild cognitive impairment. *Dementia and Neuropsychologia*, 17, e20200096.

Breijyeh, Z., and Karaman, R. (2020). Comprehensive review on Alzheimer's disease: Causes and treatment. *Molecules*, 25(24), e20200096.

Cacace, R., Heeman, B., Van Mossevelde, S., De Roeck, A., Hoogmartens, J., De Rijk, P., Gossye, H., De Vos, K., De Coster, W., Strazisar, M., De Baets, G., Schymkowitz, J., Rousseau, F., Geerts, N., De Pooter, T., Peeters, K., Sieben, A., Martin, J. J., Engelborghs, S., Salmon, E., Santens, P., Vandenberghe, R., Cras, P., De Deyn, P. P., Swieten, J. C., Duijn, C. M., Zee, J., Sleegers, K., and Van Broeckhoven, C. (2019). Loss of *DPP6* in neurodegenerative dementia: A genetic player in the dysfunction of neuronal excitability. *Acta Neuropathologica*, 137, 901–918.

Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. California, USA.

Chen, Y.-X., Liang, N., Li, X.-L., Yang, S.-H., Wang, Y.-P., and Shi, N.-N. (2021a). Diagnosis and treatment for mild cognitive impairment: A systematic review of clinical practice guidelines and consensus statements. *Frontiers in Neurology*, 12, 719849.

Chen, Z., Boehnke, M., Wen, X., and Mukherjee, B. (2021b). Revisiting the genome-wide significance threshold for common variant GWAS. *G3 Genes Genomes Genetics*, 11(2), jkaa056.

Cho, K. H., Park, H. J., Kim, S. J., and Kim, J. R. (2019). Decrease in HDL-C is associated with age and

household income in adults from the Korean National Health and Nutrition examination survey 2017: Correlation analysis of low HDL-C and poverty. *International Journal of Environmental Research and Public Health*, 16(18), 3329.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.

Devroye, L., Györfi, L., and Lugosi, G. (1996). Introduction. In *A Probabilistic Theory of Pattern Recognition. Stochastic Modelling and Applied Probability, vol 31* (Devroye, L., Györfi, L., and Lugosi, G., Eds.), pp. 1–8. New York: Springer.

Hammond, R. K., Pahl, M. C., Su, C., Cousminer, D. L., Leonard, M. E., Lu, S., Doege, C. A., Wagley, Y., Hodge, K. M., Lasconi, C., Johnson, M. E., Pippin, J. A., Hankenson, K. D., Leibel, R. L., Chesi, A., Wells, A. D., and Grant, S. F. (2021). Biological constraints on GWAS SNPs at suggestive significance thresholds reveal additional BMI loci. *eLife*, 10, e62206.

Haykin, S. S. (2009). *Neural Networks and Learning Machines*, 3rd, New Jersey: Pearson Prentice Hall, pp. 122–200.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pp. 278–282. Quebec, Canada.

Jeong, D., Yoo, C., Yeh, S.-W., Yoon, J.-H., Lee, D., Lee, J.-B., and Choi, J.-Y. (2022). Statistical seasonal forecasting of winter and spring PM2.5 concentrations over the Korean peninsula. *Asia-Pacific Journal of Atmospheric Sciences*, 58, 549–561.

Jo, T., Nho, K., Bice, P., and Saykin, A. J. (2022). Deep learning-based identification of genetic variants: Application to Alzheimer's disease classification. *Briefings in Bioinformatics*, 23(2), bbac022.

Kim, D. H., Yeo, S. H., Park, J.-M., Choi, J. Y., Lee, T.-H., Park, S. Y., Ock, M. S., Eo, J., Kim, H.-S., and Cha, H.-J. (2014). Genetic markers for diagnosis and pathogenesis of Alzheimer's disease. *Gene*, 545(2), 185–193.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3), 559–575.

Rasmussen, J., and Langerman, H. (2019). Alzheimer's disease - Why we need early diagnosis. *Degenerative Neurological and Neuromuscular Disease*, 9, 123–130.

Salcedo-Sanz, S., Cornejo-Bueno, L., Prieto, L., Paredes, D., and García-Herrera, R. (2018). Feature selection in machine learning prediction systems for renewable energy applications. *Renewable and Sustainable Energy Reviews*, 90, 728–741.

Schmidt-Morgenroth, I., Michaud, P., Gasparini, F., and Avrameas, A. (2023). Central and peripheral inflammation in mild cognitive impairment in the context of Alzheimer's disease. *International Journal of Molecular Sciences*, 24(13), 10523.

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Gunes, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J. A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., Ramachandran, S., Stefancsik, R., Stewart, J., Whetzel, P., Wilson, R., Hindorff, L., Cunningham, F., Lambert, S. A., Inouye, M., Parkinson, H., and Harris, L. W. (2023). The NHGRI-EBI GWAS Catalog: Knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1), D977–D985.

Tangmanussukum, P., Kawichai, T., Suratanee, A., and Plaimas, K. (2022). Heterogeneous network propagation with forward similarity integration to enhance drug–target association prediction. *PeerJ Computer Science*, 8, e1124.

Yamazaki, Y., Zhao, N., Caulfield, T. R., Liu, C. C., and Bu, G. (2019). Apolipoprotein E and Alzheimer's disease: Pathobiology and targeting strategies. *Nature Reviews Neurology*, 15, 501–518.

Zhou, Q., Zhao, F., Lv, Z. P., Zheng, C. G., Zheng, W. D., Sun, L., Wang, N. N., Pang, S., de Andrade, F. M., Fu, M., He, X. H., Hui, J., Jiang, W., Yang, C. Y., Shi, X. H., Zhu, X. Q., Pang, G. F., Yang, Y. G., Xie, H. Q., Zhang, W. D., Hu, C. Y., and Yang, Z. (2014). Association between *APOC1* polymorphism and Alzheimer's disease: A case-control study and meta-analysis. *PLoS One*, 9(1), e87017.