Mathematical Biosciences and Engineering

MBE, 20(9): 16648–16662.
DOI: 10.3934/mbe.2023741
Received: 31 March 2023
Revised: 08 June 2023
Accepted: 04 July 2023
Published: 18 August 2023

http://www.aimspress.com/journal/MBE

Research article

Identification of image genetic biomarkers of Alzheimer's disease by orthogonal structured sparse canonical correlation analysis based on a diagnostic information fusion

Wei Yin†, Tao Yang†, GuangYu Wan* and Xiong Zhou

Department of Radiology, Xianning Central Hospital, The First Affiliated Hospital of Hubei University of Science and Technology, Hubei 437000, China

* **Correspondence:** Email: wgy397260449@163.com; Tel: +07158896013; Fax: +07158896013.

† These two authors contributed equally.

Abstract: Alzheimer's disease (AD) is an irreversible neurodegenerative disease, and its incidence increases yearly. Because AD patients will have cognitive impairment and personality changes, it has caused a heavy burden on the family and society. Image genetics takes the structure and function of the brain as a phenotype and studies the influence of genetic variation on the structure and function of the brain. Based on the structural magnetic resonance imaging data and transcriptome data of AD and healthy control samples in the Alzheimer's Disease Neuroimaging Disease database, this paper proposed the use of an orthogonal structured sparse canonical correlation analysis for diagnostic information fusion algorithm. The algorithm added structural constraints to the region of interest (ROI) of the brain. Integrating the diagnostic information of samples can improve the correlation performance between samples. The results showed that the algorithm could extract the correlation between the two modal data and discovered the brain regions most affected by multiple risk genes and their biological significance. In addition, we also verified the diagnostic significance of risk ROIs and risk genes for AD. The code of the proposed algorithm is available at https://github.com/Wanguangyu111/OSSCCA-DIF.

Keywords: Alzheimer's disease; image genetics; canonical correlation analysis; structural constraints; diagnostic model

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disease caused by many factors, and its incidence is increasing yearly [1]. The change of gene expression in genetics involves gene variation. However, the brain structure and function of AD patients are also different from those of the control group. Image genetics explores changes in brain structure and function from the perspective of genetic variation. Through a correlation analysis of imaging and genetics, we can explore the potential markers of AD.

Machine learning algorithms have been widely used in various bioinformatics fields, such as miRNA-disease relationship prediction. Ha and Park [2] proposed the metric learning for predicting miRNA-disease association (MLMD) algorithm, which can reveal not only novel miRNAs associated with disease, but also miRNA-miRNA and disease-disease similarities. Moreover, they proposed the matrix factorization with disease similarity constraint (MDMF) algorithm based on matrix factorization, which incorporates disease similarity constraints to improve the prediction performance [3]. In addition, they proposed the simple yet effective computational framework (SMAP) algorithm to predict the relationship between and accurately predict the association between miRNA-diseases, which combines miRNA functional similarity, disease semantic similarity and Gaussian interaction spectrum kernel similarity [4]. Recently, they combined a deep neural network to propose the node2vec-based neural collaborative filtering for predicting miRNA-disease association (NCMD) algorithm, which uses Node2vec to understand the low-dimensional vector representation of miRNA and disease, and combines the linear ability of generalized matrix factorization and the nonlinear ability of multi-layer perceptron [5]. They tested and confirmed the effectiveness of the algorithm on three datasets of breast cancer, lung cancer and pancreatic cancer. In the search for biomarkers of AD, many scholars have proposed algorithms related to the diagnosis and prediction of AD. Park et al. [6] proposed a deep learning model to integrate large-scale gene expression and DNA methylation data to predict AD. This method is superior to traditional machine learning algorithms in that it uses typical dimensionality reduction methods and improves the accuracy of prediction. Wang et al. created the multi-task sparse canonical correlation analysis and regression (MT-SCCAR) model, combined with the annual total score of depression level, the clinical dementia rating scale, the functional activity questionnaire, and the neuropsychiatric symptom questionnaire. These four clinical data were used as compensation information and embedded in the algorithm by a linear regression. They confirmed the superiority and robustness of the algorithm on real and simulated data [7].

Canonical correlation analysis (CCA) is an algorithm to obtain the maximum correlation between two kinds of data. However, it is not suitable for an association analysis of high-dimensional data. For this reason, some scholars put forward a sparse canonical correlation analysis (SCCA) algorithm [8]. Based on CCA, the SCCA algorithm assists the CCA algorithm in feature selection in high-dimensional features through l1 norm constraints. However, because the l1 norm constraints only considers sparsity at the individual level, it is only partially applicable to image data. Lesions in different brain regions may play a role at the same time; therefore, it is necessary to add structural constraints to the SCCA algorithm. Du et al. [9] proposed a graph-guided pairwise group lasso (GGL) and applied it to image data. GGL can be used in a data-driven mode that does not provide prior knowledge. It thinks that each group consists of only two variables, and they will be extracted with similar or equal weights. They found that the performance of this algorithm with actual data is due to other competitive algorithms. However, they did not consider the diagnostic information of the subjects. Previous studies showed that the addition of diagnostic information could effectively improve the correlation analysis

performance of the algorithm [10,11]. In addition, there may be feature redundancy in imaging and genetic data, and orthogonal constraints can be added to the algorithm by linear programming.

Thus, it was suggested to integrate structural magnetic resonance imaging (sMRI) and the gene expression data of AD patients and its control group using an orthogonal structured SCCA algorithm based on diagnostic information fusion. Specifically, after preprocessing sMRI data, we extracted the gray matter volume of each region of interest (ROI) as a feature. Then, we picked the expression of differentially expressed genes (DEGs) between the sick group and the control group as characteristics from the gene expression data. Based on the SCCA algorithm, this method added GGL constraints on images and orthogonal constraints on two kinds of data, which could improve the performance of the association analysis and prevent the influence of feature redundancy on the results. The experimental results showed that this algorithm was superior to other CCA-based algorithms and had a stronger correlation analysis ability. Top ROIs and top genes with biological and diagnostic significance can be obtained. The selected top biomarkers can provide a reference for the diagnosis of AD and drug target discovery.

2. Materials and methods

2.1. *SCCA*

The SCCA algorithm adds sparse constraints to the CCA algorithm. Given n samples, p ROIs and q genes, sMRI data can be expressed as $X \in R^{n\times p}$, and gene expression data can be expressed as $Y \in R^{n\times q}$. The objective of the SCCA algorithm is to adjust the typical correlation weights $u \in R^{p\times l}$ and $v \in R^{q\times l}$ to maximize the correlation between Xu and Yv, and its objective function is shown in Formula (1):

$$\min_{\mathbf{u},\mathbf{v}} -\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} + \lambda_1\left|\left|\mathbf{u}\right|\right|_1 + \lambda_2\left|\left|\mathbf{v}\right|\right|_1$$

$$\text{s.t.} \parallel \mathbf{Xu} \parallel^2 \leq 1, \parallel \boldsymbol{Y}\mathbf{v} \parallel^2 \leq 1 \tag{1}$$

where λ_1 and λ_2 control the sparsity of $\mathbf{u}$ and $\boldsymbol{v}$, respectively.

2.2. *GGL*

The graph-guided fusion lasso differs from the conventional group lasso in that it does not rely on prior knowledge; however, the graph-guided fusion lasso will introduce estimation bias. GGL uses group lasso and graph-guided fusion lasso. It can be defined as follows:

$$\Omega_{\text{GGL}}(\mathbf{u}) = \sum_{(i,j)\in E}\sqrt{u_i^2 + u_j^2} \tag{2}$$

where E is the edge set of the graph which highly related features are connected.

2.3. *Orthogonal structured sparse canonical correlation analysis for diagnostic information fusion (OSSCCA-DIF)*

This paper presented an OSSCCA-DIF algorithm, which uses GGL as a structural constraint based on the SCCA algorithm. Orthogonal constraints were used to prevent the redundant characteristics of

image genetics data from affecting the results. In addition, we added the diagnostic information of samples as prior knowledge to the algorithm to improve its correlation performance. The objective function of the OSSCCA-DIF algorithm is given as follows:

$$\min_{\mathbf{u},\mathbf{v}} -\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} + \lambda_1||\mathbf{u}||_1 + \lambda_2||\mathbf{v}||_1 + \mathbf{X}^T z + \beta_1\|\mathrm{uu^T} - \mathrm{I}\|_2^2 + \beta_2\|\mathrm{vv^T} - \mathrm{I}\|_2^2 + \mu\Omega_{\mathrm{GGL}}(\mathbf{u})$$

$$\text{s.t.} \parallel \mathbf{Xu} \parallel^2 \leq 1, \parallel \boldsymbol{Y}\mathbf{v} \parallel^2 \leq 1 \quad (3)$$

where $z \in R^{n\times 1}$ is used to store the diagnostic information of the sample, β_1 and β_2 control the orthogonal constraint strength of $\mathbf{u}$ and $\boldsymbol{v}$, respectively, and γ is applied to control the strength of the GGL constraint.

2.4. *The optimization algorithm*

For the optimization of Formula (3), the Lagrange multiplier method can be used to solve the partial derivatives of the weight $\mathbf{u}$ of the ROI and the weight $\mathbf{v}$ of the gene, respectively. Firstly, $\mathbf{u}$ is regarded as a constant term. Then, the objective function can be rewritten as Eq (4):

$$\mathcal{L}(\mathbf{u},\mathbf{v}) = -\mathbf{u}^T\mathbf{X}^T\boldsymbol{Y}\mathbf{v} + \beta_1\|\mathrm{uu^T} - \mathrm{I}\|_2^2 + \mathbf{X}^T z + \beta_2\|\mathrm{vv^T} - \mathrm{I}\|_2^2 + \parallel \mathbf{Xu} \parallel^2 + \parallel \boldsymbol{Y}\mathbf{v} \parallel^2 + \mu\Omega_{\mathrm{GGL}}(\mathbf{u}) + \lambda_1||\mathbf{u}||_1 + \lambda_2||\mathbf{v}||_1 + \mathbf{X}^T z. \quad (4)$$

For Eq (4), take the derivative of $\mathbf{v}$ and make it zero, and we can get Eq (5):

$$\mathrm{V}^T\mathrm{Y}\mathbf{v}\text{-}\mathrm{Y}^T\mathrm{Xu} + 2\beta_2(\mathrm{vv^T} - \mathrm{I})\mathrm{v} + \lambda_2\mathrm{D_v}\mathbf{v} + 2\mathrm{Y}^T\mathrm{Y}\mathbf{v} = 0. \quad (5)$$

The iterative solution formula of $\mathbf{v}$ can be written as Formula (6):

$$\mathbf{v} = (\mathrm{Y}^T\mathrm{Y} + \lambda_2\mathrm{D_v} + \mathbf{X}^T z + 2\mathrm{Y}^T\mathrm{Y} + 2\beta_2(\mathrm{vv^T} - \mathrm{I})\mathrm{v})^{-1}(\mathrm{Y}^T\mathrm{Xu}). \quad (6)$$

Similarly, for u, if v is regarded as a constant term, then the solution formula of u can be obtained, as shown in Eq (7):

$$\mathbf{u} = (\mathrm{X}^T\mathrm{X} + \lambda_1\mathrm{D_u} + 2\mathrm{X}^T\mathrm{X} + \mu\mathrm{G_u} + 2\beta_1(\mathrm{uu^T} - \mathrm{I})\mathrm{u})^{-1}(\mathrm{X}^T\mathrm{Yv}). \quad (7)$$

In Eq (7), $\boldsymbol{D}_{\mathrm{u}}$ and $\boldsymbol{D}_{\mathrm{v}}$ are diagonal matrices, and their diagonal elements in the ith row can be expressed as $\frac{1}{|u_{1i}|}(i = 1, \cdots, p)$ and $\frac{1}{|v_{1i}|}(i = 1, \cdots, q)$, respectively. $\mathrm{G_u}$ is also a diagonal matrix, and the diagonal elements in the ith row can be expressed as $\frac{1}{\sqrt{u_{1(i-1)}^2 + u_{1i}^2}} + \frac{1}{\sqrt{u_{1i}^2 + u_{1(i+1)}^2}}(i = 2, \cdots, p-1)$.

3. Results

3.1. *Data acquisition and pretreatment*

We obtained 296 samples of AD, mild cognitive impairment (MCI), and healthy controls (HC) from the Alzheimer's Disease Neuroimaging Disease (ADNI) database (https://adni.loni.usc.edu/). Table 1 provides the statistical information for each set of samples. We collected sMRI and gene expression data from these samples. For sMRI, we first calibrated the head movement using the DiffusionKit software and then segmented the images using the Statistical Parametric Mapping

(SPM) software package of the Computational Anatomy Toolbox (CAT) toolkit of Matlab software and divided them into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF). In this paper, we used a GM volume of 140 brain regions divided by the brain as the ROI. For the gene expression data, we utilized the limma algorithm to examine the differential expression and obtained 962 DEGs. We extracted the expression of DEGs as the genetic feature of the sample.

Table 1. Sample statistics.

	Gender	MMSE	Age
AD	10/15	20.48 ± 10.22	75.99 ± 4.28
MCI	90/115	28.13 ± 1.734	71.48 ± 7.57
HC	31/34	28.94 ± 1.25	75.17 ± 5.86

3.2. *Parameter selection and results of the algorithm*

In order to obtain the best result, this paper took the canonical correlation coefficient (CCC) as the performance measure of the algorithm, and the solution method of the CCC was as follows. Using the grid search method [0.01 0.1 1], the parameters of super parameters (β_1, β_2, λ_1, λ_2, and μ) of the OSSCCA-DIF algorithm were selected for the real data sets (Figure 1). Finally, the results of the sixth parameter selection were taken as the best parameters ($\beta_1 = 0.01$, $\beta_2 = 0.01$, $\lambda_1 = 0.01$, $\lambda_2 = 0.1$ and $\mu = 1$). We detail the optimal parameter information in Table S1 of the Supplementary material. It can be seen from the figure that the CCC obtained by different parameter combinations is quite different. The proposed algorithm is sensitive to the parameters and thus affects the stability of the model. Therefore, it is helpful to select the optimal results by appropriately enlarging the selection range of parameters in the practical application. Figure 2 shows the heat map of weights u and v. Specifically, the abscissa of the graph represents either an ROI or a gene. The distribution of different colors represents the weight of the ROIs or genes; the darker the color, the higher the weight. After taking the absolute values of U and V, we give the names and weight information of the top 10 ROIs and the top 10 genes with the greatest weight in Table 2. We will discuss the biological significance of these ROIs and genes in detail in the discussion section. Figure 3 is a visual display of the top 10 ROI. Figure 4 displays the enrichment analysis results of the top 10 genes. We will discuss the relationship between these channels and AD in detail in the discussion section.

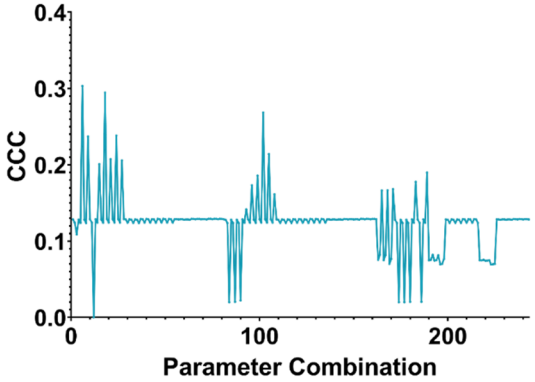

Figure 1. The line chart of optimal parameter selection.

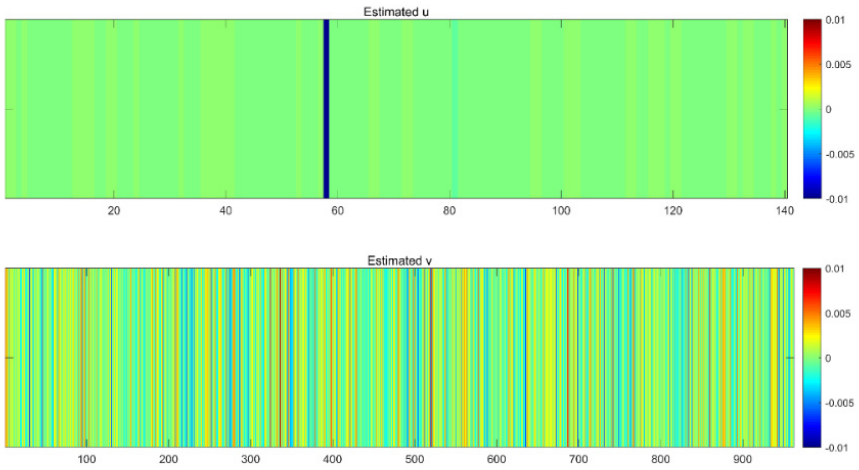

Figure 2. The weight heatmap of **u** and **v**.

Table 2 Weight information of the top ROIs and top genes.

ROI	Weights	Gene	Weights
rEnt	0.06482	ADAM17	0.006038
lAmy	0.000438	TBX4	0.00569
rAmy	1.98×10^{-06}	ZC2HC1C	0.005581
lFroOpe	2.94×10^{-07}	CCND1	0.005188
lMidOccGy	8.01×10^{-08}	POM121L12	0.004765
rInfTemGy	3.23×10^{-08}	HIST1H2BM	0.004729
rParHipGy	3.10×10^{-08}	ALDH3A1	0.004531
lHip	2.67×10^{-08}	ANO5	0.004486
rTemPo	2.38×10^{-08}	SPDYE4	0.004477
rFusGy	2.34×10^{-08}	HERC2P9	0.00443

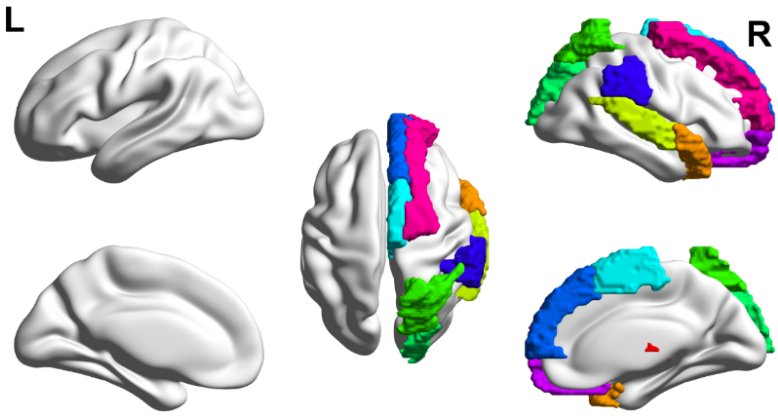

Figure 3. Visualization of the top ROIs.

To prove that the proposed algorithm has a good correlation analysis ability, we compared the CCC of the proposed algorithm with the SCCA-FGL, SCCA, and CCA algorithms (Table 3), and the CCC of the proposed algorithm was higher than the other three algorithms. In addition, we present the Pearson correlation heat map of the top 10 ROIs and top 10 genes in Figure 5. Among them, lHip and ZC2HC1C reached the maximum positive correlation (0.4551), while lMidOccGy and POM121L12 reached the maximum negative correlation (−0.3227).

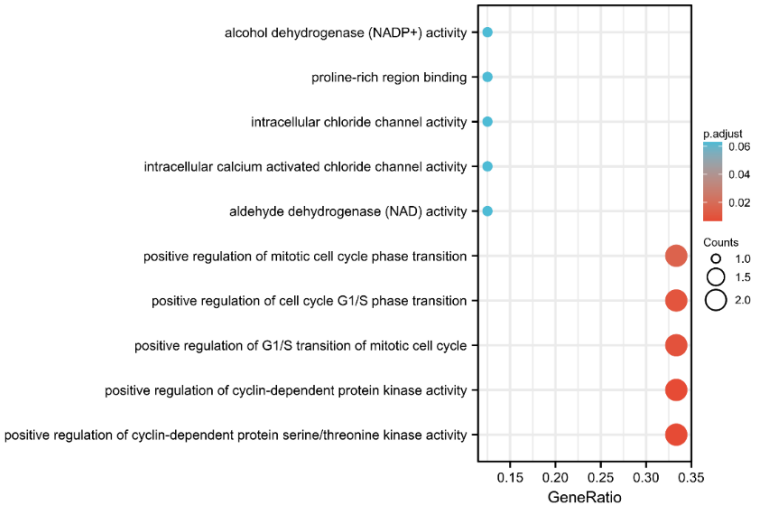

Figure 4. Enrichment analysis results of the top genes.

Table 3 Performance comparison of CCA-based algorithms.

Algorithm	CCC
OSSCCA-DIF	**0.3033**
SCCA_FGL	0.2357
SCCA	0.1563
CCA	0.0427

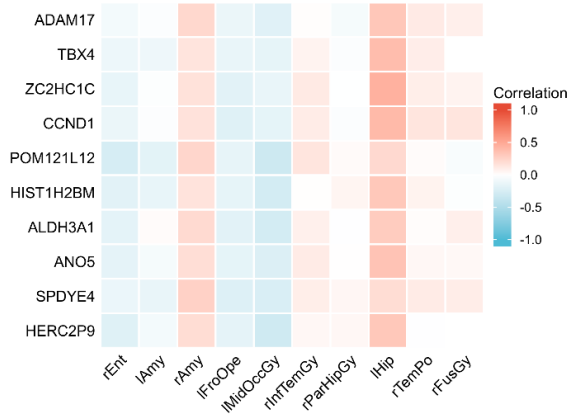

Figure 5. Correlation thermogram between the top ROIs and top genes.

Note: rEnt = Right Entorhinal Area, lAmy = Left Amygdala, rAmy = Right Amygdala, lFroOpe = Left Frontal Operculum, lMidOccGy = Left Middle Occipital Gyrus, rInfTemGy = Right Inferior Temporal Gyrus, rParHipGy = Right Parahippocampus Gyrus, lHip = Left Hippocampus, rTemPo = Right Temporal Pole, rFusGy = Right Fusiform Gyrus.

3.3. *Diagnostic performance verification of top markers*

In this section, we used the Receiver Operating Characteristic (ROC) curve to verify the diagnostic performance of the top markers (Figure 6). ROC curves have been widely used in the biomedical field [12,13]. Based on the logistic regression algorithm, we constructed the diagnosis model using the top 10 ROIs and top 10 genes. Among them, the AUC of the diagnosis model constructed by the ROIs reached 0.898. The AUC of the diagnosis model created by the genes reached 0.853. In addition, we present the diagnostic models constructed using the top 10 ROIs and the top 10 genes from several other algorithms in Figure S1 of the Supplementary material. The AUC of the top 10 ROIs selected by the proposed algorithm was higher than that of several other algorithms. The AUC of the top 10 genes selected was slightly lower than that of the CCA algorithm. In addition, we present the details of the ROC curves of the diagnostic models constructed by the top ROIs/genes selected by the OSSCCA-DIF, SCCA-FGL, SCCA, and CCA algorithms in Table S2 and Table S3 of the Supplementary material.

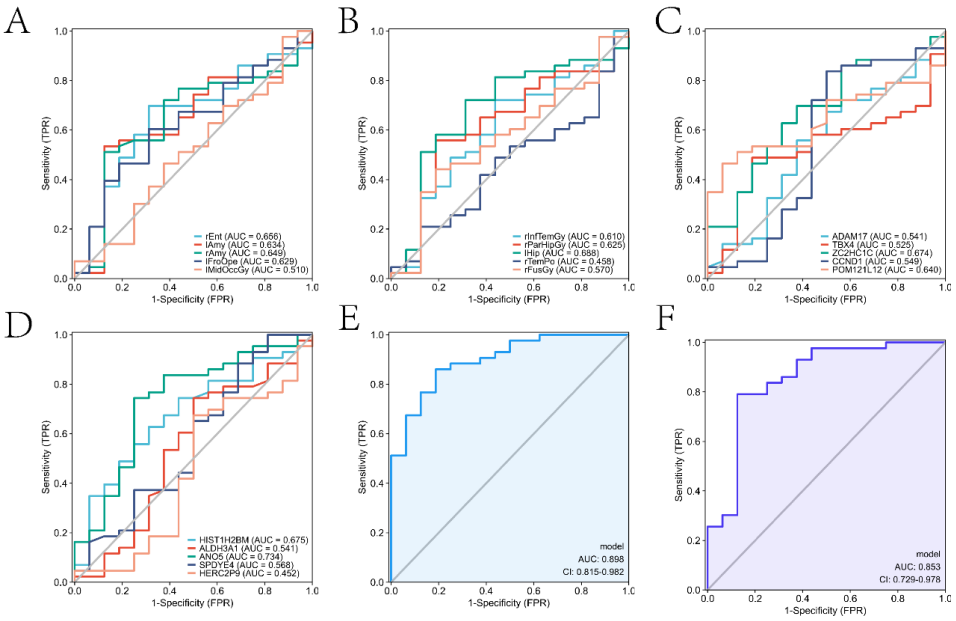

Figure 6. Diagnostic performance verification of TOP markers. A and B are ROC curves of the first five ROIs and the last five ROIs, respectively. C and D are ROC curves of the first five genes and the last five genes, respectively. E and F are ROC curves of the diagnosis model constructed by the top 10 ROIs and top 10 genes, respectively.

3.4. *Experimental results on the synthetic dataset*

In order to further verify the effectiveness of the algorithm, we constructed two synthetic data

sets ($\boldsymbol{X_1} \in R^{n1\times p1}$, $\boldsymbol{Y_1} \in R^{n1\times q1}$, $\boldsymbol{X_2} \in R^{n2\times p2}$, $\boldsymbol{X_2} \in R^{n2\times q2}$), and generated two sets of weights ($\boldsymbol{u_1} \in R^{p1\times 1}$, $\boldsymbol{v_1} \in R^{q1\times 1}$, $\boldsymbol{u_2} \in R^{p2\times 1}$, $\boldsymbol{v_2} \in R^{q2\times 1}$). Here, $n1 = 100$, $n2 = 500$, $p1$=300, $q1 = 500$, and $p2 = 800$. In addition, we generated the variable $\boldsymbol{z} \in R^{n\times 1}$. $\boldsymbol{X_k}$ and $\boldsymbol{Y_k}$ can be generated by $(x_{ij})_k \sim \mathcal{N}(z_i u_{kj}, l * \sigma_k)$. Here, $k = 1, 2$. σ_k and l denote the variance and noise level of the noise, respectively. We set l to 10 and present the CCCS of several algorithms on two synthetic datasets in Table 4. As can be seen from the table, the CCC of the proposed algorithm is larger than that of the other algorithms on both datasets.

Table 4. Performance comparison of CCA-based algorithms in synthetic data sets.

Algorithm	CCC (synthetic dataset 1)	CCC (synthetic dataset 2)
OSSCCA-DIF	0.2733	0.2320
SCCA_FGL	0.0372	0.1119
SCCA	0.1510	0.0836
CCA	0.1692	0.0424

3.5. Results of ablation experiments on real and simulated datasets

Additionally, we conducted ablation experiments on the proposed OSSCCA-DIF algorithm. Specifically, by removing each part of the OSSCCA-DIF algorithm either individually or in pairs (except for sparse constraints), we compare the CCC between the real and simulation datasets under the same parameter conditions (Table 5). The objective functions to be compared (scenarios 1–6) are given below, as shown in Eqs (8)–(13):

$$\mathcal{L}(\mathbf{u},\mathbf{v}) = -\mathbf{u}^T\mathbf{X}^T\boldsymbol{Y}\mathbf{v} + \| \mathbf{X}\mathbf{u} \|^2 + \| \boldsymbol{Y}\mathbf{v} \|^2 + \mu\Omega_{\mathrm{GGL}}(\mathbf{u}) + \lambda_1||\mathbf{u}||_1 + \lambda_2||\mathbf{v}||_1, \quad (8)$$

$$\mathcal{L}(\mathbf{u},\mathbf{v}) = -\mathbf{u}^T\mathbf{X}^T\boldsymbol{Y}\mathbf{v} + \beta_1\|\mathrm{uu}^{\mathrm{T}} - \mathrm{I}\|_2^2 + \beta_2\|\mathrm{vv}^{\mathrm{T}} - \mathrm{I}\|_2^2 + \| \mathbf{X}\mathbf{u} \|^2 + \| \boldsymbol{Y}\mathbf{v} \|^2 + \lambda_1||\mathbf{u}||_1 + \lambda_2||\mathbf{v}||_1, \quad (9)$$

$$\mathcal{L}(\mathbf{u},\mathbf{v}) = -\mathbf{u}^T\mathbf{X}^T\boldsymbol{Y}\mathbf{v} + \| \mathbf{X}\mathbf{u} \|^2 + \| \boldsymbol{Y}\mathbf{v} \|^2 + \lambda_1||\mathbf{u}||_1 + \lambda_2||\mathbf{v}||_1 + \mathbf{X}^T z, \quad (10)$$

$$\mathcal{L}(\mathbf{u},\mathbf{v}) = -\mathbf{u}^T\mathbf{X}^T\boldsymbol{Y}\mathbf{v} + \beta_1\|\mathrm{uu}^{\mathrm{T}} - \mathrm{I}\|_2^2 + \beta_2\|\mathrm{vv}^{\mathrm{T}} - \mathrm{I}\|_2^2 + \| \mathbf{X}\mathbf{u} \|^2 + \| \boldsymbol{Y}\mathbf{v} \|^2 + \mu\Omega_{\mathrm{GGL}}(\mathbf{u}) + \lambda_1||\mathbf{u}||_1 + \lambda_2||\mathbf{v}||_1, \quad (11)$$

$$\mathcal{L}(\mathbf{u},\mathbf{v}) = -\mathbf{u}^T\mathbf{X}^T\boldsymbol{Y}\mathbf{v} + \beta_1\|\mathrm{uu}^{\mathrm{T}} - \mathrm{I}\|_2^2 + \beta_2\|\mathrm{vv}^{\mathrm{T}} - \mathrm{I}\|_2^2 + \| \mathbf{X}\mathbf{u} \|^2 + \| \boldsymbol{Y}\mathbf{v} \|^2 + \lambda_1||\mathbf{u}||_1 + \lambda_2||\mathbf{v}||_1 + \mathbf{X}^T z, \quad (12)$$

$$\mathcal{L}(\mathbf{u},\mathbf{v}) = -\mathbf{u}^T\mathbf{X}^T\boldsymbol{Y}\mathbf{v} + \mu\Omega_{\mathrm{GGL}}(\mathbf{u}) + \lambda_1||\mathbf{u}||_1 + \lambda_2||\mathbf{v}||_1 + \mathbf{X}^T z + \| \mathbf{X}\mathbf{u} \|^2 + \| \boldsymbol{Y}\mathbf{v} \|^2. \quad (13)$$

Table 5. Shows six scenarios and the CCC of OSCCA-DIF on real datasets and two simulated datasets.

Case	CCC (real dataset 1)	CCC (synthetic dataset 1)	CCC (synthetic dataset 2)
Case 1	0.1427	0.1236	0.1198
Case 2	0.1542	0.1468	0.1333
Case 3	0.1642	0.1539	0.1248
Case 4	0.1344	0.1326	0.1164
Case 5	0.1529	0.1432	0.1409
Case 6	0.1243	0.1142	0.1022
OSCCA-DIF	**0.3033**	**0.2733**	**0.2320**

4. Discussion

As a neurodegenerative disease, there is no effective treatment for AD. Alongside aging, the incidence of AD is increasing year by year. Image genetics can mine disease-related markers by integrating image genomics and genetic data through a series of correlation analysis algorithms. Therefore, this paper proposed an OSSCCA-DIF algorithm. Based on the SCCA algorithm, this algorithm added orthogonal constraints on weight vectors u and v and GGL structural constraints on image feature weight u. In addition, the diagnostic information of the sample was added to the algorithm. The experimental results showed that by integrating sMRI and ROI data in real data, the performance of this algorithm was better than other CCA-based algorithms.

Most of the top ROIs mined by the proposed algorithm have proven to be closely related to AD. First, our algorithm determined that rEnt was the brain region with the most significant weight, and the weight value reached 0.06482. The entorhinal cortex (EC) is unanimously considered to be the earliest pathological structure of AD [14,15]. Thaker et al. [16] analyzed the relationship between the thickness of the olfactory cortex (sMRI) and pathological changes (autopsy) in 50 AD patients and found that the thickness of the olfactory cortex may be related to the severity of AD. The experiment confirmed that, compared with the control group, the volume and average thickness of the right inner olfactory cortex in AD patients were lower, and the expression level of lncRNA BACE1-AS in plasma exosomes isolated from AD patients was significantly increased. Therefore, Wang et al. [17] proposed that the level of BACE1-AS in peripheral blood exosomes should be combined with the volume and thickness of the right entorhinal cortex as a potential biomarker of AD. Second, our algorithm identified that lAmy and rAmy were top ROIs. These two brain regions are both sides of the amygdala. As an important structure of emotional learning and memory, the amygdala is related to a series of mental diseases such as AD. The MRI volume of the amygdala may be related to the severity of dementia in AD, and it shows neuron loss and atrophy in AD patients [18–20]. The amygdala has an excellent diagnostic value for sMRI of AD [21]. Finally, our algorithm found that the top brain regions (rParHipGy and lHip) also proved to be closely related to AD. In the memory system of the human brain, it is essential to connect the posterior cingulate cortex with the hippocampus, either directly or indirectly through the parahippocampal gyrus. These brain regions all play a vital role in the early progression of AD [22]. In the experiment evaluating the correlation between brain metabolism and the orientation of AD, it was found that its improved orientation performance was related to the more significant brain metabolism in brain regions such as the left middle occipital gyrus, and the higher CERAD identification score was more related to the metabolic activity in the left medial temporal lobe regions (including the hippocampus, parahippocampal gyrus, and left fusiform gyrus) [23].

The top genes (ADAM17, CCND1, HIST1H2BM, ALDH3A1) determined by our algorithm are proven to be either directly or indirectly related to AD. It is common knowledge that a feature of AD is the accumulation of extracellular Amyloid-β (Aβ) plaques and neurofibrillary tangles composed of tau in neurons [24–26]. Among them, Aβ is a protein hydrolysate separated from its precursor, namely the amyloid precursor protein (APP), by β- and γ-secretase, and tau is a microtubule-associated protein involved in microtubule stability. The main manifestations of AD patients are decreased memory, attention, spatial orientation, language ability, and olfactory function, which are all related to the deposition of tau protein and APP. It has been proven that ADAM17 is a potential therapeutic target for AD because ADAM17 can be used as an α-secretase regulating APP, thereby affecting the production of Aβ.

Additionally, protease encoded by ADAM17 plays a role in the shedding of tumor necrosis factor-α (TNF-α). As a key pro-inflammatory cytokine in inflammation, TNF-α's signal transduction aggravates Aβ and tau pathology in vivo [27]. In order to explore the role of propionic acid (PPA) in the pathogenesis of AD, Aliashrafi et al. [28] selected 284 genes related to PPA and AD and identified CCND1 as an important hub gene, bottleneck gene, and seed gene through a network analysis and an Molecular Complex Detection (MCODE) analysis. Zeng et al. [29] also determined that CCND1 can be the core goal of AD treatment. H2BC14 (HIST1H2BM) is the core component of the nucleosome. The nucleosome assembly protein 1-like 5(NAP1L5) is downregulated in the brain tissue of AD patients, and the overexpression of NAP1L5 can alleviate APP metabolism and Tau phosphorylation [30]. ALDH3A1, a protein-coding gene, belongs to A1, a member of the aldehyde dehydrogenase 3 family. In the enrichment analysis of the top 10 genes, we also found that aldehyde dehydrogenase is closely related to AD. The relationship between other genes and AD needs further study.

Through the enrichment analysis of the top 10 genes, we found that many significant pathways were related to the occurrence and development of AD. McKibben and Rhoades [31] studied the role of the proline-rich region (PRR) in regulating the interaction between Tau and soluble tubulin. Aldehyde dehydrogenase can balance the amine metabolism of neurodegenerative diseases such as AD [32]. Tao et al. [33] indicated that aldehyde dehydrogenase-2 could be a potential target for AD treatment. Rapamycin can block G1/S conversion between AD patients and normal controls. Experiments have confirmed that compared with the control group, AD patients who used rapamycin still progressed to the late cell cycle [34]. Cyclin dependent kinase 5 (Cdk5) can also be used as a potential therapeutic target for AD [35].

Finally, we also determined the diagnostic significance of the top markers for AD. We verified the AUC of each top ROI and gene in the test set and found that the AUCs of all ROIs and genes were more significant than 0.5. In addition, through the logistic regression algorithm, we found that the AUC of the diagnosis model constructed by the top 10 ROIs and top 10 genes reached 0.898 and 0.853, respectively, and were both within reasonable confidence intervals.

5. Conclusions

In this paper, we discussed the risk brain regions and risk genes closely related to the occurrence and development of AD by studying the relationship between sMRI and bivariate variables of gene expression data. The proposed OSSCCA-DIF algorithm has advantages in correlation analysis performance and biomarker selection. However, most algorithms based on the SCCA algorithm assume that the image genetic data is linear; however, this assumption is not necessarily valid in real data. Therefore, in future research, we will try to introduce a deep-learning algorithm to make up for this defect.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflict of interest.

Data Availability Statement

The data used in this paper came the AD Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu).

References

1. R. Au, R. J. Piers, L. Lancashire, Back to the future: Alzheimer's disease heterogeneity revisited, *Alzheimer's Dementia: Diagn. Assess. Dis. Monit.*, **1** (2015), 368–370. https://doi.org/10.1016/j.dadm.2015.05.006
2. J. Ha, C. Park, MLMD: Metric learning for predicting MiRNA-disease associations, *IEEE Access*, **9** (2021), 78847–78858, https://doi.org/10.1109/ACCESS.2021.3084148
3. J. Ha, MDMF: Predicting miRNA-Disease association based on matrix factorization with disease similarity constraint, *J. Pers. Med.*, **12** (2022), 885. https://doi.org/10.3390/jpm12060885
4. J. Ha, SMAP: Similarity-based matrix factorization framework for inferring miRNA-disease association, *Knowl.-Based Syst.*, **263** (2023), 110295. https://doi.org/10.1016/j.knosys.2023.110295
5. J. Ha, S. Park, NCMD: Node2vec-based neural collaborative filtering for predicting MiRNA-disease association, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **20** (2023), 1257–1268. https://doi.org/10.1109/TCBB.2022.3191972
6. C. Park, J. Ha, S. Park, Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset, *Expert Syst. Appl.*, **140** (2019), 112873. https://doi.org/10.1016/j.eswa.2019.112873
7. S. Wang, H. Chen, W. Kong, F. Ke, K.Wei, Identify biomarkers of alzheimer's disease based on multi-task canonical correlation analysis and regression model, *J. Mol. Neurosci.*, **72** (2022), 1749–1763. https://doi.org/10.1007/s12031-022-02031-9
8. D. M. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, *Biostatistics*, **10** (2009), 515–534. https://doi.org/10.1093/biostatistics/kxp008
9. L. Du, T. Zhang, K. Liu, J. Yan, X. Yao, S. L. Risacher, et al., Identifying associations between brain imaging phenotypes and genetic factors via a novel structured SCCA approach, in *International Conference on Information Processing in Medical Imaging*, **10265** (2017), 543–555. https://doi.org/10.1007/978-3-319-59050-9_43
10. L. Du, F. Liu, K. Liu, X. Yao, S. L. Risacher, J. Han, et al., Identifying diagnosis-specific genotype-phenotype associations via joint multitask sparse canonical correlation analysis and classification, *Bioinformatics*, **36** (2020), 371–379. https://doi.org/10.1093/bioinformatics/btaa434
11. X. Hao, Q. Tan, Y. Guo, Y. Xiao, M. Yu, M. Wang, et al., Identifying modality-consistent and modality-specific features via label-guided multi-task sparse canonical correlation analysis for neuroimaging genetics, *IEEE Trans. Biomed. Eng.*, **70** (2023), 831–840. https://doi.org/10.1109/TBME.2022.3203152
12. N. Q. K. Le, D. T. Do, T. T. Nguyen, Q. A. Le, A sequence-based prediction of Kruppel-like factors proteins using XGBoost and optimized features, *Gene*, **787** (2021), 145643. https://doi.org/10.1016/j.gene.2021.145643

13. Q. H. Kha, Q. T. Ho, N. Q. K. Le, Identifying SNARE proteins using an alignment-free method based on multiscan convolutional neural network and PSSM profiles, *J. Chem. Inf. Model.*, **62** (2022), 4820–4826. https://doi.org/10.1021/acs.jcim.2c01034
14. J. Zhan, M. Brys, L. Glodzik, W. Tsui, E. Javier, J. Wegiel, et al., An entorhinal cortex sulcal pattern is associated with Alzheimer's disease, *Hum. Brain Mapp.*, **30** (2009), 874–882. https://doi.org/10.1002/hbm.20549
15. M. Zhou, F. Zhang, L. Zhao, J. Qian, C. Dong, Entorhinal cortex: a good biomarker of mild cognitive impairment and mild Alzheimer's disease, *Rev. Neurosci.*, **27** (2016), 185–195. https://doi.org/10.1515/revneuro-2015-0019
16. A. A. Thaker, B. D. Weinberg, W. P. Dillon, C. P. Hess, H. J. Cabral, D. A. Fleischman, et al., Entorhinal cortex: Antemortem cortical thickness and postmortem neurofibrillary tangles and amyloid Pathology, *Am. J. Neuroradiol.*, **38** (2017), 961–965. https://doi.org/10.3174/ajnr.A5133
17. D. Wang, P. Wang, X. Bian, S. Xu, Q. Zhou, Y. Zhang, et al., Elevated plasma levels of exosomal BACE1-AS combined with the volume and thickness of the right entorhinal cortex may serve as a biomarker for the detection of Alzheimer's disease, *Mol. Med. Rep.*, **22** (2020), 227–238. https://doi.org/10.3892/mmr.2020.11118
18. T. H. L. G. Vereecken, O. J. M. Vogels, R. Nieuwenhuys, Neuron loss and shrinkage in the amygdala in Alzheimer's disease, *Neurobiol. Aging*, **15** (1994), 45–54. https://doi.org/10.1016/0197-4580(94)90143-0
19. C. L. Grady, M. L. Furey, P. Pietrini, B. Horwitz, S. I. Rapoport, Altered brain functional connectivity and impaired short-term memory in Alzheimer's disease, *Brain*, **124** (2001), 739–756. https://doi.org/10.1093/brain/124.4.739
20. D. Horínek, A. Varjassyová, J. Hort, Magnetic resonance analysis of amygdalar volume in Alzheimer's disease, *Curr. Opin. Psychiatry*, **20** (2007), 273–277. https://doi.org/10.1097/YCO.0b013e3280ebb613
21. D. W. Wang, S. L. Ding, X. L. Bian, S. Y. Zhou, H. Yang, P. Wang, Diagnostic value of amygdala volume on structural magnetic resonance imaging in Alzheimer's disease, *World J. Clin. Cases*, **9** (2021), 4627–4636. https://doi.org/10.12998/wjcc.v9.i18.4627
22. J. Soldner, T. Meindl, W. Koch, A. L. W. Bokde, M. F. Reiser, H. Möller, et al., Strukturelle und funktionelle neuronale Konnektivität bei der Alzheimer-Krankheit, *Nervenarzt*, **83** (2012), 878–887. https://doi.org/10.1007/s00115-011-3326-3
23. G. H. Weissberger, R. J. Melrose, C. M. Fanale, J. V. Veliz, D. L. Sultzer, Cortical Metabolic and Cognitive Correlates of Disorientation in Alzheimer's Disease, *J. Alzheimer's Dis.*, **60** (2017), 707–719. https://doi.org/10.3233/JAD-170420
24. M. A. Busche, B. T. Hyman, Synergy between amyloid-β and tau in Alzheimer's disease, *Nat. Neurosci.*, **23** (2020), 1183–1193. https://doi.org/10.1038/s41593-020-0687-6
25. C. W. Chang, E. Shao, L. Mucke, Tau: Enabler of diverse brain disorders and target of rapidly evolving therapeutic strategies, *Science*, **371** (2021). https://doi.org/10.1126/science.abb8255
26. K. Stefanoska, M. Gajwani, A. R. P. Tan, H. I. Ahel, P. R. Asih, A. Volkerling, et al., Alzheimer's disease: Ablating single master site abolishes tau hyperphosphorylation, *Sci. Adv.*, **8** (2022). https://doi.org/10.1126/sciadv.abl8809
27. B. Decourt, D. K. Lahiri, M. N. Sabbagh, Targeting tumor necrosis factor alpha for Alzheimer's disease, *Curr. Alzheimer Res.*, **14** (2017), 412–425. https://doi.org/10.2174/1567205013666160930110551

28. M. Aliashrafi, M. Nasehi, M. R. Zarrindast, M. T. Joghataei, H. Zali, S. D. Siadat, Association of microbiota-derived propionic acid and Alzheimer's disease bioinformatics analysis, *J. Diabetes Metab. Disord.*, **19** (2020), 783–804, https://doi.org/10.1007/s40200-020-00564-7
29. P. Zeng, H. F Su., C. Y. Ye, S. W. Qiu, Q. Tian, Therapeutic mechanism and key alkaloids of uncaria rhynchophylla in alzheimer's disease from the perspective of pathophysiological processes, *Front. Pharmacol.*, **12** (2021), 806984. https://doi.org/10.3389/fphar.2021.806984
30. B. Wang, W. Liu, F. Sun, Nucleosome assembly protein 1-like 5 alleviates Alzheimer's disease-like pathological characteristics in a cell model, *Front. Mol. Neurosci.*, **15** (2022), 1034766. https://doi.org/10.3389/fnmol.2022.1034766
31. K. M. McKibben, E. Rhoades, Independent tubulin binding and polymerization by the proline-rich region of Tau is regulated by Tau's N-terminal domain, *J. Biol. Chem.*, **294** (2019), 19381–19394. https://doi.org/10.1074/jbc.RA119.010172
32. E. Grünblatt, P. Riederer, Aldehyde dehydrogenase (ALDH) in Alzheimer's and Parkinson's disease, *J. Neural Transm.*, **123** (2016), 83–90. https://doi.org/10.1007/s00702-014-1320-1
33. R. Tao, M. Liao, Y. Wang, H. Wang, Y. Tan, S. Qin, et al., In situ imaging of formaldehyde in live mice with high spatiotemporal resolution reveals aldehyde dehydrogenase-2 as a potential target for Alzheimer's disease treatment, *Anal. Chem.*, **94** (2022), 1308–1317. https://doi.org/10.1021/acs.analchem.1c04520
34. M. Song, Y. A. Kwon, Y. Lee, H. Kim, J. H. Yun, S. Kim, et al., G1/S cell cycle checkpoint defect in lymphocytes from patients with Alzheimer's disease, *Psychiatry Invest.*, **9** (2012), 413–417. https://doi.org/10.4306/pi.2012.9.4.413
35. A. S. Bhounsule, L. K. Bhatt, K. S. Prabhavalkar, M. Oza, Cyclin dependent kinase 5: A novel avenue for Alzheimer's disease, *Brain Res. Bull.*, **132** (2017), 28–38. https://doi.org/10.1016/j.brainresbull.2017.05.006

Supplementary

Table S1. Detailed parameter Settings of the model on real datasets.

Hyperparameter	Value
β_1	0.01
β_2	0.01
λ_1	0.01
λ_2	0.1
μ	1
Random seed	1

Table S2. Details of the ROC curves of the diagnostic models constructed by several algorithms based on top ROIs.

Algorithm	Area under curve	Confidence interval	Sensitivity	Specificity	Youden index
OSSCA-DIF	0.898	0.815–0.912	0.86	0.812	0.673
SCCA-FGL	0.85	0.741–0.960	0.791	0.875	0.666
SCCA	0.712	0.562–0.863	0.605	0.812	0.417
CCA	0.89	0.807–0.972	0.767	0.875	0.642

Table S3. Details of the ROC curves of the diagnostic models constructed by several algorithms based on top Genes.

Algorithm	Area under curve	Confidence interval	Sensitivity	Specificity	Youden index
OSSCA-DIF	0.853	0.729–0.978	0.791	0.875	0.666
SCCA-FGL	0.763	0.635–0.891	0.581	0.875	0.456
SCCA	0.760	0.628–0.893	0.907	0.5	0.407
CCA	0.865	0.770–0.959	0.791	0.812	0.603

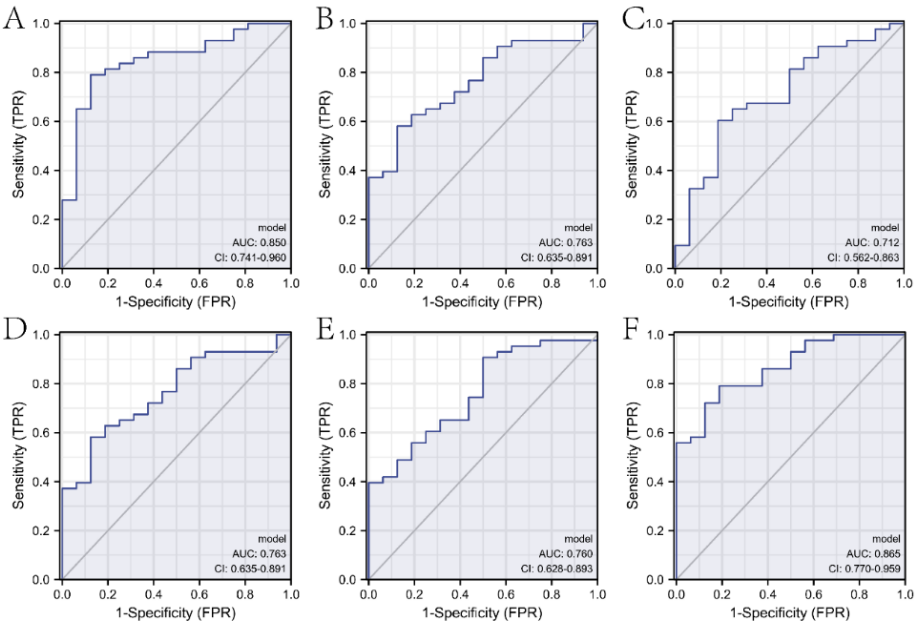

Figure S1. ROC curves of the diagnostic models constructed by the top 10 ROIs and top 10 genes selected by the SCCA-FGL, SCCA and CCA algorithms. A, C, and E are the ROC curves of the diagnostic models constructed by the top 10 ROIs selected by the CCA-FGL, SCCA, and CCA algorithms, respectively. B, D, and F are the ROC curves of the diagnostic models constructed by the top 10 genes selected by the CCA-FGL, SCCA, and CCA algorithms, respectively $\beta_1 = 0.01$, $\beta_2 = 0.01$, $\lambda_1 = 0.01$, $\lambda_2 = 0.1$ and $\mu = 1$.