Expert Systems With Applications 140 (2020) 112873

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset

Chihyun Park, Jihwan Ha, Sanghyun Park*

Dept. of Computer Science, Yonsei University, 134, Shinchon-dong, Seodaemun-gu, Seoul, South Korea

ARTICLE INFO

Article history:
Received 8 February 2019
Revised 14 August 2019
Accepted 14 August 2019
Available online 15 August 2019

Keywords:
Alzheimer's disease
Omics data integration
Biological feature selection
Deep neural network
Machine learning

ABSTRACT

Motivation: The molecular mechanism of Alzheimer's disease (AD) has not been clearly revealed and there is no clinically reliable genetic risk factor. Therefore, diagnosis of AD has been mostly performed by analyzing brain images such as magnetic resonance imaging and neuropsychological tests. Identifying the molecular-level mechanism of AD has been lacking data owing to the difficulty of sampling in the posterior brains of normal and AD patients; however, recent studies have produced and analyzed large-scale omics data for brain areas such as prefrontal cortex. Therefore, it is necessary to develop AD diagnosis or prediction methods based on these data.

Results: This paper proposed a deep learning-based model that can predict AD using large-scale gene expression and DNA methylation data. The most challenging problem in constructing a model to diagnose AD based on the multi-omics dataset is how to integrate different omics data and how to deal with high-dimensional and low-sample-size data. To solve this problem, we proposed a novel but simple approach to reduce the number of features based on a differentially expressed gene and a differentially methylated position in the multi-omics dataset. Moreover, we developed a deep neural network-based prediction model that improves performance compared to that of conventional machine learning algorithms. The feature selection method and the prediction model presented in this paper outperformed conventional machine learning algorithms, which utilize typical dimension reduction methods. In addition, we demonstrated that integrating gene expression and DNA methylation data could improve the prediction accuracy.

Availability: https://github.com/ChihyunPark/DNN_for_ADprediction.

1. Introduction

Advances in producing high-throughput omics have enabled us to understand disease mechanisms and pathways at a detailed molecular level and contributed toward development of advanced treatments based on this understanding (Karczewski & Snyder, 2018). In addition, these omics data have led to the development of various predictive models for identifying disease risk or prognosis of cancer patients (Choi et al., 2017; Karczewski & Snyder, 2018). With the development of machine learning methods, omics data have been used in various research fields, including complex diseases such as cancer. Recently, there have been many attempts to integrate and utilize different types of omics data. This means simultaneously using both gene expression and copy number data as an input for training the model. The integrated use of multi-omics data provides more opportunities to overcome limitations regarding the etiology of a disease, and leads to the development of more accurate models that reflect the nature of biology (Huang, Chaudhary, & Garmire, 2017). However, it is still challenging to integrate the dataset and information of different biological layers because the biological process of different layers is rather interdependent or interactive. Nevertheless, in the case of cancer, many studies have been performed owing to the emergence of datasets such as TCGA (Cancer Genome Atlas Research et al., 2013). However, there have been relatively fewer attempts for other diseases. In other words, fewer studies have proposed a model for predicting disease risk using omics data for other diseases.

Late-onset Alzheimer's disease (AD) is a disease that requires such study. Globally, the number of AD patients is expected to increase considerably, posing a huge threat to public health (Brookmeyer et al., 2007). Several studies have been conducted to identify genetic risk factors that can elucidate the complexity of AD pathogenesis (Park et al., 2017). However, the causal mechanisms of AD occurrence and progression have not been precisely revealed

* Corresponding author.
E-mail addresses: chihyun.park@yonsei.ac.kr (C. Park), jihwanha@yonsei.ac.kr (J. Ha), sanghyun@yonsei.ac.kr (S. Park).

https://doi.org/10.1016/j.eswa.2019.112873
0957-4174/

and the only treatments available are those to relieve symptoms (Zhang et al., 2013). In terms of AD prediction or diagnosis studies using the molecular layer dataset, only few studies have been conducted because it is difficult to obtain an omics dataset from AD patients' brain tissue. In these circumstances, a method for obtaining risk scores using AD-associated single nucleotide polymorphism (SNP) derived from genome-wide association study (GWAS) and APOE status has recently been published (Desikan et al., 2017). APOE is a representative gene known to be associated with AD. It has been revealed that the risk of AD can be increased or decreased according to the genetic variants of APOE (Liu et al., 2013). Desikan et al. (2017) used the Cox proportional-hazards model and large-scaled genomic dataset produced from the International Genomics of Alzheimer's Project to quantify genetic risks for AD. They demonstrated that this model can quantify individual differences in age-specific genetic risk for AD. However, they only focused only on APOE and AD-associated SNPs identified by GWAS. As AD has a complex genetic mechanism, there are various AD-related genetic factors that cannot be explained by SNP alone and need to be included in the prediction model.

AD prediction studies that do not use genetic data mostly use phenotypic data, such as magnetic resonance imaging, and machine learning methods are appropriately utilized for these datasets (Lee et al., 2018). From a clinical aspect, these phenotypic data and neuropsychological tests have been utilized together to diagnose AD; a recent study focused on AD diagnosis with biomarkers in living persons (Jack et al., 2018). According to the studies conducted so far, these biomarkers are molecules belonging to the pathway known to be involved in AD, such as β amyloid deposition, pathologic tau, and neurodegeneration. In addition, the proton-magnetic resonance spectroscopy technique, which provides a noninvasive imaging biomarker, has been utilized to classify AD (Munteanu et al., 2015). In this study, a multilayer perceptron-based model showed the best performance. However, these studies did not consider the interactions among different molecular layers such as the association between gene expression and DNA methylation. In summary, currently, few studies use computa models based on multi-omics data to diagnose or predict AD.

This paper proposes a deep learning-based model to predict AD that uses multiple heterogeneous omics datasets. We previously conducted a study to construct an AD-specific gene network by integrating omics data from different layers and confirmed that it is possible to predict AD from omics data (Park et al., 2017). In particular, we integrated gene expression and DNA methylation and found that it can help to explain the AD mechanism. However, building the AD prediction model with two different omics datasets was challenging. As the data had high-dimensional and low-sample-size (HDLSS) characteristics, we had to apply a method to reduce the feature, and it was unfeasible to use a common feature selection algorithm because the two omics data had different biological characteristics. Here, we propose a feature selection method that is simple but can retain biological characteristics well. Based on this, we propose a deep neural network-based model that can derive the best prediction performance compared to typical machine learning algorithms.

2. Materials and methods

In this section, we introduce the proposed algorithm with an explanation of the dataset being used, how we integrate two heterogeneous omics datasets, and why we apply the deep learning approach. As shown in Fig. 1, our approach generally comprises of three parts: feature selection, model training, and model testing.

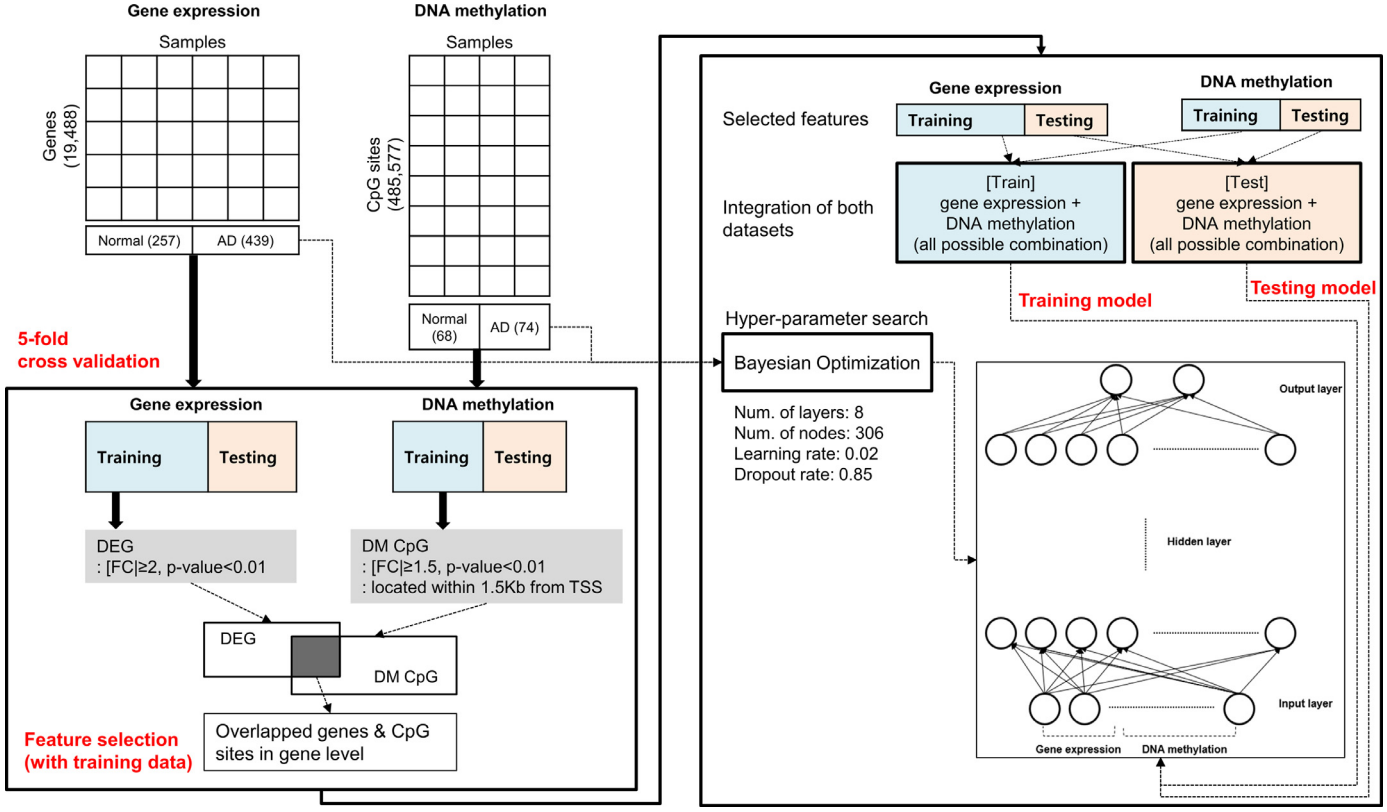

Fig. 1. Workflow of the proposed method. The proposed method comprises three parts. After splitting data into training and testing sets, the feature space was reduced by the proposed approach, and then, the AD prediction model was trained with the integrated omics dataset. The hyperparameters of the proposed deep neural network model were determined by Bayesian optimization. Finally, we validated the performance of the constructed model with testing dataset.

2.1. Datasets

In this study, we used two types of heterogeneous omics datasets: gene expression and DNA methylation profiles. We integrated two large-scale gene expression profiles GSE33000 and GSE44770 (Narayanan et al., 2014; Zhang et al., 2013) to increase the sample size, focusing exclusively on the prefrontal cortex. The integrated dataset was composed of 257 non-demented, i.e. normal, and 439 AD samples. These two gene expressions were normalized by the z-score. In addition, we used recently published DNA methylation profiles corresponding to the same brain region, i.e., prefrontal cortex GSE80970 (Smith et al., 2018). This dataset uses the Illumina HumanMethylation 450 Beadchip and comprises 68 normal and 74 AD samples. From raw *Beta*-values, we calculated the *M*-value. As suggested in the previous study (Du et al., 2010), *M*-values were used for differential analysis of methylation levels to further increase the statistical validity. In addition, quantile normalization was applied for the following analysis. Table 1 summarizes the datasets used in the paper.

2.2. The proposed feature selection approach

We used a multi-omics dataset for building the AD prediction model. It is difficult to simply combine gene expression and DNA methylation profiles to form a feature set, because they have different characteristics. Not only their biological meaning but also the manner in which the values are derived differ, and their distribution might also differ. Therefore, we define two types of features. The first is gene and the second is probing of CpG site from gene expression and DNA methylation.

As shown in Table 1, the numbers of both feature types are considerably high to be used for building the model. In particular, they have HDLSS characteristic with a much higher number of features than the number of samples. HDLSS characteristic is a well-known problem for phenotype prediction using genetic data. If the dataset that has HDLSS characteristic is directly used to build a prediction model, severe overfitting and high variance of gradients can occur. This is challenging for most machine learning algorithms. Therefore, an appropriate feature selection algorithm is required to reduce the size of the feature space and risk of overfitting.

To train the model, we propose differentially expressed gene (DEG)- and differentially methylated position (DMP)-based approaches. The reason we do not use conventional feature selection or dimension reduction algorithms, such as Lasso, Relief-F, or principal component analysis (PCA), is that they cannot reflect biological processes. This means that these methods can reduce the number of features or dimensions, but do not guarantee that the reduced features will retain their biological meaning. Moreover, these methods are not suitable to be applied to the multi-omics dataset. These methods cannot consider the relationship between two different omics datasets. Especially, DNA methylation is well known as a molecular factor that can control and regulate the level of expression of genes located nearby in the CpG site. Conventional feature selection or dimension reduction algorithms cannot handle this characteristic.

Our feature selection method comprises two steps. The first step is to identify DEG and DMP. We use the Limma package, which provides solution for analysis of both DEG and DMP, and filter the results using our criterion |fold change| ≥ 2, P-value < 0.01 (Ritchie et al., 2015). Limma package is the most widely used package to confirm the degree of DEG based on a t-test in microarray-based experiments, and has recently been actively used for DNA methylation analysis (Maksimovic, Phipson, & Oshlack, 2016). The same criterion |fold change| ≥ 1.5, P-value < 0.01 is used to identify DMPs; then, we only select DMPs whose CpG sites are located inside of the 1500 base pairs from the transcript start site. The reason why we choose these DMPs is that methylation at the transcription factor binding site has high potential to regulate gene expression. The second step is to integrate DEGs and DMPs by intersecting both. The reason we apply intersection is that DNA methylation in promoters is closely linked to downstream gene repression. The level of gene expression can be regulated if the promoter site is methylated. Therefore, we hypothesize that if a gene is differentially expressed and its promoter site is hyper- and hypomethylated simultaneously, the gene is significantly associated with the disease. The regulatory relationship, which means that the hyper/hypomethylated gene in its promoter or gene body region will affect down/up-regulation of its expression, depends on the disease and not all genes show this pattern. Therefore, we do not account for the details of DMP and DMG, such as hypermethylation or down-regulation, to include various situations without considering the ideal situation alone. We consider the gene as a feature if its level of methylation and expression is significantly different compared to the normal status. All DMPs have their corresponding genes, and we can intersect these two results at the gene level.

A characteristic of the used data is that the gene expression and DNA methylation profiles are not measured from the same sample. As described above, however, both gene expression and DNA methylation data are generated from the tissues extracted from the prefrontal region of late-onset AD patients. Therefore, we decide to use all possible pairs of gene expression and DNA methylation profile for each label, i.e., normal and AD. Ideally, to be more specific, the number of gene expressions and DNA methylation profiles for a normal sample are 257 and 68, respectively. From these samples, we obtained all possible combinations of 17,476 samples and regard them as input data with normal labels. Similarly, in the case of AD, the gene expression has 439 samples and the DNA methylation has 74 samples. All possible combinations of 32,486 samples are obtained and regarded as input data labeled with AD. Because we split the training and testing data sets through k-fold cross-validation, the number of actual AD and normal samples by the combination approach are fewer than 17,476 and 32,486, respectively. Several studies have performed integrated analysis with gene expression and DNA methylation dataset from different samples but the same disease. However, they also focused on the overlapping genes between DEG and DMG (Song et al., 2018; Zhao et al., 2018).

2.3. Deep neural network for prediction model

This paper proposes a deep neural network (DNN) model. To investigate the optimized hyperparameters of the proposed model, Bayesian optimization (Snoek, Larochelle, & Adams, 2012) is uti-

Table 1
Summary of the used datasets. Two microarray-based gene expression datasets and an array-based DN methylation dataset were used.

Dataset	Gene expression Rosetta/Merck Human 44k 1.1 microarray		DNA methylation Illumina Human Methylation 450 Bead Chip
GEO ID	GSE33000	GSE44770	GSE80970
Number of normal samples	157	100	68
Number of AD samples	310	129	74
Number of features	19,488 genes		485,577 probes for CpG site

lized because it efficiently uncovers the global maxima of the black-box function, such as the accuracy value for a validation set in the defined parameter space. The Bayesian approach keeps track of previous evaluation results and infers the probabilistic model, and then, selects the next candidate of parameters based on this model. Therefore, Bayesian optimization can efficiently search the optimal hyperparameters. Our objective function value is accuracy of test dataset. We search for a combination of parameters with the best test accuracy in the specified bound region for the following four parameters: number of the hidden layers (7–11), number of the nodes per layer (250–350), learning rate (0.01–0.2) and dropout rate (0.6–0.9). From the entire gene expression and DNA methylation dataset, the proposed feature selection approach and Bayesian optimization are consequently performed with 5-fold split data. After the learning for each fold until 400 epochs, a combination of parameters with the highest test accuracy is identified. The average of the hyperparameters from the five learnings is finally applied to the proposed model.

The basic structure of the proposed DNN model is shown in Fig. 1. The input layer comprises two parts, which correspond to the gene expression and the DNA methylation dataset. There are two nodes in the output layer as our problem is binary classification and one hot encoding is used for the output variable. ReLU is used as an activation function and a softmax regression layer with logit scores is added to the output layer to convert and normalize the output value to be between 0 and 1. The model is composed of 8 hidden layers with 306 nodes and one bias node for each one. We use the reduced mean of cross-entropy as the cost function. Then, we perform gradient descent optimization to minimize the cost. The cost function uses the following formula:

$$cost\left(y,\hat{y}\right) = -\sum_{k=1}^{2} y_k \log\left(\hat{y}_k\right) \tag{1}$$

$$\hat{y}_k = \frac{exp(W_k x + b_k)}{\sum_{j=1}^{2} exp\left(W_j + b_j\right)} \tag{2}$$

where y and $\hat{y}$ denote the known and predicted values, respectively. y is one hot encoded and $\hat{y}$ is a result of softmax regression. W, x, and b denote weight, input, and bias, respectively, and exp denotes an exponential function. The learning and dropout rates of the proposed model are set to 0.02 and 0.85, respectively. The maximum number of epochs is 1500.

To avoid overfitting, we apply not only cross-validation but also early stopping based on the test dataset during the model training. We define a simple rule to stop training: after 100 epochs, the average of the ten most recent test accuracy values is calculated for every epoch, and this value is compared with the current test accuracy value to check whether it is converged or decreasing. Simultaneously, in the same way, the current training accuracy is compared with the average value of the training of the recent ten epochs to check whether it is increasing. If both of these rules are satisfied, the learning is terminated. The used thresholds were 0.001 and 0.01 for the convergence or decrease and the increase, respectively. The proposed DNN model is implemented with the API of Google TensorFlow (version 1.4.1).

3. Results

3.1. Experimental design

Our approach has three components—dataset, feature selection or dimension reduction, and training algorithm. Through the experiments, we tried to verify the following three hypotheses assumed in this paper.

(1) The use of multi-omics data, rather than single-omics data, will lead to an improvement in the accuracy of AD prediction.
(2) The proposed feature selection method used for biological analysis will lead to a higher AD prediction than a general dimension reduction method.
(3) Applying deep learning over conventional machine learning methods may improve predictive performance.

To verify the above assumptions and to demonstrate the superiority of the proposed method, we performed various comparative experiments while changing these components. For comparison, we defined a baseline method that uses each gene expression and the DNA methylation dataset, as well as a typical dimension reduction algorithm and a conventional training algorithm based on machine learning. Based on this method, we performed a comparative experiment by changing the proposed method for each element. In this section, we present the result of the baseline method. Both dimension reduction algorithms and conventional machine learning algorithms were used, which were provided by the Scikit-learn package (version 0.21.2).

3.2. Single-omics datasets

First, we applied PCA and t-Stochastic Nearest Neighbor (t-SNE), which are typically used to reduce the features. Using these algorithms, we reduced the dimensions to the same number of features as those obtained through DEG or DMP. As, for each fold, the numbers of DEG and DMP identified from the training data were different, the size of the reduced dimension was also different, as shown in Table 2. The dimensions were reduced to 17.8 for DEG and 156.4 for DMP, on average. In this experiment, each gene expression and DNA methylation dataset were used as a training dataset. Therefore, PCA and t-SNE were performed with each gene expression and DNA methylation dataset. There were the following four training cases: PCA-gene expression, PCA-DNA methylation, t-SNE-gene expression, and t-SNE-DNA methylation. As a prediction algorithm, we used random forest (Breiman, 2001), SVM (Keerthi et al., 2001), and naïve Bayesian (John & Langley, 1995) with the following random forest: criterion = 'entropy', max depth = 6, number of trees = 100; SVM: kernel = RBF (Radial Basis Function, complexity = 1, gamma = auto; naïve Bayesian: no parameter) throughout the remaining comparison tests in this paper. As shown in Table 2, there were 12 different baseline prediction models. The highest average accuracy was 0.632 when t-SNE was applied to gene expression and the prediction model was constructed with SVM. However, the performance of all these baseline prediction models was insufficient to be used as the AD prediction model.

3.3. Integrating gene expression and DNA methylation datasets

Instead of using each gene expression and DNA methylation dataset, we integrated them as input of the prediction model. As shown in Fig. 1, we first conducted the proposed feature selection approach with the training dataset after splitting the input data into training and testing datasets. Then, we integrated both omics data into one for each training and testing datasets. While integrating, we obtained all possible combinations of samples for each label as described in the Method section. Consequently, on average, 11,178 normal samples and 20,376 AD samples were generated by this approach during five folds as a training dataset and average 693 normal samples and 1192 AD samples were generated as a testing dataset. With this dataset, PCA and t-SNE were performed. We reduced the dimension using PCA and t-SNE with the same number of dimensions as that of the features obtained by applying DEG and DMG. The average value of the reduced dimension was 174.2. After performing dimension reduction with PCA and t-SNE,

Table 2
Performance comparison for three AD prediction algorithms varying the dimension reduction algorithm for each gene expression and DNA methylation dataset.

5-fold CV	Training algorithm	Num. of. genes (Num. of samples)	Accuracy (AUROC)		Num. of CpGs (Num. of. samples)	Accuracy (AUROC)	
			PCA (gene expression)	*t*-SNE (gene expression)		PCA (DNA methylation)	*t*-SNE (DNA methylation)
$K=1$	Random Forest	29 (556/140)	0.579 (0.500)	0.593 (0.506)	266 (113/29)	0.483 (0.481)	0.517 (0.519)
	SVM		0.600 (0.500)	0.579 (0.485)		0.517 (0.500)	0.483 (0.500)
	Naïve Bayesian		0.629 (0.613)	0.414 (0.476)		0.552 (0.550)	0.586 (0.598)
$K=2$	Random Forest	24 (557/139)	0.698 (0.565)	0.604 (0.485)	212 (113/29)	0.379 (0.394)	0.429 (0.427)
	SVM		0.669 (0.500)	0.633 (0.473)		0.448 (0.500)	0.571 (0.500)
	Naïve Bayesian		0.583 (0.556)	0.568 (0.500)		0.448 (0.442)	0.500 (0.448)
$K=3$	Random Forest	2 (557/139)	0.547 (0.466)	0.604 (0.489)	11 (114/28)	0.357 (0.358)	0.536 (0.537)
	SVM		0.554 (0.485)	0.640 (0.500)		0.500 (0.428)	0.429 (0.497)
	Naïve Bayesian		0.640 (0.500)	0.640 (0.500)		0.571 (0.615)	0.607 (0.564)
$K=4$	Random Forest	2 (557/139)	0.619 (0.576)	0.633 (0.593)	17 (114/28)	0.607 (0.604)	0.429 (0.427)
	SVM		0.554 (0.521)	0.633 (0.595)		0.571 (0.500)	0.571 (0.500)
	Naïve Bayesian		0.568 (0.500)	0.568 (0.500)		0.500 (0.521)	0.500 (0.448)
$K=5$	Random Forest	32 (557/139)	0.579 (0.546)	0.604 (0.487)	276 (114/28)	0.500 (0.513)	0.536 (0.556)
	SVM		0.676 (0.554)	0.676 (0.500)		0.464 (0.500)	0.464 (0.500)
	Naïve Bayesian		0.612 (0.568)	0.640 (0.514)		0.500 (0.503)	0.536 (0.556)
Avg.	Random Forest	**17.8**	**0.624 (0.531)**	**0.608 (0.512)**	**156.4**	**0.465 (0.470)**	**0.472 (0.481)**
	SVM	**(556.8/139.2)**	**0.611 (0.501)**	**0.632 (0.511)**	**(113.6/28.4)**	**0.500 (0.486)**	**0.479 (0.499)**
	Naïve Bayesian		**0.606 (0.547)**	**0.584 (0.497)**		**0.514 (0.526)**	**0.535 (0.533)**

AUROC: Area Under the Receiver Operating Characteristic.
Num. of samples: number of training samples/number of testing samples.

Table 3
Performance comparison for three AD prediction algorithms varying the dimension reduction algorithm using the integrated dataset of gene expression and DNA methylation.

5-fold CV	Training algorithm	Num. of genes and CpGs (Num. of samples)	Accuracy (AUROC)	
			PCA (gene expression + DNA methylation)	*t*-SNE (gene expression + DNA methylation)
$K=1$	Random Forest	295 (31,386/1946)	0.526 (0.508)	0.487 (0.465)
	SVM		0.597 (0.500)	0.597 (0.500)
	Naïve Bayesian		0.599 (0.593)	0.557 (0.616)
$K=2$	Random Forest	236 (31,672/1840)	0.659 (0.636)	0.49 (0.476)
	SVM		0.600 (0.500)	0.600 (0.500)
	Naïve Bayesian		0.598 (0.597)	0.600 (0.500)
$K=3$	Random Forest	13 (31,343/1958)	0.590 (0.542)	0.694 (0.657)
	SVM		0.281 (0.500)	0.281 (0.500)
	Naïve Bayesian		0.590 (0.682)	0.717 (0.577)
$K=4$	Random Forest	19 (31,495/1890)	0.586 (0.565)	0.568 (0.475)
	SVM		0.381 (0.500)	0.578 (0.467)
	Naïve Bayesian		0.599 (0.606)	0.604 (0.504)
$K=5$	Random Forest	308 (31,876/1791)	0.614 (0.583)	0.439 (0.414)
	SVM		0.623 (0.500)	0.623 (0.500)
	Naïve Bayesian		0.565 (0.566)	0.636 (0.547)
Avg.	Random Forest	**174.2**	**0.595 (0.567)**	**0.536 (0.497)**
	SVM	**(31,554.4/1885.0)**	**0.496 (0.500)**	**0.536 (0.493)**
	Naïve Bayesian		**0.590 (0.609)**	**0.623 (0.549)**

three machine learning algorithms were used to build a prediction model.

As shown in Table 3, the accuracy and area under the receiver operating characteristics (AUROC) were measured. The performance in the case of using the integrated omics data was improved over the case of using the DNA methylation data alone, but it was confirmed that, when we used random forest and SVM as the prediction models, the overall accuracy was lower than that obtained using the gene expression data alone.

As shown in Table 4, the performance of the proposed feature selection approach was better than that of DNA methylation alone, but somewhat lower than that of gene expression alone. However, as shown in Tables 2 and 3, we confirmed that our feature selection approach performed better than the case where we used PCA or *t*-SNE in each dataset, including gene expression, DNA methylation and integration of both.

When using only gene expression or DNA methylation, there are limitations in training and testing the model with only a few samples. We supposed that the integrating approach is meaningful because it may be possible to explain the biological mechanism with relatively similar performance. If we use gene expression and DNA methylation together, the causal relationship can be inferred.

3.4. Deep learning-based prediction model

Table 5 shows the result of hyperparameter search. To investigate the best combination of parameters for the proposed deep neural network model, we performed Bayesian optimization as we mentioned in the Materials and method section. Before conducting our feature selection and training the model, we split our input data using the five-fold cross-validation approach. Then, for each fold, we performed Bayesian optimization and obtained a combination of parameters with the best training accuracy. Finally, the average value of each parameter from the five folds was applied to our model. The final dropout rate, learning rate, size of hidden layers, and number of nodes per layer were 0.85, 0.02, 8, and 306, respectively.

Table 4
Performance comparison for three AD prediction algorithms with our feature selection approach. To show the contribution of data integration, we compared the performance of the prediction.

5-fold CV	Training algorithm	Num. of genes and CpGs (Num. of samples)	Accuracy (AUROC) Our approach (DEG + DMP)	Num. of genes (Num. of samples)	Accuracy (AUROC) DEG	Num. of CpGs (Num. of samples)	Accuracy (AUROC) DMP
$K=1$	Random Forest	295 (31,386/1946)	0.659 (0.652)	29 (556/140)	0.829 (0.821)	266 (113/29)	0.448 (0.445)
	SVM		0.393 (0.401)		0.829 (0.818)		0.483 (0.479)
	Naïve Bayesian		0.485 (0.464)		0.771 (0.777)		0.483 (0.479)
$K=2$	Random Forest	236 (31,672/1840)	0.745 (0.739)	24 (557/139)	0.849 (0.821)	212 (113/29)	0.414 (0.433)
	SVM		0.609 (0.598)		0.827 (0.789)		0.586 (0.589)
	Naïve Bayesian		0.575 (0.573)		0.806 (0.778)		0.586 (0.589)
$K=3$	Random Forest	13 (31,343/1958)	0.610 (0.562)	2 (557/139)	0.799 (0.764)	11 (114/28)	0.500 (0.492)
	SVM		0.480 (0.513)		0.806 (0.769)		0.643 (0.642)
	Naïve Bayesian		0.778 (0.721)		0.806 (0.774)		0.714 (0.684)
$K=4$	Random Forest	19 (31,495/1890)	0.763 (0.750)	2 (557/139)	0.734 (0.714)	17 (114/28)	0.643 (0.635)
	SVM		0.514 (0.523)		0.727 (0.709)		0.750 (0.719)
	Naïve Bayesian		0.727 (0.683)		0.734 (0.714)		0.714 (0.688)
$K=5$	Random Forest	308 (31,876/1791)	0.721 (0.710)	32 (557/139)	0.799 (0.753)	276 (114/28)	0.607 (0.618)
	SVM		0.634 (0.624)		0.820 (0.780)		0.571 (0.574)
	Naïve Bayesian		0.564 (0.558)		0.770 (0.737)		0.571 (0.574)
Avg.	Random Forest	**174.2 (31,554.4/1885.0)**	**0.700 (0.683)**	**17.8 (556.8/139.2)**	**0.802 (0.775)**	**156.4 (113.6/28.4)**	**0.522 (0.525)**
	SVM		**0.526 (0.532)**		**0.802 (0.773)**		**0.607 (0.600)**
	Naïve Bayesian		**0.626 (0.600)**		**0.777 (0.756)**		**0.614 (0.603)**

Table 5
Results of hyperparameters search by Bayesian optimization. For each fold, we performed Bayesian optimization and average value of the parameters was finally applied to our model.

5-fold CV	Dropout rate	Learning rate	Size of hidden layers	Number of nodes per layer	Best test accuracy from Bayesian Optimization
$K=1$	0.897	0.019	7	340	0.999857
$K=2$	0.900	0.010	11	303	0.999571
$K=3$	0.900	0.010	7	340	1.000000
$K=4$	0.871	0.083	9	271	0.999857
$K=5$	0.722	0.017	8	277	0.999285
Average	**0.85**	**0.02**	**8**	**306**	0.999710

Table 6
Performance comparison while varying dimension reduction/feature selection algorithm in deep learning-based prediction model.

5-fold CV	PCA (gene expression + DNA methylation)			*t*-SNE (gene expression + DNA methylation)			Our approach (gene expression + DNA methylation)		
	Training Accuracy	Test cost	Test Accuracy (Test AUROC)	Training Accuracy	Test cost	Test Accuracy (Test AUROC)	Training Accuracy	Test cost	Test Accuracy (Test AUROC)
$K=1$	0.939	1.307	0.681 (0.670)	0.850	1.016	0.576 (0.511)	0.850	0.430	0.823 (0.815)
$K=2$	0.856	1.760	0.576 (0.570)	0.829	0.681	0.657 (0.612)	0.844	0.521	0.804 (0.779)
$K=3$	0.764	0.907	0.688 (0.655)	0.665	0.566	0.750 (0.577)	0.788	0.337	0.872 (0.820)
$K=4$	0.939	1.978	0.507 (0.495)	0.857	1.059	0.637 (0.524)	0.818	0.445	0.826 (0.806)
$K=5$	0.931	1.007	0.676 (0.668)	0.870	1.322	0.452 (0.405)	0.863	0.556	0.788 (0.765)
average	**0.886**	**1.392**	**0.626 (0.612)**	**0.814**	**0.929**	**0.614 (0.526)**	**0.832**	**0.458**	**0.823 (0.797)**

As shown in Tables 3 and 4, when we integrated the multi-omics dataset, the highest accuracy of the conventional machine learning algorithm achieved was 0.7. We assumed that the deep learning model could improve the predictive performance. Table 6 shows the accuracy and AUROC of a deep neural network-based prediction model while varying the feature selection approach. As with the previous experiments, we applied five-fold cross-validation and obtained the average value of accuracy and AUROC.

As described in the Method section above, early stopping was applied when training the model. Fig. 2 shows the relationships among training loss, training accuracy, and testing accuracy obtained by training and testing datasets from the first fold. As shown in Fig. 2, after about 900th epochs, training loss was decreased and training accuracy was increased, but testing accuracy was decreased. That is, overfitting occurred. In our approach, overfitting was detected in the 897th epoch and the learning was stopped.

Consequently, 0.823 was the average accuracy of the proposed deep learning with our feature selection method. In addition, we found that the deep learning-based method performed slightly better when integrating the omics dataset. Figs. 3 and 4 show the entire average accuracy and AUROC values while varying the dataset and feature selection approach. The red bar indicates the performance of the proposed deep learning model. We also constructed deep neural network model using the features only from DEGs because machine learning algorithms in this case showed 0.802 as a test accuracy as shown in Table 4. However, the test accuracy when we applied deep learning approach to this case was 0.737. Supporting Table 1 and 2 and Supporting Fig. 1 include the result of this evaluation. It was demonstrated that the proposed approach, which uses biological feature selection such as DEG and DMP with the deep learning model, was superior to all comparative approaches while changing the dataset.

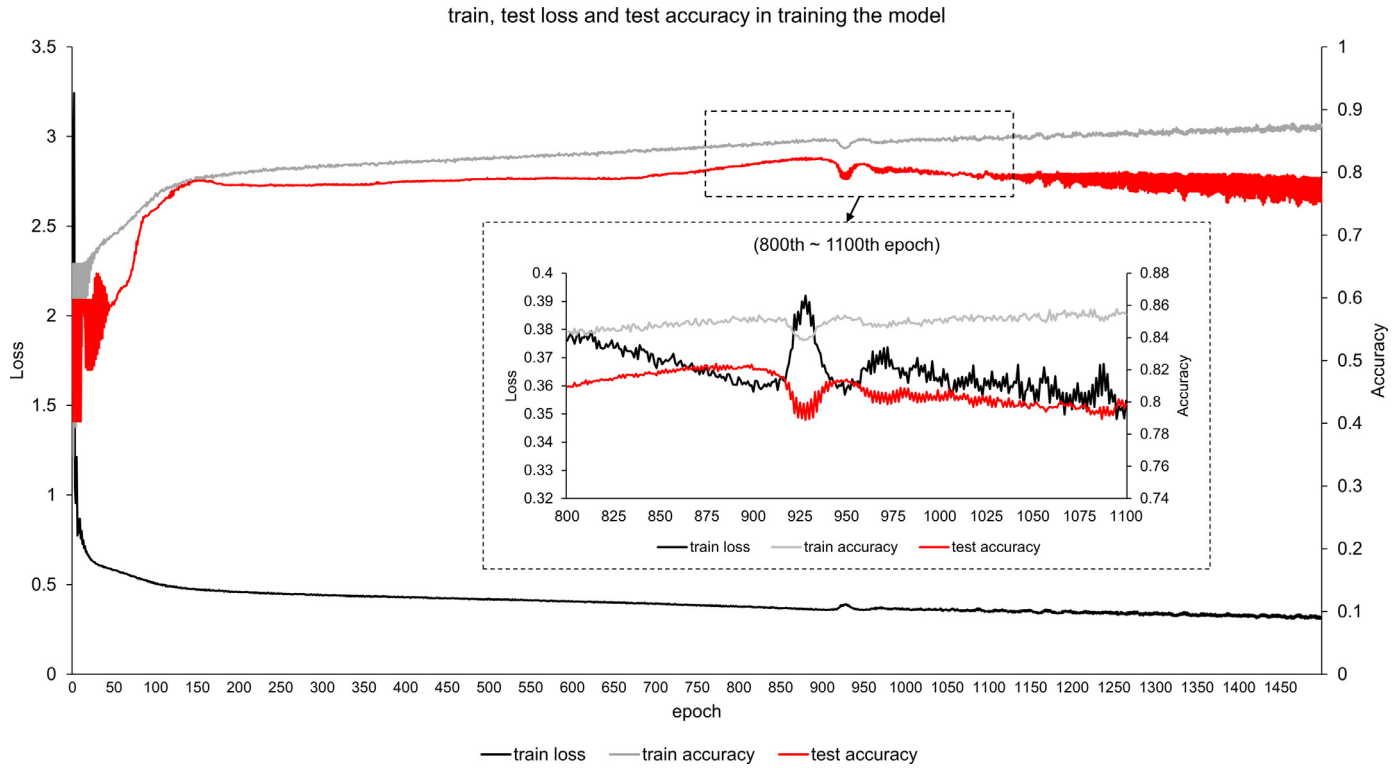

Fig. 2. Importance of early stopping to avoid overfitting when training the model. In this figure, there are training loss, training accuracy, and testing accuracy of the first fold dataset. After approximately 900 epochs, the training was successful, but the testing accuracy did not increase. In the proposed method, the training was stopped at the 897th epoch.

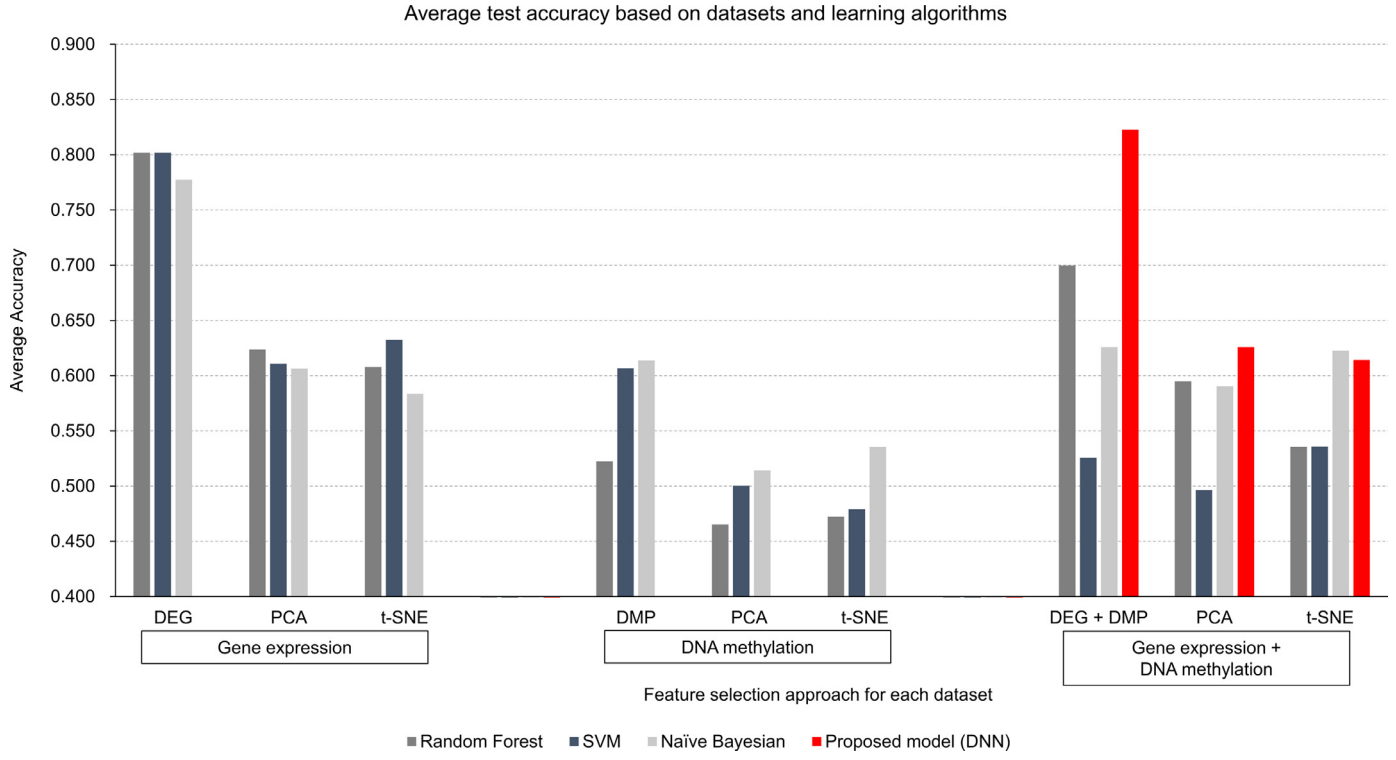

Fig. 3. Average accuracy of all comparisons and the proposed model. X-axis presents the dimension reduction or feature selection approaches for single-omics and multi-omics datasets. Each colored bar indicates the difference prediction algorithm, especially the red bar presents the proposed deep learning model. The proposed approach showed the best accuracy among all comparisons.

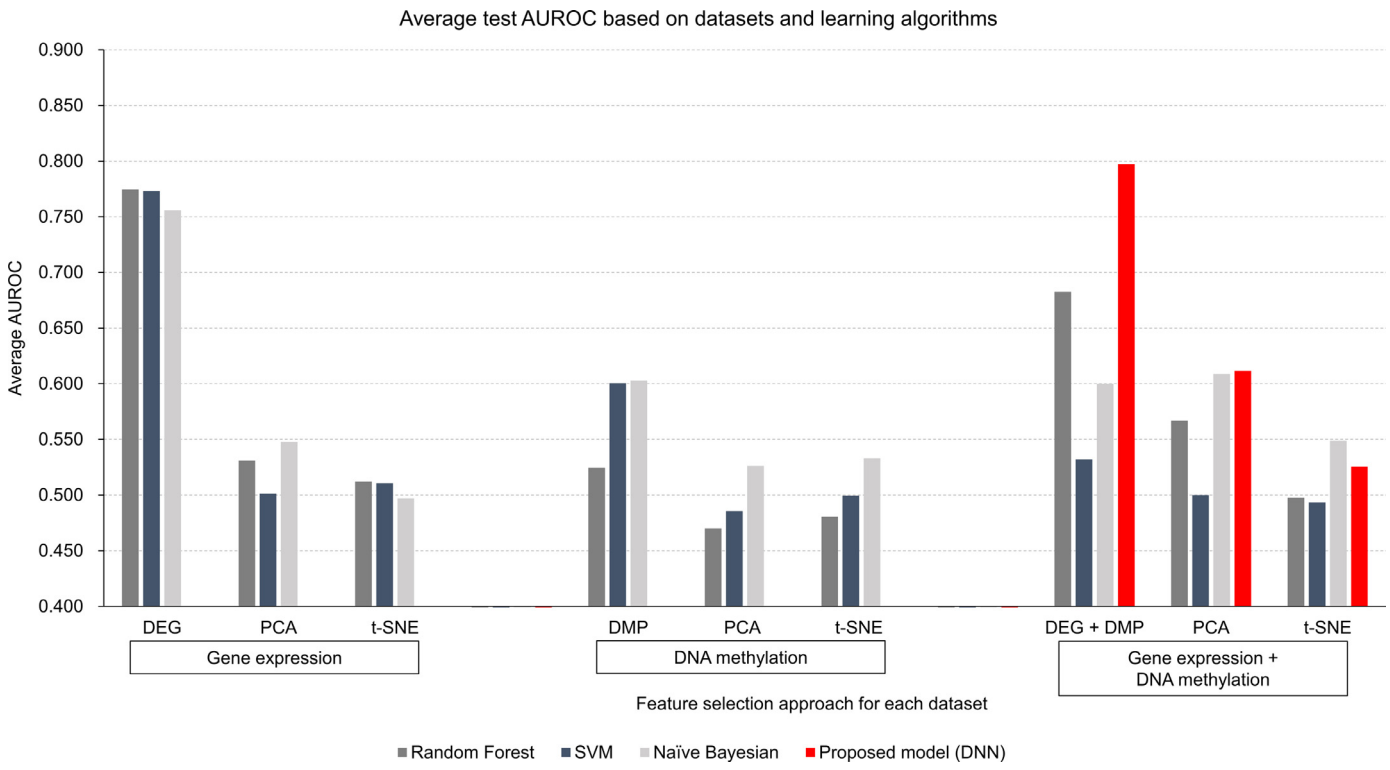

Fig. 4. Average AUROC of all comparisons and the proposed model.

Table 7
Selected genes from gene expression and DNA methylation and comparison with AlzGene database.

Each fold	$K=1$ (29 genes)	$K=2$ (24 genes)	$K=3$ (2 genes)	$K=4$ (2 genes)	$K=5$ (32 genes)
Selected genes	ARMCX5 BEX2 BEX5 CHM CITED1 CNKSR2 DDX3Y ELF4 ELK1 FLNA GPRASP1 MAGED1 MAP7D2 MCTS1 MID1IP1 MORF4L2 NGFRAP1 PAK3 PIGA PJA1 RBMX RPGR RRAGB SAT1 SLC25A5 SLC9A6 SRPX SYTL4 TAZ	ARMCX5 BEX2 CNKSR2 DDX3Y ELF4 ELK1 FLNA GPRASP1 MAGED1 MCTS1 MID1IP1 NGFRAP1 PAK3 PIGA PIR PJA1 RBMX RRAGB SAT1 SLC25A5 SLC9A6 SRPX STAG2 TAZ	BEX2 MS4A4A	BEX2 ELF4	ARMCX5 BEX2 BEX5 CNKSR2 DDX3Y ELF4 ELK1 FLNA GPRASP1 MAGED1 MAP7D2 MCTS1 MID1IP1 MORC4 MORF4L2 NGFRAP1 PAK3 PIGA PIR PJA1 RBMX RNF128 RPGR RRAGB SAT1 SLC25A5 SLC9A6 SRPX STAG2 SYTL4 TAZ TMSL3
Union of the selected genes across 5 folds	ARMCX5 BEX2 BEX5 CHM CITED1 CNKSR2 DDX3Y ELF4 ELK1 FLNA GPRASP1 MAGED1 MAP7D2 MCTS1 MID1IP1 MORC4 MORF4L2 MS4A4A NGFRAP1 PAK3 PIGA PIR PJA1 RBMX RNF128 RPGR RRAGB SAT1 SLC25A5 SLC9A6 SRPX STAG2 SYTL4 TAZ TMSL3				
Reported in AlzGene DB	MS4A4A				

4. Discussion

As mentioned in the Method section, on average, 17.8 genes and 156.4 CpG sites corresponded to 35 genes selected by the proposed approach. We demonstrated that the 35 genes had discriminative ability to classify the AD status compared to the normal status through various experiments. In this part, we investigated these 35 genes. Table 7 shows the list of genes and whether they have been already reported in the AlzGene (Bertram et al., 2007) database, which curated all genes from more than 1390 genetic association studies. In the AlzGene database, 695 genes were reported by GWAS meta-analysis in 2013. Among the 35 genes, only one gene (MS4A4A) was overlapped. Recently, CpG-related SNP, which has the potential to perturb DNA methylation near MS4A4A, has been reported to be significantly associated with AD risk (Ma et al., 2019) by performing a genome-wide study on large-sized cohorts.

One interesting observation from the experimental results is that in the five-fold tests, the third fold showed the highest test accuracy, as shown in Table 6, where the used genes included only MS4A4A and BEX2. We could infer that these two genes had higher predictive power than the other genes. BEX2 is known to play a role in cell cycle progression and inhibit neuronal differentiation. Although the study is still lacking, it can be supposed that this gene is also associated with AD because the biomedical significance of MS4A4A has been revealed. Through this literature study, we conclude that the features we selected can be new AD markers that are different from those previously known to be involved in AD.

To identify the functions of these 35 genes, several enrichment tests were performed using GSEA (Subramanian et al., 2005). First, gene ontology-based enrichment tests were conducted. Table 8 shows their results. As a result, regulation of cell death was significantly enriched from our gene set with a low p-value

Table 8
Result of functional enrichment test with gene ontology and KEGG pathway by using GSEA.

Category	GO term or KEGG pathway	*P*-value
Gene Ontology Biological Process	Regulation of Cell Death	9.75e−8

Table 9
AD-related result from the enrichment test with chemical and genetic perturbations CGP in GSEA.

CGP	Description	*P*-value
Blalock Alzheimers Disease DN	Genes down-regulated in brain from patients with Alzheimer's disease	1.92e−08

(9.75e-8). Neuronal cell death in AD is a frequently observed pathological feature. It has been revealed that Amyloid β (Aβ), the major component of senile plaques, plays a central role in neuronal cell death. There are many studies that have tried to reveal the mechanism of Aβ in AD. In terms of drug development, inhibiting Aβ production and aggregation is one of the representative trials (Hampel et al., 2010). Genes that can control cell death are important in AD and can also be used as diagnostic markers.

We also performed a functional enrichment test with chemical and genetic perturbations, CGP in GSEA. As shown in Table 9, the 35 genes significantly overlapped with groups of AD-related down-regulated genes. All these results of functional enrichment tests imply that the feature we extracted was significantly related to AD-related functions or the pathway.

Multi-omics approaches have been widely applied to solve biomedical problems. Integration of different omics data types can elucidate potential causative changes that lead to diseases or phenotypic changes (Hasin, Seldin, & Lusis, 2017). Regression-based and data mining- or machine learning-based approaches have been widely used, and network approaches that can reflect biological interconnections in complex diseases have recently emerged (Yan et al., 2018). Most studies for integrating transcriptomic and epigenetic data have focused on identifying causal epigenetic factors that can affect changes in gene expression, such as mQTL approaches (Taylor et al., 2019). However, there is still little attempt to integrate multi-omics data into disease prediction problems. Two recent studies have integrated multi-omics data based on deep learning to distinguish subtypes in diseases or predicting cancer patients' survival (Chaudhary et al., 2018; Zhang et al., 2018). In both these studies, multi-omics data were integrated by applying an autoencoder and a prediction model was constructed by applying a conventional machine learning algorithm such as SVM.

The most important advantage of integrating multi-omics into prediction models is that we can identify molecular factors that are considered important for the predictive models. Based on this, mechanism studies can be conducted to explain the changes in phenotypes or diseases. Although there are not many such approaches in disease prediction studies, as well as AD, we suppose that such study will receive much attention owing to the possibility of biomedical explanation and connection to the mechanism study.

5. Conclusions

We proposed a novel AD prediction algorithm that uses a multi-omics dataset. We demonstrated that the proposed feature selection approach, which considers the biological context, was more effective than typical dimension reduction algorithms. We also proposed a deep learning-based prediction model that enhanced the prediction performance. To demonstrate the proposed method, we performed various comparative experiments by changing each component of the method. The limitation of our study is that we focused on integrating two molecular layers and did not use the multi-omics dataset from the same sample group. To overcome these limitations, we created all possible sample pairs for each label and a new dataset from it, possibly overfitting the model in this process. Until now, there has been no database that has produced considerable omics data from one sample like that from the TCGA database. In the future, we intend to apply our approach to other multi-omics datasets that have the same sample group, provided we can obtain these datasets.

Declaration of Competing Interest

We declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit authorship contribution statement

Chihyun Park: Conceptualization, Data curation, Formal analysis, Methodology, Validation, Writing - original draft, Writing - review & editing, Software. **Jihwan Ha:** Conceptualization, Investigation, Validation. **Sanghyun Park:** Conceptualization, Funding acquisition, Project administration, Supervision.

Acknowledgments

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the SW Starlab support program (IITP-2017-0-00477) supervised by the IITP (Institute for Information & communications Technology Promotion).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eswa.2019.112873.

References

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., & Tanzi, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet, 39*, 17–23.
Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.
Brookmeyer, R., Johnson, E., Ziegler-Graham, K., & Arrighi, H. M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimers Dement, 3*, 186–191.
Cancer Genome Atlas Research, N., Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet, 45*, 1113–1120.
Chaudhary, K., Poirion, O. B., Lu, L., & Garmire, L. X. (2018). Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res, 24*, 1248–1259.
Choi, J., Park, S., Yoon, Y., & Ahn, J. (2017). Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers. *Bioinformatics, 33*, 3619–3626.
Desikan, R. S., Fan, C. C., Wang, Y., Schork, A. J., Cabral, H. J., Cupples, L. A., et al. (2017). Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLoS Med, 14*, e1002258.
Du, P., Zhang, X., Huang, C. C., Jafari, N., Kibbe, W. A., Hou, L., & Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics, 11*, 587.
Hampel, H., Shen, Y., Walsh, D. M., Aisen, P., Shaw, L. M., Zetterberg, H., et al. (2010). Biological markers of amyloid beta-related mechanisms in Alzheimer's disease. *Exp Neurol, 223*, 334–346.
Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol, 18*, 83.
Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Front Genet, 8*, 84.
Jack, C. R., Jr., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., et al. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement, 14*, 535–562.
John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (pp. 338–345). Canada: Morgan Kaufmann Publishers Inc. Montréal, Qué.

Karczewski, K. J., & Snyder, M. P. (2018). Integrative omics for health and disease. *Nat Rev Genet, 19*, 299–310.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation, 13*, 637–649.

Lee, J. S., Kim, C., Shin, J. H., Cho, H., Shin, D. S., Kim, N., et al. (2018). Machine learning-based individual assessment of cortical atrophy pattern in Alzheimer's disease spectrum: development of the classifier and longitudinal evaluation. *Sci Rep, 8*, 4161.

Liu, C. C., Liu, C. C., Kanekiyo, T., Xu, H., & Bu, G. (2013). Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nat Rev Neurol, 9*, 106–118.

Ma, Y., Jun, G. R., Chung, J., Zhang, X., Kunkle, B. W., Naj, A. C., et al. (2019). CpG-related SNPs in the MS4A region have a dose-dependent effect on risk of late-onset Alzheimer disease. *Aging Cell, 18*(4), e12964.

Maksimovic, J., Phipson, B., & Oshlack, A. (2016). A cross-package Bioconductor workflow for analysing methylation array data. *F1000Res, 5*, 1281.

Munteanu, C. R., Fernandez-Lozano, C., Abad, V. M., Fernández, S. P., Álvarez-Linera, J., Hernández-Tamames, J. A., et al. (2015). Classification of mild cognitive impairment and Alzheimer's Disease with machine-learning techniques using 1H Magnetic Resonance Spectroscopy data. *Expert Syst. Appl., 42*, 6205–6214.

Narayanan, M., Huynh, J. L., Wang, K., Yang, X., Yoo, S., McElwee, J., et al. (2014). Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases. *Mol Syst Biol, 10*, 743.

Park, C., Yoon, Y., Min, O., Yu, S. J., & Ahn, J. (2017). Systematic identification of differential gene network to elucidate Alzheimer's disease. *Expert Syst Appl, 85*, 249–260.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res, 43*, e47.

Smith, R. G., Hannon, E., De Jager, P. L., Chibnik, L., Lott, S. J., Condliffe, D., et al. (2018). Elevated DNA methylation across a 48-kb region spanning the HOXA gene cluster is associated with Alzheimer's disease neuropathology. *Alzheimers Dement, 14*(12), 1580–1588.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2* (pp. 2951–2959). Lake Tahoe, Nevada: Curran Associates Inc.

Song, D., Qi, W., Lv, M., Yuan, C., Tian, K., & Zhang, F. (2018). Combined bioinformatics analysis reveals gene expression and DNA methylation patterns in osteoarthritis. *Mol Med Rep, 17*, 8069–8078.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A, 102*, 15545–15550.

Taylor, D. L., Jackson, A. U., Narisu, N., Hemani, G., Erdos, M. R., Chines, P. S., et al. (2019). Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. *Proc Natl Acad Sci U S A, 116*, 10883–10888.

Yan, J., Risacher, S. L., Shen, L., & Saykin, A. J. (2018). Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief Bioinform, 19*, 1370–1381.

Zhang, B., Gaiteri, C., Bodea, L. G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., et al. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell, 153*, 707–720.

Zhang, L., Lv, C., Jin, Y., Cheng, G., Fu, Y., Yuan, D., et al. (2018). Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. *Front Genet, 9*, 477.

Zhao, C., Zou, H., Zhang, J., Wang, J., & Liu, H. (2018). An integrated methylation and gene expression microarray analysis reveals significant prognostic biomarkers in oral squamous cell carcinoma. *Oncol Rep, 40*, 2637–2647.