Machine Intelligence Research
www.mi-research.net

20(1), February 2023, 57-78
DOI: 10.1007/s11633-022-1361-0

Machine Learning for Brain Imaging Genomics Methods: A Review

Mei-Ling Wang[1,2] Wei Shao[1,2] Xiao-Ke Hao[3] Dao-Qiang Zhang[1,2]

[1] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

[2] Key Laboratory of Pattern Analysis and Machine Intelligence, Ministry of Industry and Information Technology, Nanjing 211106, China

[3] School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China

Abstract: In the past decade, multimodal neuroimaging and genomic techniques have been increasingly developed. As an interdisciplinary topic, brain imaging genomics is devoted to evaluating and characterizing genetic variants in individuals that influence phenotypic measures derived from structural and functional brain imaging. This technique is capable of revealing the complex mechanisms by macroscopic intermediates from the genetic level to cognition and psychiatric disorders in humans. It is well known that machine learning is a powerful tool in the data-driven association studies, which can fully utilize priori knowledge (intercorrelated structure information among imaging and genetic data) for association modelling. In addition, the association study is able to find the association between risk genes and brain structure or function so that a better mechanistic understanding of behaviors or disordered brain functions is explored. In this paper, the related background and fundamental work in imaging genomics are first reviewed. Then, we show the univariate learning approaches for association analysis, summarize the main idea and modelling in genetic-imaging association studies based on multivariate machine learning, and present methods for joint association analysis and outcome prediction. Finally, this paper discusses some prospects for future work.

Keywords: Brain imaging genomics, machine learning, multivariate analysis, association analysis, outcome prediction.

Citation: M. L. Wang, W. Shao, X. K. Hao, D. Q. Zhang. Machine learning for brain imaging genomics methods: A review. *Machine Intelligence Research*, vol.20, no.1, pp.57–78, 2023. http://doi.org/10.1007/s11633-022-1361-0

1 Introduction

In recent years, with the development of cognitive neuroscience, neuroimaging has brought new vitality to the study of the working mechanism of the human brain. At the same time, with the development of noninvasive brain imaging technology, researchers hope to gain new insights into the imaging characteristics and molecular mechanisms of the brain, as well as their impact on normal and disordered brain function and behavior. Commonly used brain imaging techniques include structural magnetic resonance imaging (sMRI), functional magnetic resonance imaging (fMRI), diffusion tensor imaging (DTI), and positron emission tomography imaging (PET). In addition, with the development of genetic technology, researchers can identify genetic markers associated with neurological and psychiatric diseases from a more refined molecular level (such as single nucleotide polymorphisms (SNPs)).

With recent technological advances in acquiring multimodal brain imaging data and high-throughput genomics data, brain imaging genomics is emerging as a rapidly growing research field. Hariri and Weinberger[1] proposed the concept of imaging genomics or imaging genetics, which performs integrative studies that analyse genetic variations, such as SNPs, as well as epigenetic and copy number variations (CNVs), molecular features captured by various omics data, and brain imaging quantitative traits (QTs), coupled with other biomarker, clinical, and environmental data[2, 3].

As an emerging data science, brain imaging genomics has achieved rapid growth, which is greatly attributed to the public availability of valuable imaging and genomics datasets. Due to the open-science nature of the Alzheimer's Disease Neuroimaging Initiative (ADNI) project[4], hundreds of publications using ADNI imaging genomics data have been produced in the past decade, yielding innovative machine learning methods and novel biomedical discoveries. Similar to the ADNI, an increasing number of landmark studies are producing big data, including multidimensional imaging and omics modalities, making them available to the research community. These include the Enhancing Neuro Imaging Genetics through Meta Analysis (ENIGMA) Consortium[5], Philadelphia Neurodevel-

Review
Manuscript received April 16, 2022; accepted August 1, 2022
Recommended by Associate Editor Jun-Zhou Huang

opmental Cohort (PNC)[6] and Parkinson's Progression Markers Initiative (PPMI)[7].

Brain imaging genomics mainly uses brain imaging technology to evaluate the genetic influence on individuals by using brain structure and function as phenotypes, and explores how genes affect the neural structure and function of the brain, as well as the resulting neurological pathology. Studying the association between genetics and brain structure and function, and building a visible bridge between "genes and brain", can better reveal the pathogenesis of neuropsychiatric diseases[8–10]. Imaging genomics can also identify biological indicators or endophenotypes of a brain disease, which provides a more accurate method for predicting and diagnosing the disease. Specifically, most researchers consider SNPs as genotype data for association analysis. In the acquisition of endophenotypic data, researchers mostly use brain imaging data (i.e., MRI) in clinic for analysis. For example, sMRI, an imaging technique that measures the structural organization of the brain, can quantify abnormalities in morphology (i.e., gray matter volume). fMRI scans have been shown to be effective in revealing functional connectivity patterns of the brain. Based on different modalities of brain imaging technology, at present, imaging genomics mainly focuses on the association analysis between gene SNPs and brain structure, function, and connectivity[11–14].

Early imaging genomics approaches consisted of univariate paired statistical analysis methods, where multiple tests are employed to find the association between SNPs or genes and complex diseases or measurable quantitative traits (QTs). Genome-wide association study (GWAS) uses the whole genome high-throughput sequencing technology to classify the sequence variation in the genome of the research object, and finally selects significant SNPs via the biostatistics methods and bioinformatics methods[15]. Since the first GWAS research paper on age-related macular degeneration published in Science in 2005[16], this method has been used in the analysis of psychiatric disorders[17]. GWAS has played a great role in the study of imaging genetics, but there are also some problems, such as strict multiple correction, so that many small effect variants cannot pass the correction level. In addition, GWAS can only obtain a single degree of association between genetic variation and traits, and cannot well explain the complex molecular mechanisms of the brain.

In recent years, with the rapid development of machine learning in academia and industry, researchers have tried to use these data analysis tools to solve some problems in many fields. In the association analysis of imaging genetics, in addition to univariate statistical analysis, the multivariate machine learning model is the most widely used, and it has identified disease-sensitive imaging and genetic biomarkers. Internationally, some scholars have also written a review of related methods in imaging genetics. For example, Medland et al.[18] have raised the problems and challenges of using traditional univariate statistical models to process large-scale genome-wide brain imaging association analysis, reviewing the research results in different central databases. Liu and Calhoun[19] summarized the application of other multivariate methods such as independent component analysis in imaging genetics. Thompson et al.[20] focused on the association analysis between genetics and brain structure connectivity and functional networks. Based on the above review works, this article is devoted to providing comprehensive and up-to-date coverage of machine learning methods in brain imaging genomics. Fig. 1 is adopted to present a schematic of the topics covered in brain imaging genomics. One of the main goals of imaging genomics based on machine learning is to realize association analysis studies for understanding mechanisms and pathways. We group these imaging genomics based on machine learning methods into two categories. The first category mainly uses regression models to identify complex multi-SNP and/or multi-QT associations. Most of the regression models discussed earlier can be described using the regularized loss function framework. A sparsity-inducing regularization term is often included in these models. The motivations are twofold. First, it is reasonable to hypothesize that only a small number of markers are relevant in the resulting imaging genomics association. The sparsity term can help identify these relevant markers. Second, the sparsity constraint can reduce the model complexity and subsequently reduce the risk of overfitting. In addition to regression models, another category of prominent methods developed for brain imaging genomics studies are correlation models, such as sparse canonical correlation analysis (SCCA)[21–23] and parallel-independent component analysis (pICA)[24, 25]. Similar to the regression model discussed earlier, the sparsity is encouraged in these correlation models to reduce model complexity and the risk of overfitting, as well as identify relevant biomarkers. Overall, this article is focused on the three types of learning problems as follows. First, we will show the limitations of the univariate imaging genetics association analysis and show the univariate learning approaches for correlation analysis. Second, we will present the problem of multivariate imaging genetics association analysis and summarize the main idea and modelling in genetic-imaging association studies based on multivariate machine learning. Third, we will review methods that are used to predict an outcome of interest by combining both imaging and genomics data, and methods for joint association analysis and outcome prediction. Finally, some unsolved problems in genetic imaging and future research directions are prospected.

2 Univariate analysis method

The statistical analysis of single-genetic variables usu-

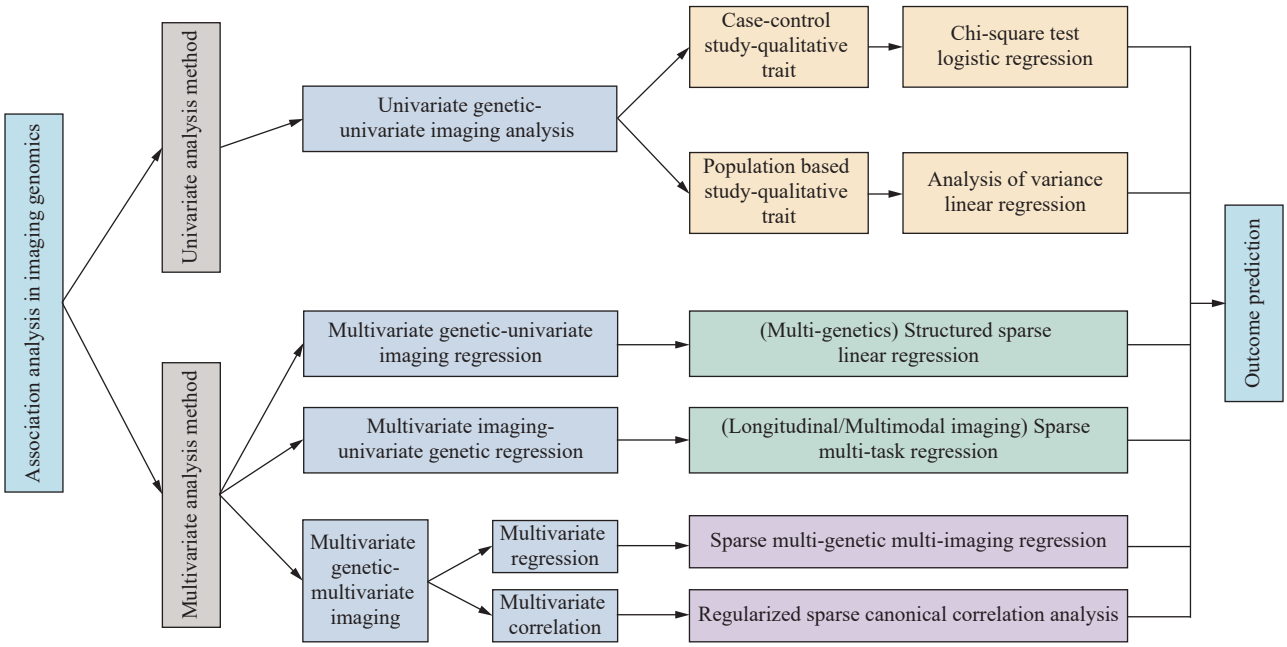

Fig. 1 Schematic of topics covered in brain imaging genomics. The goal is to present association analysis in imaging genetics based on machine learning.

ally adopts the Pearson′s chi-square test for the experimental group and the control group as the allele detection method. That is, to confirm whether the locus is associated with a genetic risk factor by analysing whether there are statistical differences between the corresponding genomic loci of a group of patients with various diseases and a group of normal controls. Imaging genetics analysis based on univariate statistical methods can use linear regression and analysis of variance models as allele association analysis methods[26]. In addition, for the multiple univariate models, firstly, $p \times q$ linear regression models ($y_j = \beta_{jk} x_k$, where p is the gene feature dimension and q is the imaging feature dimension) are fitted. Then, $p \times q$ null hypotheses ($H_0 : \beta_{jk} = 0$) are tested. Finally, the p-values are sorted to select the smaller p-values. For example, in 2009, Potkin et al.[27] performed a genome-wide association study (GWAS) on patients, normal controls, and imaging phenotypes. That is, the effect of SNPs on quantitative phenotypes of brain areas can be calculated by a generalized linear model, which is constructed by imaging phenotypes, disease diagnosis and gene data. The expression is as follows:

$$Y = b_0 + b_1 SNP + b_2 APOEe4 + b_3 gender + b_4 age + b_5 diagnosis + b_6 SNP \times diagnosis + \epsilon \quad (1)$$

where Y denotes the neuroimaging QT, b_i represents the coefficient of each variable, and $SNP \times diagnosis$ represents the interaction relationship. The p-value obtained is the association result between SNP and QT[27].

In the univariate imaging genetics association analysis, according to different scales[28], we summarize as follows: for the genetic level, it includes 1) candidate genetics/SNPs[29–32], 2) related biological functions characteristic pathways/networks[33–35], 3) whole genome[27, 36–39]. For the brain imaging level, it includes 1) individual regions of interest (ROI)[27, 29, 33, 36], 2) multiple ROI[30, 34, 37], 3) whole brain[31, 32, 35, 38, 39]. Whether it is the association analysis between candidate genetic locus SNP and neuroimaging[40] (cerebrospinal fluid[41], cognitive score[42], and any other QT), or the association analysis between whole genome and neuroimaging or even the association analysis between whole genome and smaller voxel-wise brain imaging, linear regression and analysis of variance can solve the problems of imaging genetics association analysis at different scales. In addition, some researchers have released relevant statistical analysis software, such as Plink[43].

GWAS genetic statistical analysis needs to find the association with disease phenotypes from millions or even tens of millions of SNPs. Although the Bonferroni correction can be used to strictly control the significance[44, 45], this strategy will lead to many small effect variations that cannot pass the correction level, and multiple such small effect variations may act together to have a great impact on the traits. The application of univariate analysis methods in imaging genetics has a more intuitive explanation, and can simply and quickly detect the association between a single SNP and a single QT. However, due to the high-dimensional characteristics of data variables, a large number of multiple comparisons eventually make the statistical test results not significant, and the above test method is based on a strict hypothesis. That is, genetic loci or imaging characteristic variables are statistic-

ally independent, while the important information of the association between variables is ignored. Therefore, for high-dimensional features, the univariate approach still has some limitations in dealing with the problem of imaging genetics association analysis.

3 Multivariate analysis method

Following the univariate voxel-wise genome-wide association analysis (vGWAS)[39], Hibar et al.[46, 47] proposed a multivariate voxel-wise gene-wide association study (vGeneWAS), which solves the problem of variable collinearity by principal components regression (PCReg) to all SNPs in a genome. Specifically, principal component analysis (PCA) was first used to obtain the mutually orthogonal factors that maximize the variance on the SNP regression variable set. Then, the standard partial F-test was used on these orthogonal factors. Finally, following the related work proposed by Stein et al.[39] in 2010, the same genetic and brain imaging dataset were used to group SNPs and detect the association between grouped SNPs with voxel-wise imaging. Experimental results show that this method achieves better association performance and reduces the number of statistical tests. Therefore, in order to enhance the ability to detect the association between genetics and quantitative traits (QTs), some researchers have used multivariate methods to address the association of multi-genetic or multi-locus combined effects in imaging genetics[19, 48]. Recently, research on machine learning based imaging genetics has attracted much attention, which aims to identify the association between genetics and imaging features by using regression models. We can use different criteria to divide these methods into regression models (including multivariate genetic-univariate imaging regression, multivariate imaging-univariate genetic regression, and multivariate genetic-multivariate imaging regression) and correlation models (i.e., multivariate genetic-multivariate imaging correlation). In the next subsection, several classic and state-of-the-art association models will be introduced by the above division strategy.

3.1 Regression models

3.1.1 Multivariate genetic-univariate imaging regression

We usually use a sparse regularized regression model to realize multivariate genetic-univariate imaging regression. The main motivation is twofold. First, assuming that only a few markers are associated with imaging genomics, sparse terms assist to identify these related markers. Second, sparse constraints can reduce the complexity of the model and the risk of overfitting. In [49, 50], regression models based on L1 norm penalty constraints have been successfully applied to multivariate genetic data analysis. They aim to identify sparse SNP loci that are highly associated with specific brain regions. These models provide a general technical framework to deal with the small sample regression problem of detecting and identifying high-dimensional genetic SNPs. However, the constraints based on the L1 norm do not fully consider the structural relationship between the feature variables, therefore the optimal regression results cannot be achieved in theory. Considering the spatial structure relationship between SNP features, Silver et al.[51–53] proposed the group sparse model or fusion sparse model to select SNP loci in the same group or adjacent feature variables, and the models based on group sparsity or fusion sparse are as follows:

$$\min_{w} \|y - Xw\|_2^2 + \lambda \sum_{i=1}^{g} \sqrt{\sum_{j \in G(i)} w_j^2} \tag{2}$$

$$\min_{w} \|y - Xw\|_2^2 + \lambda \sum_{i<j} \mid w_i - w_j \mid \tag{3}$$

where these two equations are utilized for identifying a set of SNPs from X and predicting a single imaging phenotype y. In (2), w_j in the group sparsity term represents all the SNP loci features belonging to the group $G(i)$, and the goal is to control the selected loci to include the characteristics of clustering. For example, there will be a linkage disequilibrium (LD) effect[54] between gene loci, that is, SNPs linked on different genes will appear in the same LD block nonrandomly. This provides domain knowledge for the feature selection model based on group sparsity so that SNPs in the same LD group can be detected simultaneously. In (3), the fusion Lasso term can control the weight contribution of adjacent position features w_i and w_j to be as similar as possible, that is, the feature variables selected by the fusion Lasso term have spatial continuity. The empirical study was performed on an ADNI sample.

In addition, there is not only a flat spatial relationship between SNP loci, but also a hierarchical relationship in the actual gene structure. For example, in a certain pathway, the interaction of specific gene loci can affect protein synthesis and functional transformation, and some SNP loci under the same gene also have certain correlations (such as LD). Therefore, making full use of the prior knowledge of this hierarchical structure to perform the imaging genetic analysis will often reduce the error in the regression analysis and learn more explanatory feature patterns[55–57], as shown in Fig. 2. As shown, the model uses a tree-guided sparse learning (TGSL) method to identify the association between genotype and phenotype. When constructing a tree structure, the SNP loci are used as leaf nodes, the LD block and the gene block are used as intermediate nodes, and all genes in the pathway are used as the final root nodes. The structure tree has d layers and each layer has n_i nodes. The node of the

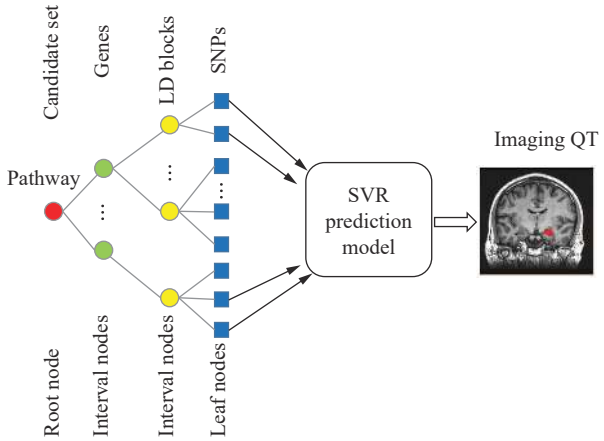

Fig. 2 Tree-guided sparse regression model, which aims to identify a set of SNPs for predicting a single imaging phenotype[55–57].

i-th layer is $\{G_1^i, \cdots, G_j^i, \cdots, G_{n_i}^i\}$, and the tree-guided sparse regression model is as follows:

$$\min_w \|y - Xw\|_2^2 + \lambda \sum_{i=1}^{d} \sum_{j=1}^{n_i} \alpha_j^i \|w_{G_j^i}\|_2 \tag{4}$$

which also aims to identify a set of SNPs for predicting a single imaging phenotype y. α_j^i is the weight of any node G_j^i predefined according to prior knowledge. $w_{G_j^i}$ is the weight of any node G_j^i in the learned tree structure. It is worth noting that when the weight of a node is zero, its child nodes are all zero, that is, all the features of the subtree have nothing to do with the regression task and are not selected. Compared with the traditional Lasso method, the SNPs obtained by the optimization of the model have smaller errors in predicting the gray matter volume of the brain, and these SNP loci associated with MRI brain regions have a hierarchical clustering. The empirical study was performed on an ADNI sample to identify sparse SNP patterns at the block level to better guide the biological interpretation.

3.1.2 Multivariate imaging-univariate genetic regression

In research on machine learning based imaging genetics, most of the works have focused on discovering and detecting multivariate SNP loci associated with imaging phenotypes. However, few studies have explored how SNP values change when phenotypic measurement variables change, that is, using multivariate imaging to regress univariate genetic features. For example, Shen et al.[38] proposed a task-related time series multivariate sparse regression model based on the group structure information between prediction variables. The model is as follows:

$$\min_W \sum_t^T \|Y - X_t^{\mathrm{T}} W_t\|_F^2 + \lambda_1 \sum_{k=1}^{d} \sqrt{\sum_{t=1}^{T} \|w_t^k\|_2^2} + \lambda_2 \mathrm{Tr}(WW^{\mathrm{T}})^{\frac{1}{2}} \tag{5}$$

where brain imaging data is $X = \{X_1, X_2, \cdots, X_T\} \in \mathbf{R}^{d \times n \times T}$, SNP data is $Y = \{y_1, y_2, \cdots, y_n\} \in \mathbf{R}^{n \times c}$, and w_t^k represents the k-th row of the coefficient matrix W_t at time t. $\mathrm{Tr}(\cdot)$ is the trace of the matrix. The tensor coefficient matrix is $W = \{W_1, W_2, \cdots, W_T\} \in \mathbf{R}^{d \times c \times T}$. The tensor coefficient matrix reveals the time series of brain imaging QTs. When $c = 1$, the algorithm is the association between time series imaging MRI and risk genes, as shown in Fig. 3. It can be found, through the joint constraints of feature weights in multiple regression tasks and multiple time points, task-related longitudinal imaging phenotype markers can be selected. This model performs a new perspective from phenotype to genotype analysis to study the impact of individual genes on changes in brain structure and function. The empirical study was performed on an ADNI sample.

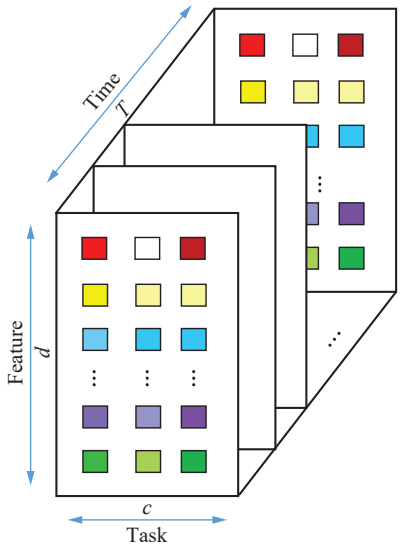

Fig. 3 Task-correlated longitudinal sparse regression model, which aims to study the impact of individual genes on changes in brain structure and function[38]

In addition, in univariate genetic-multivariate imaging association analysis, most studies focused on the single modal imaging phenotype QT. In order to study the association between genetics and multimodal brain imaging QTs, Hao et al.[58, 59] realized the association analysis between multimodal imaging QTs Y and candidate risk gene loci x by introducing the group sparse regularization term to construct the diagnosis-guided multimodal (DGMM) regression model (see Fig. 4) as

$$\min_W \sum_m^M \|x - Y_m w_m\|_2^2 + \lambda_1 \sum_i \sqrt{\sum_j w_{ij}^2} + \lambda_2 R(w) \tag{6}$$

where $W = [w_{VBM}, w_{FDG}, w_{AV45}] \in \mathbf{R}^{q \times 3}$ $(j = \{1, 2, 3\}, i = \{1, \cdots, q\})$ is the association weight matrix of multimodal imaging QTs (VBM, FDG, AV45, where VBM is voxel-based morphometry obtained by prepro-

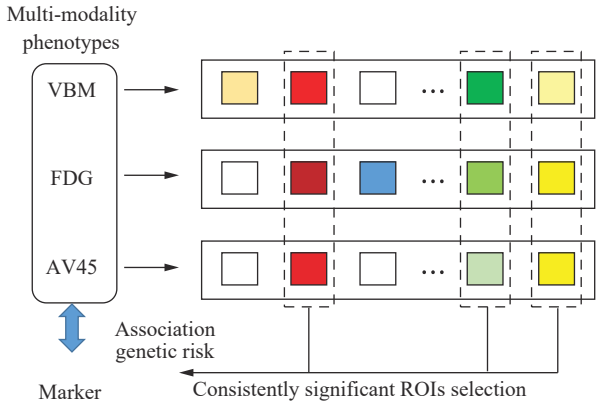

Fig. 4 Multi-modality association model[58, 59]. The goal of the regression model is to realize the association analysis between multimodal imaging QTs and candidate risk gene loci.

cessing structural magnetic resonance imaging data. FDG is the fluorodeoxyglucose positron emission tomography. AV45 is the F-18 florbetapir PET scans amyloid imaging.) and candidate risk gene APOE e4. The second term is the group sparse regularization term. $R(w)$ is the Laplacian regularization term. This method realizes the feature selection of multimodal imaging biomarkers associated with risk genes through a generalized linear regression function. The multimodality association method can identify robust consistent brain regions and has strong antinoise ability compared with the single-modality association method. Therefore, this method can be applied to the association analysis between other risk genes and multimodal imaging QTs. The empirical study was performed on an ADNI sample, where the response is the APOE e4 SNP and the predictors include three modalities of ROI measures: 1) VBM measure from structural MRI, 2) FDG measure from PET, and 3) AV45 measure from PET. For example, Wang et al.[60] presented a diagnosis-aligned multimodal (DAMM) strategy for the regression of a candidate risk gene APOE e4 x on multimodal imaging QTs Y_m (VBM measure from structural MRI and hypergraph-based clustering coefficient measure from fMRI) as follows:

$$\min_{W} \sum_{m}^{M} \|x - Y_m w_m\|_2^2 + \lambda_1 R_1(w) + \lambda_2 R_2(w) \qquad (7)$$

where $R_1(w)$ is the same group sparse regularization term as (6) so as to jointly select a few features associated with risk SNP loci in multimodal brain imaging QTs. $R_2(w)$ is the graph Laplacian regularization term to fully use the information between different modalities. The model uses structural voxel information and network connection information as an intermediate feature of bridging risk gene locus and disease status to find the disease-specific associations between the risk SNPs and the brain network. It is worth noting that this work is an initial attempt to explore the relationship between connectivity traits and genetic variation. The empirical study was performed on an ADNI sample, where the response is the APOE e4 SNP and the predictors include two modalities of ROI measures: 1) VBM measure from structural MRI and 2) hypergraph-based clustering coefficient measure from fMRI.

3.1.3 Multivariate genetic-multivariate imaging regression

The above multivariate analysis of SNP loci is only aimed at the regression of imaging univariate features, and can not make full use of the correlation among imaging multivariate features. In recent years, methods have been introduced for the analysis of high dimensional genetic and imaging data able to cope with multivariate genetic input and multivariate imaging output. Some researches have used sparse regression[61] to discover a low-dimensional subset of genetic data significantly associated with the imaging QT in the original high-dimensional data[48]. For example, Vounou et al.[62] proposed a sparse reduced-rank regression (SRRR) model as follows:

$$\min_{A,B} \|Y - XBA^{\mathrm{T}}\|_F^2 + \lambda_1 \|A\|_1 + \lambda_2 \|B\|_1 \qquad (8)$$

where $W = BA^{\mathrm{T}}$ and W denote the product of a matrix $B \in p \times r$ and a matrix $A \in q \times r$. The major goal of the model is to find the minimization of the rank of W. The L1 norm is imposed on A and B to sparsely select features. Fig. 5 presents a visual framework of the sparse reduced rank regression model. The empirical study was performed on an ADNI sample.

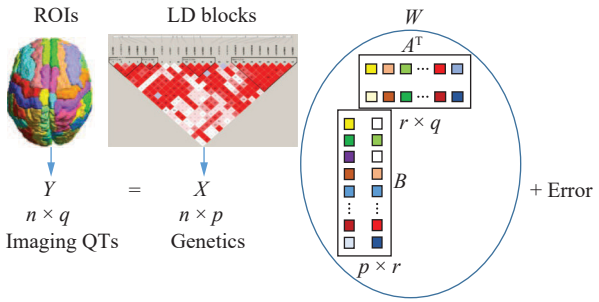

Fig. 5 Sparse reduced rank regression model. The goal is to identify a set of SNPs from X to predict a set of Alzheimer's disease (AD)-related imaging QT Y[62].

After that, Vounou et al.[62] improved their previous work. Specifically, first, the discriminant analysis method was used in the structural brain imaging QTs to find the discriminant multivariate biomarker as the phenotype, and the existing SRRR model was then used for genome-wide association analysis, which finally achieved good results[63]. Subsequently, Silver et al.[64] presented a pathways sparse reduced-rank regression (P-SRRR) method, which integrates the pathways group lasso with adaptive weights (P-GLAW) idea into the SRRR method, i.e.,

$$\min_{A,B} \|Y - XBA^{\mathrm{T}}\|_F^2 + \lambda \sum_{g \in G} d_g \|B_g\| \tag{9}$$

where G is defined as the grouping structure of B. This model identifies a set of SNPs from X, so that a set of AD-related imaging QT Y can be predicted. The pathway knowledge is utilized to group the SNPs for the sake of selecting features at the pathway level. The empirical study was performed on an ADNI sample.

In addition, some work considers multivariate multi-task regression models[52]. For example, Wang et al.[65] proposed a group-sparse multi-task regression and feature selection (G-SMuRFS) strategy (see Fig. 6):

$$\min_{W} \|Y - XW\|_F^2 + \lambda_1 \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j} w_{ij}^2} + \lambda_2 \sum_{i} \sqrt{\sum_{j} w_{ij}^2} \tag{10}$$

where the second term considers the linkage disequilibrium (LD) structural relationship between SNP loci, embedding the prior information of the grouping relationship of SNPs, so that SNPs in the same LD group are detected simultaneously. The third term uses the same L21 norm in (6) as a regularization term, which is also used to jointly select a few features associated with risk SNP loci in multimodal brain imaging QTs. The empirical study was performed on an ADNI sample, where 1 224 SNPs from 37 AD genes were used to predict ten VBM measures and SNPs were grouped by LD blocks.

3.1.4 Discussions

In addition to regularized multivariate regression models in brain imaging genomics, many Bayesian algorithms have been presented. For instance, motivated by G-SMuRFS[65], a Bayesian group sparse multitask regression (BGSMTR) model was constructed to identify multivariate genetic-multivariate imaging regression associations, and the group structure (e.g., LD blocks and genes) within the SNP data was simultaneously embraced. Compared with G-SMuRFS, which only provides a point estimate of the regression coefficients, the BGS-MTR follows full posterior inference such as interval estimates for the regression parameters. This model can be viewed as an expansion of the Bayesian group lasso[66, 67] for accommodating multivariate responses and variable selection at the SNP and gene levels. In [68], a Bayesian generalized low-rank regression (GLRR) model was built to analyze high-dimensional imaging responses and covariates, which uses a low-rank representation for approximating the high-dimensional weight matrix. This GLRR model was further extended into a Bayesian longitudinal low-rank regression (L2R2) form [69] to examine genetic effects on longitudinal imaging responses.

The methods introduced in the above three subsections are collectively referred to as multivariate regression models. We only introduce some typical examples. Table 1 summarizes the multivariate regression methods used in the studies discussed in recent years. Such approaches are devoted to revealing complex imaging genomics associations between multivariate SNP data and imaging QT data. They share a common rationale: these methods all utilize a regularized regression model to identify the association between SNPs and imaging QTs. It should be noted that two common advantages are included in these models: 1) the regression coefficients directly capture the association between SNPs and imaging QTs, which is easy to interpret; 2) the genetic markers and imaging markers obtained by using a single model do not require multiple test corrections, increasing detection power. However, due to the high dimensionality of the data, there is an increased risk of overfitting these models. In order to remedy such deficiency, various regularization forms are added to reduce model complexity and biologically meaningful structures are introduced to decrease the risk of overfitting. For example, sparsity can simplify complexity models (i.e., G-SMuRFS and SRRR) by using the L1 or L21 norm. Biologically meaningful structures (i.e., LD and pathways) could be achieved via adopting group lasso or group L21 norm (i.e., P-GLAW and P-SRRR). Additionally, the low-rank constraint can be used as a regularization term (i.e., task-correlated longitudinal sparse regression (TCLSR) model, temporal structure autolearning (TSAL) model, and joint projec-

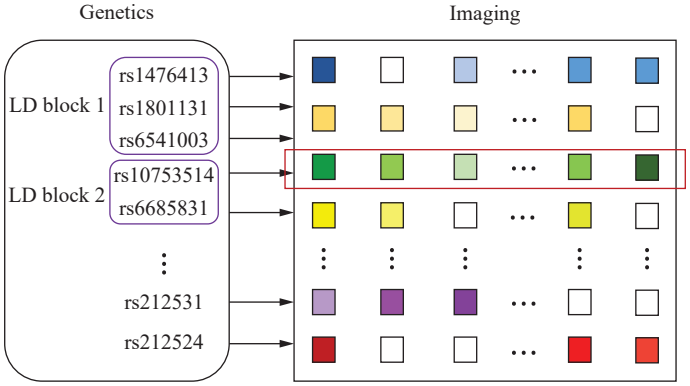

Fig. 6 Group-sparse multi-task regression and feature selection strategy, which is a structured sparse model[65]

Table 1 Example studies using multivariate regression, which aim to reveal complex imaging genomics associations between multivariate SNP data and imaging QT Data

Related studies	Method category	Dataset	Year	Publication
P-GLAW	Multivariate genetic-univariate imaging regression	ADNI	2012	*Statal Applications in Genetics and Molecular Biology*
TGSL[56, 57]	Multivariate genetic-univariate imaging regression	ADNI	2019	*IEEE/ACM Transactions on Computational Biology and Bioinformatics*
TCLSR[70]	Multivariate imaging-univariate genetic	ADNI	2012	*Bioinformatics*
DGMM[58, 59]	Multivariate imaging-univariate genetic	ADNI	2016	*Neuroinformatics*
DAMM[60]	Multivariate imaging-univariate genetic	ADNI	2019	*Bioinformatics*
SRRR[62]	Multivariate genetic-multivariate imaging regression	ADNI	2010	*NeuroImage*
P-SRRR[64]	Multivariate genetic-multivariate imaging regression	ADNI	2012	*NeuroImage*
G-SMuRFS[65]	Multivariate genetic-multivariate imaging regression	ADNI	2012	*Bioinformatics*
TSAL[71]	Multivariate imaging-univariate genetic	ADNI	2018	*Journal of Computational Biology*
JPLSR[72]	Multivariate genetic-multivariate imaging regression	ADNI	2019	*IEEE Transactions on Biomedical Engineering*
S-SRRR[73]	Multivariate genetic-multivariate imaging regression	ADNI	2016	*Medical Image Computing and Computer-Assisted Intervention (MICCAI)*
GRS-SRRR[74]	Multivariate genetic-multivariate imaging regression	ADNI	2017	*IEEE Transactions on Big Data*
RGRS-SRRR[75]	Multivariate genetic-multivariate imaging regression	ADNI	2018	*Neuroinformatics*

tion learning and sparse regression (JPLSR) model), which has a strong ability to handle spatial or temporal correlations and decrease model complexity.

3.2 Correlation models

3.2.1 Multivariate genetic-multivariate imaging correlation

In imaging genomics research, multivariate regression models have been able to solve the problem of feature selection. For multi-output variable features, the multi-task regression model[65, 76] can consider the covariance structure relationship among multiple regression output variables. However, the high-dimensional regression output will generate high computational time costs and the multivariate output structure is complex. The model that simply considers the group constraints of multiple regression tasks for association analysis is often too strict. In order to fully consider the covariance structure between two variables, Liu et al.[77, 78] proposed using the parallel independent component analysis (PICA) method to analyze the association mechanism between genetic and imaging data, so as to find the most relevant independent component of the two modal data. However, this method does not restore the contributing SNPs and important brain regions, resulting in the loss of reasonable biomarker interpretation of these components. Another bimultivariate model, such as canonical correlation analysis (CCA)[79, 80] or partial least squares regression least squares region (PLS)[81, 82] can find the linear combination of two group variables respectively, so that the correlation or covariance between genetics and imaging data is the largest. This model can better solve the problem of multivariate genetic-multivariate imaging association analysis compared with the regression model. However, in high-dimensional data, feature variables often have noise and redundancy, that is, not all SNP and QT characteristic variables are associated. Therefore, in order to select a small number of relevant genetics and imaging features with explanatory significance, the sparsity is introduced into the classical bimultivariate correlation analysis, namely SCCA[21–23] and sparse partial least squares regression (SPLS)[83, 84]. Similar to the mentioned-above regression methods, the regularization is also encouraged in these correlation models. On the one hand, as is known, only a small amount of markers are relevant in the imaging genomics association, which can be effectively identified by the regularization terms. On the other hand, the regularization is capable of reducing the model complexity, so that over-fitting is avoided. Here, denote $X \in \mathbf{R}^{n\times p}$ as the genetic data with p variables on n subjects, and $Y \in \mathbf{R}^{n\times q}$ as the imaging data with q variables on n subjects, where all columns of X and Y have been normalized with zero mean and unit variance. As the most popular bimultivariate correlation models for brain imaging genomics, the SCCA and its expansion with various regularizers can be expressed by

$$\begin{aligned} &\max_{u,v} u^{\mathrm{T}} X^{\mathrm{T}} Y v - \sum_{i=1}^{k} \lambda_i R_i(u,v) \\ &\quad \text{s.t.} \ \ \|Xu\|_2^2 = \|Yv\|_2^2 = 1. \end{aligned} \tag{11}$$

The goal is to find a linear combination of the SNPs

Xu and a combination of the imaging QTs Yv for the sake of maximizing the correlation (i.e., $u^{\mathrm{T}}X^{\mathrm{T}}Yv$ s.t. $\|Xu\|_2^2 = \|Yv\|_2^2 = 1$) on the condition of one or more regularization forms $R_i(u, v)$. For instance, the traditional SCCA model introduces $R_1(u) = \|u\|_1$ and $R_2(v) = \|v\|_1$. There are other regularizers, such as incorporating a group/network structure or other prior knowledge in brain imaging genomics data, to complete different tasks. In what follows, several classic and state-of-the-art studies using these regularized SCCA strategies will be introduced, which are widely utilized to identify complex multivariate genetic-multivariate imaging correlations. It can be divided into prior knowledge-induced SCCA models, and sample correlation-induced SCCA models according to their distinct regularization terms.

By introducing the sparsity into multivariate genetic-multivariate imaging association analysis, the model can automatically select relevant sparse SNP and QT feature variables from high-dimensional bimultivariate. However, a major problem of SCCA is that the model still does not fully consider the structural relationship between characteristic variables, that is, a lot of prior information is not used in the establishment of the model. For example, SNP loci in the same LD block may have some common characteristics, and the brain needs multiple brain regions to work together to complete a certain function. Therefore, in the research of multivariate genetic-multivariate imaging association, in order to make up for the shortcomings of traditional SCCA, many scholars have expanded and improved the SCCA model by using various priori information as regularization terms[85–89]. For example, a structure-aware SCCA (S2SCCA) model[89] was built via adding the following two group L1 norm terms into (11) i.e.,

$$R_1(u) = \sum_{g \in G_1} \|u_g\|_2 \tag{12}$$

$$R_2(v) = \sum_{g \in G_2} \|v_g\|_2 \tag{13}$$

where the LD blocks are employed for the construction of the SNP grouping structure G_1 in (12). In (13), the ROIs are utilized for forming the voxelwise imaging QT grouping structure G_2. The goal is to identify multivariate genetic-multivariate imaging associations between APOE SNPs and the voxelwise QTs by using the prior knowledge as regularization terms. The empirical study was performed on an ADNI sample to identify multi-SNP-multi-QT associations between the voxelwise QTs and APOE SNPs.

In addition, Yan et al.[85] presented a knowledge-guided SCCA (KG-SCCA), as shown in Fig. 7. This model uses group sparse regularization constraints to embed the LD block grouping prior information SNP loci into

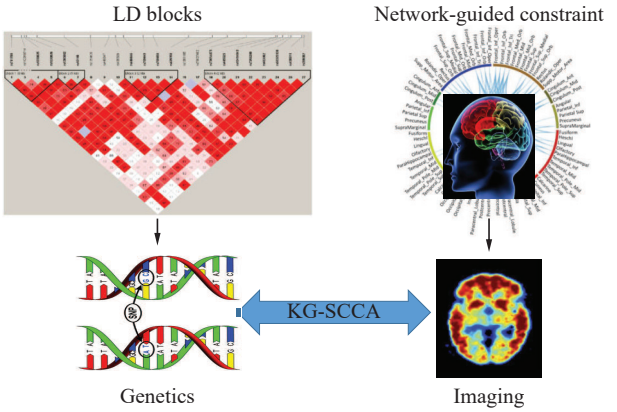

Fig. 7 Schematic of KG-SCCA, which aims to identify multivariate genetic-multivariate imaging associations between APOE SNPs and the voxelwise QTs[85]

(11) and the expression is as follows:

$$R_1(u) = \sum_{i=1}^{g} \sqrt{\sum_{j} \in G(i) u_j^2} \leq c_3 \tag{14}$$

where the feature variables in the L2 norm constraint group have the same weight contribution as much as possible, that is, SNP loci in the same LD block are more likely to be selected in association analysis. The L1 norm selects the few LD blocks with strong correlation by constraining the sparsity between groups. At the same time, the model also introduces the brain function network information as the prior knowledge of the feature similarity in brain region. In other words, when the connection weight in the brain network is high, the two brain region nodes have similar characteristics (gene expression is highly correlated), and the expression form of regularization constraints is as follows:

$$R_2(v) = \sum_{(i,j) \in E, i<j} \tau(w_{ij}) \times \|v_i - \mathrm{sgn}(w_{ij}) v_j\|_2 \leq c_4 \tag{15}$$

where v_i and v_j represent the feature weights of any two nodes on the brain network respectively. $\mathrm{sgn}(w_{ij})$ is the sgn of the correlation between v_i and v_j. When $\mathrm{sgn}(w_{ij})$ is positive, there is a positive correlation between v_i and v_j. When $\mathrm{sgn}(w_{ij})$ is negative, there is a negative correlation v_i and v_j. $\tau(w_{ij})$ is the connection intensity between v_i and v_j. The higher intensity of $\tau(w_{ij})$ indicates that the two brain region variables v_i and v_j tend to be selected simultaneously. An empirical study was performed on an ADNI sample to identify multi-SNP-multi-QT associations between amyloid imaging QTs and APOE SNPs.

Furthermore, some work considers bi-multivariate multi-task prior knowledge-induced SCCA models. For example, in [90, 91], a multitask SCCA (MTSCCA) was presented for identifying bimultivariate associations between SNP data and multimodal imaging data as follows:

$$\max_{u,v} \sum_{j=1}^{M} u_j^{\mathrm{T}} X^{\mathrm{T}} Y_j v_j - \lambda_1 \|U\|_{2,1} - \lambda_2 \|U\|_{G_{2,1}} - \lambda_3 \|V\|_{2,1}$$
$$\text{s.t.} \ \ \|Xu_j\|_2^2 = \|Y_j v_j\|_2^2 = 1 \tag{16}$$

where $X \in \mathbf{R}^{n\times p}$ is SNP data and $Y_j \in \mathbf{R}^{n\times q}$ $(j \in [1, M])$ is the imaging data of M modalities. $U = [u_1, u_2, \cdots, u_M]$ and $V = [v_1, v_2, \cdots, v_M]$. It is worth noting that the L21 norm regularization term is a "group-sparsity" regularizer, which forces only a small number of features to be selected from different modalities. The first regularization is a L21 norm to select SNP features. The second one is the group L21 norm that can select SNP features at the LD block level. The third regularization is a L21 norm, so as to select imaging features across all the modalities. A fast optimization algorithm has been implemented and applied to an ADNI sample to identify associations between over 150 000 SNPs from chromosome 19 and ROI-based QTs from three imaging modalities (VBM, FDG-PET, and Amyloid-PET).

In real biomedical studies, providing precise prior knowledge is a difficult task. Therefore, the above expanded forms of SCCA may be invalid once the biological priori knowledge is unavailable or incomplete. In general, the sample correlation was utilized rather than priori knowledges to define the graph or network constraint. Specifically, there are three types of regularizations used in (11): 1) L1 norm for flat sparsity, 2) group L1 norm for group sparisity, and 3) graph Laplacian-type norm to jointly select features connected in a graph. For example, in [92], a generic non-convex penalty based SCCA (GNC-SCCA) was designed as follows:

$$R(u) = \sum_{i=1}^{p} P_{\lambda,\gamma}(\mid u_i \mid) \tag{17}$$

where λ and γ represent nonnegative parameters, and $P_{\lambda,\gamma}(\mid u_i \mid)$ denotes a non-convex function. Seven non-convex penalties were added into the L1-based SCCA for sake of reducing the estimation bias. An empirical study was performed on an ADNI sample to identify multi-SNP-multi-QT associations between voxelwise QTs and 163 SNPs from AD genes.

Since the SCCA has the powerful ability to identify bi-multivariate relationships coupled with feature selection, it has become a popular tool in such field. The L0 norm is a sparsity-inducing tool, but it is a NP-hard problem. In practice, the L1 norm or its variants are usually introduced to replace the L0 norm for the sake of inducing sparsity. For instance, in [93], both truncated L1 norm penalized SCCA (TLP-SCCA) and truncated group lasso SCCA (TGL-SCCA) were presented, which respectively used truncated L1 norm and truncated group lasso below:

$$R(u) = \sum_{i=1}^{p} J_\tau(\mid u_i \mid), \ \ \text{where} \ \ J_\tau(u_i) = \min\left(\frac{\mid u_i \mid}{\tau}, 1\right) \tag{18}$$

$$R(u) = \sum_{k=1}^{K} J_\tau(\mid G_k \mid), \ \ \text{where} \ \ J_\tau(G_k) = \min\left(\frac{\mid G_k \mid}{\tau}, 1\right) \tag{19}$$

where τ denotes a tuning parameter. Selecting an appropriate value of τ, $R(u)$ can achieve a balance between the L0 norm and the L1 norm. It should be noted that, G_k denotes a subset of u at the k-th group $(k \in [1, K])$, and u represents the concatenation of all G_k. An empirical study was performed on an ADNI sample to identify multi-SNP-multi-QT associations between voxelwise QTs and 58 SNPs from AD-related genes, where QTs were grouped by ROI and SNPs were grouped by LD block.

In addition, inspired by GraphNet[94], Du et al.[86] proposed an absolute value-based GraphNet SCCA (AGN-SCCA), where an extended version of GraphNet regularization is added into the SCCA model. The forms of the AGN regularizations can be described into

$$R_1(u) = \mid u \mid^{\mathrm{T}} L_1 \mid u \mid + \beta_1 \|u\|_1 \tag{20}$$

$$R_2(v) = \mid v \mid^{\mathrm{T}} L_2 \mid v \mid + \beta_2 \|v\|_1 \tag{21}$$

where both L_1 and L_2 denote Laplacian matrices of the correlation matrices of X and Y. It should be noted that the data-driven correlation is here employed as a graph constraint, so that correlated features can be selected together. Additionally, by the added absolute value operation, both positively and negatively correlated features are allowed to be jointly selected. An empirical study was performed on an ADNI sample to identify multi-SNP-multi-QT associations between ROI-based imaging QTs and 58 SNPs from AD-related genes.

Furthermore, the sample correlation-induced SCCA models are devoted to the association between the SNP data and the imaging data of one modality at single time point. Fortunately, these models have been extended to focus on the longitudinal imaging data. For example, in order to identify genetic associations with longitudinal phenotypic markers, Hao et al.[95] designed a temporally constrained group SCCA (TGSCCA) framework, which is modelled as

$$\max_{u,V} \sum_{t=1}^{T} u^{\mathrm{T}} X^{\mathrm{T}} Y_t v_t - \lambda_1 \|u\|_1 - \lambda_2 \|V\|_{2,1} - \lambda_3 \sum_{t=1}^{T-1} \|v_{t+1} - v_t\|_1$$
$$\text{s.t.} \ \ \|Xu\|_2^2 = \|Y_t v_t\|_2^2 = 1 \tag{22}$$

where both u and v_t denote the weight vectors measuring

the contributions of the SNP loci and imaging phenotype ROIs at time-point t. v_{t+1} and v_t are weight vectors at adjacent time-points. λ_1, λ_2 and λ_3 are three regularization parameters. As can be found, the fused Lasso regularization is capable of constraining the gaps between two successive canonical weight vectors from adjacent time-points to be small, so that weight vectors become smooth to select neighboring features together. An empirical study was performed on an ADNI sample to identify associations between 85 APOE SNPs and longitudinal VBM QTs from 116 ROIs at four time points.

3.2.2 Discussions

Other bimultivariate correlation models are now discussed below. Fang et al.[96] designed a greedy projected distance correlation (G-PDC) strategy for the examination of pairwise gene-region of interest (ROI) associations, in which each gene and ROI contain a number of SNPs and voxels, respectively. Distance correlation is used to measure statistical dependence between two random vectors (e.g., gene versus ROI), which can model nonlinear relationships between them. Projected distance correlation is prone to measure conditional dependence based on distance correlation[97]. A gene-ROI pair is provided to test their independence and control all the other SNPs and voxels. Hao et al.[98] proposed an analytical strategy with three-way SCCA (T-SCCA) for exploring the intrinsic associations among genetic markers, imaging QTs, and clinical scores of interest. Hu et al.[99] designed a distance CCA (DCCA) algorithm by integrating distance correlation into the SCCA model. This algorithm identified a set of original SNPs and a set of original imaging QTs with the highest distance correlation, so as to reduce burden for multiple testing correction. An empirical study was performed on the PNC data to examine the pairwise association between 264 ROIs (containing 27 384 voxels) and 736 genes (containing 21 487 SNPs). Wang et al.[100] proposed a multi-modality discriminant SCCA method (MD-SCCA), where valuable discriminant similarity information is incorporated into the SCCA model to improve learning results. To be specific, the discriminant similarity information between within-class subjects was firstly obtained via the sparse representation. Then, a discriminant SCCA algorithm (D-SCCA) was constructed by enforcing the discriminant similarity information. Finally, the MD-SCCA method was employed to fully investigate the relationships among different modalities of different subjects. To consider the underlying complex multi-subspace structure of the original data, Wang et al.[101] utilized the self-expressiveness reflecting the similarity structure of the data for reconstructing the original input before the association analysis. Concretely, the within-class similarity information was firstly applied to the construction of self-expressive networks by sparse representation. The fusion method was then used to iteratively fuse the self-expressive networks from multi-modality brain phenotypes into one network. At last, a practical solution was provided for constructing and using the fused self-expressive network, so that the association between single modality phenotype and genotype could be mined by our method with L1 norm as well as the association between multi-modality phenotypes and genotype was explored by the form with the L21 norm. In addition, the deep learning technique has achieved great success in data-driven problems in biology and medicine. For instance, Wang et al.[102, 103] proposed a novel deep self-reconstruction sparse canonical correlation analysis (DS-SCCA) method for the identification of genetic associations with functional connectivity phenotypic markers. They focused on identifying the connectome, consisting of the brain region features and connectivity features, of functional brain networks derived from the fMRI data by realizing the relationships between genetic variants (i.e., the single nucleotide polymorphism, SNP) and brain networks (i.e., quantitative trait, QT). Furthermore, the main contribution of such work was an initial attempt to discover how genetic factors affect brain connectivity.

Table 2 summarizes bimultivariate correlation methods for the studies discussed earlier, which are devoted to identifying multivariate genetic-multivariate imaging associations from high-dimensional imaging genomic data. Similar to the regression models described earlier, the sparsity is also encouraged in these correlation models to reduce model complexity and the risk of overfitting, as well as identify relevant biomarkers. Most of these methods are based on regularized SCCA. In these SCCA models, the L1 or L21 norm is employed for feature selection, group L1 or L21 norm is used to select features at the group level, and the graph Laplacian is utilized for graph-guided learning. Note that the L21 norm is usually included in multimodal and longitudinal SCCA methods to select features across modalities or time points, and fussed lasso or fused pairwise L21 norm is often used to smooth neighboring weights along the temporal dimension. Here, we adopt these studies using these strategies to identify complex multi-SNPCmulti-QT associations. We will cover: 1) fundamental SCCA models (i.e., S2SCCA, KG-SCCA); 2) enhanced SCCA models (i.e., an SCCA framework using a generic nonconvex penalty (GNC-SCCA), TLP-SCCA, TGL-SCCA, absolute value-based GraphNet SCCA (AGN-SCCA), FDR-corrected SCCA); 3) multimodal and longitudinal SCCA models (i.e., MTSCCA, TGSCCA); 4) other bimultivariate correlation models (i.e., T-SCCA, MD-SCCA, FSN-SCCA, FSN-GSCCA, DS-SCCA).

4 Outcome prediction

In imaging genomics, how to integrate brain imaging and genomics data for the prediction of outcomes of interest, such as impairment score, disease stage, and progression status, is also an interesting topic. At present,

Table 2 Example studies using bimultivariate correlation methods, which aim to identify multivariate genetic-multivariate imaging correlation from high-dimensional imaging genomic data

Related studies	Method feature	Dataset	Year	Publication
S2CCA[89]	Structure aware SCCA	ADNI	2014	*MICCAI*
KG-SCCA[85]	Knowledge-guided by LD block	ADNI	2014	*Bioinformatics*
GNC-SCCA[92]	Genetic non-convex penalty SCCA	ADNI	2017	*Scientific Reports*
TLP-SCCA, TGL-SCCA[93]	Truncated L1 norm penalized SCCA, truncated group lasso SCCA	ADNI	2017	*Bioinformatics*
AGN-SCCA[86]	Absolute value based GraphNet SCCA	ADNI	2016	*Bioinformatics*
FDR-corrected SCCA[104]	Incorporation of FDR concept into SCCA	PNC	2018	*Transactions on Medical Imaging*
MTSCCA[91]	Multi-task SCCA	ADNI	2021	*IEEE/ACM Transactions on Computational Biology and Bioinformatics*
TGSCCA[95]	Temporally constrained group SCCA	ADNI	2017	*Bioinformatics*
T-SCCA[98]	Three-way SCCA	ADNI	2017	*Scientific Reports*
MD-SCCA[100]	Multi-modality discriminant SCCA	ADNI	2021	*IEEE/ACM Transactions on Computational Biology and Bioinformatics*
FSN-SCCA, FSN-GSCCA[101]	Fusion self-expressive network based SCCA, fusion self-expressive network based group SCCA	ADNI	2021	*Transactions on Medical Imaging*
DS-SCCA[103]	Deep self-reconstruction SCCA	ADNI	2022	*Bioinformatics*

most methods usually apply conventional learning methods or develop new learning models to combine imaging and genomics data for outcome prediction[76, 105–109]. For instance, Dukart et al.[105] investigated the role of multimodal imaging (FDG-PET, MRI, and Amyloid-PET), neuropsychological, and genetic data as potential biomarkers to identify mild cognitive impairment (MCI) patients that will suffer from AD in the future. To be specific, naive Bayes classifiers were firstly constructed for distinguishing AD and CN participants by different combinations of the data modalities mentioned above. Then, the learned classifier was applied to MCI cohort to predict AD conversion status. Related experimental results indicated that 76% accuracy is obtained by FDG-PET data and 87% is acquired via multimodal imaging and genetic data. In [106], a composite imaging genetic score was created to predict MCI conversion to AD. On the imaging hand, a nonlinear pattern recognition method[107] was firstly exploited for identifying AD-relevant volumetric regions. Then, an imaging score for each individual was obtained by applying a nonlinear support vector machine (SVM) to imaging measures from these regions. On the genomic hand, this technique utilized a linear SVM for classifying AD versus clinically normal (CN), so that a polygenic AD-related genetic score for each subject was exported. Finally, this technique created a composite imaging genetic score to be a weighted sum of the imaging score and the genetic score. Relevant results validated that the proposed composite score can effectively improve the prediction accuracy. Peng et al.[108] constructed a structured sparse kernel learning (SSKL) model used for AD prediction. In this model, each feature was expressed by a kernel and the modality information was adopted to group kernels, so that variables could be selected at both the feature and group levels. Furthermore, an innovative structured sparsity regularization was added for ensuring feature sparsity within each modality but encouraging nonsparse solution modality. The empirical study provided promising results.

In addition, for sake of understanding the biological pathway from genetics to brain structure and function, and to cognitive, behavior, and diagnostic outcomes, many studies have explored the associations among genomics, imaging, and outcomes, and there are some methods for association analysis-based outcome prediction[101, 110–114] at present. For example, a discriminative SCCA model was constructed in [110] for the identification of disease-relevant imaging proteomics associations. Without SNP data, the protein expression data collected from CSF and plasma was analyzed, as well as the relationship to imaging QTs and multiclass diagnostic labels (CN, MCI, and AD) was studied. Furthermore, in [111], a joint learning method was developed for diagnosis-relevant imaging genomics associations, which combines both SCCA and regression (SCCAR). Here, denote z as the outcome data, and such method can be described into

$$\begin{aligned}&\max_{u,v}\frac{1}{2}\|z-Yv\|_2^2-u^{\mathrm{T}}X^{\mathrm{T}}Yv+\lambda_1R_1(u)+\lambda_2R_2(v)\\&\quad\text{s.t.}\ \ \|Xu\|_2^2=\|Yv\|_2^2=1.\end{aligned}\tag{23}$$

As can be known, the imaging component Yv was jointly learned in order to predict the outcome z and correlate with the genomic component. An empirical study was performed on an ADNI sample.

Furthermore, in [112], a multi-task collaborative re-

gression (MT-CoReg) algorithm was presented, which can obtain outcome-relevant variables co-expressed in imaging and genomics modalities. This algorithm can be seen as a joint learning method via combining both SCCA and linear regression, which also uses the imaging component to predict outcome. Inspired by this work[112], Wang et al.[101] built the proposed fusion self-expressive network SCCA (FSN-SCCA) association model and used the well-known multi-kernel (MK)-SVM[115] for the classification of significant memory concern (SMC), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), AD, and NC. Fig. 8 shows a visual framework of the joint proposed FSN-SCCA association model and the well-known MK-SVM for outcome prediction. In experimental results, this method respectively achieved 93.76% and 73.85% for AD versus NC and EMCI versus LMCI on the ADNI dataset, and the corresponding area under the curve (AUC) values were 0.95 and 0.7.

Table 3 summarizes the example studies of combining both imaging and genomics data for outcome prediction or association analysis-based outcome prediction. It should be noted that, most of the above-reviewed methods usually do not consider the imaging phenotypes associated with the genotype in the clinical diagnosis. As can be well known that not all the variations in the brain are produced by genetic effects, and it is generally indeterminate which imaging phenotypes are meaningful for AD diagnosis and prediction. Relevant results presented that the association analysis-based outcome prediction approaches are helpful to guide disease interpretation and prediction.

5 Conclusions and future work

As an emerging frontier interdisciplinary field, imaging genomics involves a variety of scientific and research technologies such as neuroscience, imaging, genetics, medicine, biostatistics, data mining, and machine learning. Genomic and multimodal imaging data (including longitudinal brain imagings at different time points) also provide a rich data experimental platform for imaging genomics research, so that the pathogenic mechanism of the association between genes and brain structure or function can be presented through imaging endophenotypes with genetic properties. As a powerful tool of data-driven association analysis, machine learning technology

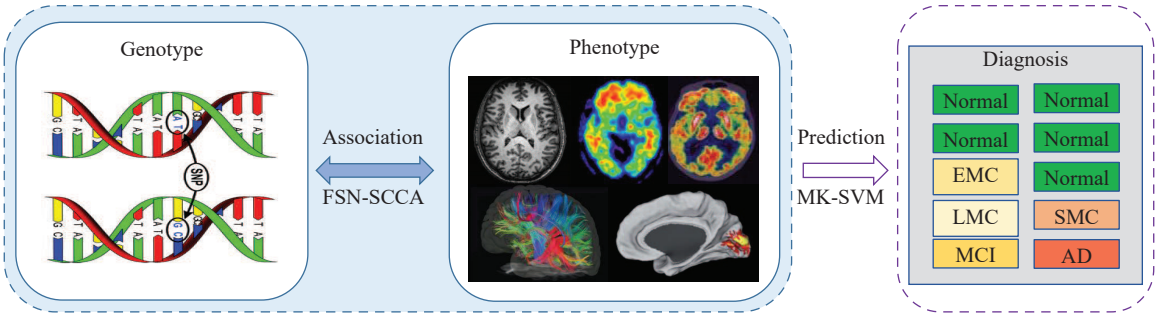

Fig. 8 Joint proposed FSN-SCCA association model and the MK-SVM to outcome prediction[101]

Table 3 Example studies of integrating imaging and genomics data for outcome prediction or association analysis-based outcome prediction

Category	Related studies	Method feature	Year	Publication
Integrate imaging and genomics data for outcome prediction	[105]	Bayes classifier	2015	*Journal of Alzheimer's Disease*
	[106]	Composite multivariate method	2012	*Journal of Alzheimer's Disease*
	SSKL[108]	AD prediction using multimodal imaging and SNP data	2016	*MICCAI*
	JCRMML[76]	Joint classification and regression framework for multimodal multitask learning	2012	*Bioinformatics*
	CaMCCo[109]	Cascaded multi-view canonical correlation	2017	*Scientific Reports*
Association analysis-based outcome prediction	DSCCA[110]	Discriminative SCCA	2017	*The 22nd Pacific Symposium on Biocomputing*
	SCCAR[111]	Combining SCCA and regression	2019	*ISBI*
	MT-CoReg[112]	Multi-task collaborative regression	2018	*Transactions on Medical Imaging*
	FSN-SCCA+MK-SVM[101]	Fusion self-expressive based network SCCA+MK-SVM	2021	*Transactions on Medical Imaging*
	[113]	Genome-wide mediation analysis	2017	*Human Brain Mapping*
	[114]	Bayesian model	2016	*Transactions on Medical Imaging*

can analyze the association between susceptible genes and brain structure or function, and better reveal the mechanism of brain cognitive behavior or related diseases by fully exploring and utilizing the internal structural information of biomarker data such as genes and imagings. This paper reviews the recent application of machine learning-based association analysis algorithms in the field of imaging genomics research. In this review, we used three databases including PubMed, Scopus, and Web of Science to select the reviewed papers. The "brain imaging genomics (genetics)" was considered as the main keyword for the selection, and the dates queries were done in March 20, 2022. A large number of experiments and reports show that some of the association results detected by the model have also been verified in the biological and medical fields.

In this paper, the structured multivariate imaging genetic association analysis methods are all based on some prior knowledge, that is, associations are made on relatively related candidate gene sets or brain area sets. Although theoretically, these methods in this paper can be applied to whole-genome or whole-brain voxel analysis, the computational efficiency is low. For the efficiency of high-dimensional feature variable gene locus detection, in addition to improving the efficiency of the algorithm itself, the calculation of big data can also be completed by introducing a distributed parallel computing method[116]. Therefore, it is necessary to further develop and construct more efficient algorithm models or working frameworks to study the imaging genetic association of genome-wide and brain wide characteristic variables.

In fact, multivariate imaging genetic research based on the structured constraint method is capable of achieving good results, since a large amount of prior knowledge is embedded in the model of data analysis. For example, as one of representative prior knowledges, LD can characterize the simple structural relationship between SNPs. On this basis, researchers can supplement the prior information and expand the model. At present, some work has considered the use of prior knowledge of biological characteristics with more genetic functions in model establishment and learning training, including gene ontology (GO), function annotation, pathway analysis system (such as KEGG (Kyoto encyclopedia of genes and genes), pathway database or OMIM (online mendelian inheritance in man) disease database)[117]. Therefore, how to design a model more suitable for practical application problems for data analysis according to these prior knowledge (such as the mechanism involved in neural regulation), that is, to realize the combination of hypothesis driven and data-driven methods[118], in order to obtain better association results, is still a current research hotspot.

Although the association results of structured multivariate genetic-multivariate imaging correlation can explain genetic effects, there may be an interactive relationship between multiple non-allelic genes among the same trait, that is, the mechanism of epistasis is not very clear. At present, there has been some work to study the interaction between SNPs on imaging QTs[119]. These methods are mainly based on traversal pairwise search methods. For example, Hibar et al.[120] used an iterative sure independence screening (SIS) algorithm to achieve and detect SNP-SNP interactions significantly associated with a brain region trait. These ergodic searches carry a considerable computational time cost, while some efficient sparse models[50] are expected to provide efficient learning algorithms for epistasis studies of multiple interactions.

In the study of outcome prediction, most of the above research on the genetic-imaging association based on supervision information are only to study the mechanism of brain cognitive behavior or disease generation and provide basis for disease diagnosis and prediction[101, 110–113]. Therefore, how to construct a multi-task unified model of joint association, regression, and classification of genetics, brain imaging, clinical scores, and diagnosis information data[112, 121], which can not only reveal the relationship between genetics and brain imaging, but also realize the diagnosis and prediction of diseases based on biomarkers. It will also become the development direction of future research in imaging genetics.

It is well-known that the deep learning has achieved great success in data-driven problems in biology and medicine. However, to the best of our knowledge, it has not been extensively applied to brain imaging genomics, which was partly caused by the limited sample size and high dimensionality of the existing imaging and genomics datasets. Recent studies have developed some deep learning based methods for outcome prediction by combining both brain imaging genomics data[122]. Since deep learning has been achieving great performance in medical image analysis[123] and multiomics research[124], we believe that developing deep learning methods to solve pressing problems in brain imaging genomics is a promising research direction.

In this article, our work focuses on the main idea and modelling in genetic-imaging association studies based on machine learning. In our brief review, the goal of imaging genomics based on machine learning is to realize the association analysis study for understanding mechanisms and pathways. At present, as the problem of data leakage in several imaging studies[125], another challenge in brain imaging genomics is how to handle data leakage leading to erroneous conclusions. Thus, data leakage in brain imaging genomics will be an important step to realize the association analysis study in order to identify some significant genetic loci and imaging phenotypic markers. Some sources of the data leakage, such as incorrect data split, late split, and the absence of an independ-

ent test set, etc., have been studied as described in several imaging studies[125]. However, systematic investigation of various data leakage factors is an underexplored topic and warrants further investigation.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Nos. 62106104, 62136004, 61902183, 61876082, 61861130366 and 61732006), the Project funded by China Postdoctoral Science Foundation (No. 2022T150320), and the National Key Research and Development Program of China (Nos. 2018YFC2001600 and 2018YFC2001602).

References

[1] A. R. Hariri, D. R. Weinberger. Imaging genomics. *British Medical Bulletin*, vol. 65, no. 1, pp. 259–270, 2003. DOI: 10.1093/bmb/65.1.259.

[2] P. M. Thompson, N. G. Martin, M. J. Wright. Imaging genomics. *Current Opinion in Neurology*, vol. 23, no. 4, pp. 368–373, 2010. DOI: 10.1097/WCO.0b013e32833b764c.

[3] D. C. Glahn, P. M. Thompson, J. Blangero. Neuroimaging endophenotypes: Strategies for finding genes influencing brain structure and function. *Human Brain Mapping*, vol. 28, no. 6, pp. 488–501, 2007. DOI: 10.1002/hbm.20401.

[4] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack Jr, W. Jagust, J. C. Morris, R. C. Petersen, A. J. Saykin, L. M. Shaw, A. W. Toga, J. Q. Trojanowski, Alzheimer′s Disease Neuroimaging Initiative. Recent publications from the Alzheimer′s disease neuroimaging initiative: Reviewing progress toward improved AD clinical trials. *Alzheimer′s & Dementia*, vol. 13, no. 4, pp. e1–e85, 2017. DOI: 10.1016/j.jalz.2016.11.007.

[5] P. M. Thompson, N. Jahanshad, C. R. K. Ching, L. E. Salminen, S. I. Thomopoulos, J. Bright, B. T. Baune, S. Bertolín, J. Bralten, W. B. Bruin, R. Bülow, J. Chen, Y. Chye, U. Dannlowski, C. G. F. de Kovel, G. Donohoe, L. T. Eyler, S. V. Faraone, P. Favre, C. A. Filippi, T. Frodl, D. Garijo, Y. Gil, H. J. Grabe, K. L. Grasby, T. Hajek, L. K. M. Han, S. N. Hatton, K. Hilbert, T. C. Ho, L. Holleran, G. Homuth, N. Hosten, J. Houenou, I. Ivanov, T. Y. Jia, S. Kelly, M. Klein, J. S. Kwon, M. A. Laansma, J. Leerssen, U. Lueken, A. Nunes, J. O′Neill, N. Opel, F. Piras, F. Piras, M. C. Postema, E. Pozzi, N. Shatokhina, C. Soriano-Mas, G. Spalletta, D. Q. Sun, A. Teumer, A. K. Tilot, L. Tozzi, C. van der Merwe, E. J. W. van Someren, G. A. van Wingen, H. Völzke, E. Walton, L. Wang, A. M. Winkler, K. Wittfeld, M. J. Wright, J. Y. Yun, G. H. Zhang, Y. Zhang-James, B. M. Adhikari, I. Agartz, M. Aghajani, A. Aleman, R. R. Althoff, A. Altmann, O. A. Andreassen, D. A. Baron, B. L. Bartnik-Olson, J. M. Bas-Hoogendam, A. R. Baskin-Sommers, C. E. Bearden, L. A. Berner, P. S. W. Boedhoe, R. M. Brouwer, J. K. Buitelaar, K. Caeyenberghs, C. A. M. Cecil, R. A. Cohen, J. H. Cole, P. J. Conrod, S. A. de Brito, S. M. C. de Zwarte, E. L. Dennis, S. Desrivieres, D. Dima, S. Ehrlich, C. Esopenko, G. Fairchild, S. E. Fisher, J. P. Fouche, C. Francks, S. Frangou, B. Franke, H. P. Garavan, D. C. Glahn, N. A. Groenewold, T. P. Gurholt, B. A. Gutman, T. Hahn, I. H. Harding, D. Hernaus, D. P. Hibar, F. G. Hillary, M. Hoogman, H. E. H. Pol, M. Jalbrzikowski, G. A. Karkashadze, E. T. Klapwijk, R. C. Knickmeyer, P. Kochunov, I. K. Koerte, X. Z. Kong, S. L. Liew, A. P. Lin, M. W. Logue, E. Luders, F. Macciardi, S. Mackey, A. R. Mayer, C. R. McDonald, A. B. McMahon, S. E. Medland, G. Modinos, R. A. Morey, S. C. Mueller, P. Mukherjee, L. Namazova-Baranova, T. M. Nir, A. Olsen, P. Paschou, D. S. Pine, F. Pizzagalli, M. E. Rentería, J. D. Rohrer, P. G. Sämann, L. Schmaal, G. Schumann, M. S. Shiroishi, S. M. Sisodiya, D. J. A. Smit, I. E. Sønderby, D. J. Stein, J. L. Stein, M. Tahmasian, D. F. Tate, J. A. Turner, O. A. van den Heuvel, N. J. A. van der Wee, Y. D. van der Werf, T. G. M. van Erp, N. E. M. van Haren, D. van Rooij, L. S. van Velzen, I. M. Veer, D. J. Veltman, J. E. Villalon-Reina, H. Walter, C. D. Whelan, E. A. Wilde, M. Zarei, Vladimir Zelman for the ENIGMA Consortium. ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. *Translational Psychiatry*, vol. 10, no. 1, Article number 100, 2020. DOI: 10.1038/s41398-020-0705-1.

[6] T. D. Satterthwaite, M. A. Elliott, K. Ruparel, J. Loughead, K. Prabhakaran, M. E. Calkins, R. Hopson, C. Jackson, J. Keefe, M. Riley, F. D. Mentch, P. Sleiman, R. Verma, C. Davatzikos, H. Hakonarson, R. C. Gur, R. E. Gur. Neuroimaging of the Philadelphia neurodevelopmental cohort. *NeuroImage*, vol. 86, pp. 544–553, 2014. DOI: 10.1016/j.neuroimage.2013.07.064.

[7] K. Marek, S. Chowdhury, A. Siderowf, S. Lasch, C. S. Coffey, C. Caspell-Garcia, T. Simuni, D. Jennings, C. M. Tanner, J. Q. Trojanowski, L. M. Shaw, J. Seibyl, N. Schuff, A. Singleton, K. Kieburtz, A. W. Toga, B. Mollenhauer, D. Galasko, L. M. Chahine, D. Weintraub, T. Foroud, D. Tosun-Turgut, K. Poston, V. Arnedo, M. Frasier, T. Sherer, Parkinson′s Progression Markers Initiative. The Parkinson′s progression markers initiative (PPMI)-Establishing a PD biomarker cohort. *Annals of Clinical and Translational Neurology*, vol. 5, no. 12, pp. 1460–1477, 2018. DOI: 10.1002/acn3.644.

[8] I. I. Gottesman, T. D. Gould. The endophenotype concept in psychiatry: Etymology and strategic intentions. *The American Journal of Psychiatry*, vol. 160, no. 4, pp. 636–645, 2003. DOI: 10.1176/appi.ajp.160.4.636.

[9] A. Meyer-Lindenberg, D. R. Weinberger. Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nature Reviews Neuroscience*, vol. 7, no. 10, pp. 818–827, 2006. DOI: 10.1038/nrn1993.

[10] T. Ge, G. Schumann, J. F. Feng. Imaging genetics-towards discovery neuroscience. *Quantitative Biology*, vol. 1, no. 4, pp. 227–245, 2013. DOI: 10.1007/s40484-013-0023-1.

[11] A. M. Winkler, P. Kochunov, J. Blangero, L. Almasy, K. Zilles, P. T. Fox, R. Duggirala, D. C. Glahn. Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. *NeuroImage*, vol. 53, no. 3, pp. 1135–1146, 2010. DOI: 10.1016/j.neuroimage.2009.12.028.

[12] S. M. Smith, P. T. Fox, K. L. Miller, D. C. Glahn, P. M.

Fox, C. E. Mackay, N. Filippini, K. E. Watkins, R. Toro, A. R. Laird, C. F. Beckmann. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 31, pp. 13040–13045, 2009. DOI: 10.1073/pnas.090526710.

[13] H. Tost, E. Bilek, A. Meyer-Lindenberg. Brain connectivity in psychiatric imaging genetics. *NeuroImage*, vol. 62, no. 4, pp. 2250–2260, 2012. DOI: 10.1016/j.neuroimage.2011.11.007.

[14] M. Rubinov, O. Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, 2010. DOI: 10.1016/j.neuroimage.2009.10.003.

[15] J. Hardy, A. Singleton. Genomewide association studies and human disease. *The New England Journal of Medicine*, vol. 360, no. 17, pp. 1759–1768, 2009. DOI: 10.1056/NEJMra0808700.

[16] R. J. Klein, C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, J. Hoh. Complement factor H polymorphism in age-related macular degeneration. *Science*, vol. 308, no. 5720, pp. 385–389, 2005. DOI: 10.1126/science.1109557.

[17] C. Esslinger, H. Walter, P. Kirsch, S. Erk, K. Schnell, C. Arnold, L. Haddad, D. Mier, C. O. von Boberfeld, K. Raab, S. H. Witt, M. Rietschel, S. Cichon, A. Meyer-Lindenberg. Neural mechanisms of a genome-wide supported psychosis variant. *Science*, vol. 324, no. 5927, Article number 605, 2009. DOI: 10.1126/science.1167768.

[18] S. E. Medland, N. Jahanshad, B. M. Neale, P. M. Thompson. Whole-genome analyses of whole-brain data: Working within an expanded search space. *Nature Neuroscience*, vol. 17, no. 6, pp. 791–800, 2014. DOI: 10.1038/nn.3718.

[19] J. Y. Liu, V. D. Calhoun. A review of multivariate analyses in imaging genetics. *Frontiers in Neuroinformatics*, vol. 8, Article number 29, 2014. DOI: 10.3389/fninf.2014.00029.

[20] P. M. Thompson, T. Ge, D. C. Glahn, N. Jahanshad, T. E. Nichols. Genetics of the connectome. *NeuroImage*, vol. 80, pp. 475–488, 2013. DOI: 10.1016/j.neuroimage.2013.05.013.

[21] D. M. Witten, R. Tibshirani, T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009. DOI: 10.1093/biostatistics/kxp008.

[22] K. A. Lê Cao, P. G. P. Martin, C. Robert-Granié, P. Besse. Sparse canonical methods for biological data integration: Application to a cross-platform study. *BMC Bioinformatics*, vol. 10, Article number 34, 2009. DOI: 10.1186/1471-2105-10-34.

[23] E. C. Chi, G. I. Allen, H. Zhou, O. Kohannim, K. Lange, P. M. Thompson. Imaging genetics via sparse canonical correlation analysis. In *Proceedings of the 10th IEEE International Symposium on Biomedical Imaging*, San Francisco, USA, pp. 740–743, 2013. DOI: 10.1109/ISBI.2013.6556581.

[24] J. Y. Liu, O. Demirci, V. D. Calhoun. A parallel independent component analysis approach to investigate genomic influence on brain function. *IEEE Signal Processing Letters*, vol. 15, pp. 413–416, 2008. DOI: 10.1109/LSP.2008.922513.

[25] V. D. Calhoun, J. Y. Liu, T. Adalimath. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage*, vol. 45, no. 1 Suppl 1, pp. S163–S172, 2009. DOI: 10.1016/j.neuroimage.2008.10.057.

[26] W. W. Daniel, C. L. Cross. *Biostatistics: A Foundation for Analysis in the Health Sciences*, 10th ed., Hoboken, USA: Wiley, 2013.

[27] S. G. Potkin, G. Guffanti, A. Lakatos, J. A. Turner, F. Kruggel, J. H. Fallon, A. J. Saykin, A. Orro, S. Lupoli, E. Salvi, M. Weiner, F. Macciardi, Alzheimer's Disease Neuroimaging Initiative. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS One*, vol. 4, no. 8, Article number e6501, 2009. DOI: 10.1371/journal.pone.0006501.

[28] L. Shen, P. M. Thompson, S. G. Potkin, L. Bertram, L. A. Farrer, T. M. Foroud, R. C. Green, X. L. Hu, M. J. Huentelman, S. Kim, J. S. K. Kauwe, Q. Q. Li, E. C. Liu, F. Macciardi, J. H. Moore, L. Munsie, K. Nho, V. K. Ramanan, S. L. Risacher, D. J. Stone, S. Swaminathan, A. W. Toga, M. W. Weiner, A. J. Saykin, Alzheimer's Disease Neuroimaging Initiative. Genetic analysis of quantitative phenotypes in AD and MCI: Imaging, cognition and biomarkers. *Brain Imaging and Behavior*, vol. 8, no. 2, pp. 183–207, 2014. DOI: 10.1007/s11682-013-9262-z.

[29] S. L. Risacher, L. Shen, J. D. West, S. Kim, B. C. McDonald, L. A. Beckett, D. J. Harvey, C. R. Jack Jr, M. W. Weiner, A. J. Saykin, Alzheimer's Disease Neuroimaging Initiative. Longitudinal MRI atrophy biomarkers: Relationship to conversion in the ADNI cohort. *Neurobiology of Aging*, vol. 31, no. 8, pp. 1401–1418, 2010. DOI: 10.1016/j.neurobiolaging.2010.04.029.

[30] S. L. Risacher, S. Kim, L. Shen, K. Nho, T. Foroud, R. C. Green, R. C. Petersen, C. R. Jack Jr, P. S. Aisen, R. A. Koeppe, W. J. Jagust, L. M. Shaw, J. Q. Trojanowski, M. W. Weiner, A. J. Saykin, Alzheimer's Disease Neuroimaging Initiative. The role of apolipoprotein E (APOE) genotype in early mild cognitive impairment (E-MCI). *Frontiers in Aging Neuroscience*, vol. 5, Article number 11, 2013. DOI: 10.3389/fnagi.2013.00011.

[31] A. J. Ho, J. L. Stein, X. Hua, S. Lee, D. P. Hibar, A. D. Leow, I. D. Dinov, A. W. Toga, A. J. Saykin, L. Shen, T. Foroud, N. Pankratz, M. J. Huentelman, D. W. Craig, J. D. Gerber, A. N. Allen, J. J. Corneveaux, D. A. Stephan, C. S. DeCarli, B. M. DeChairo, S. G. Potkin, C. R. Jack Jr, M. W. Weiner, C. A. Raji, O. L. Lopez, J. T. Becker, O. T. Carmichael, P. M. Thompson, the Alzheimer's Disease Neuroimaging Initiative, M. Weiner, L. Thal, R. Petersen, C. R. Jack Jr, W. Jagust, J. Trojanowki, A. W. Toga, L. Beckett, R. C. Green, A. Gamst, W. Z. Potter, T. Montine, D. Anders, M. Bernstein, J. Felmlee, N. Fox, P. Thompson, N. Schuff, G. Alexander, D. Bandy, R. A. Koeppe, N. Foster, E. M. Reiman, K. W. Chen, J. Trojanowki, L. Shaw, V. M. Y. Lee, M. Korecka, A. W. Toga, K. Crawford, S. Neu, D. Harvey, A. Gamst, J. Kornak, Z.

Kachaturian, R. Frank, P. J. Snyder, S. Molchan, J. Kaye, R. Vorobik, J. Quinn, L. Schneider, S. Pawluczyk, B. Spann, A. S. Fleisher, H. Vanderswag, J. L. Heidebrink, J. L. Lord, K. Johnson, R. S. Doody, J. Villanueva-Meyer, M. Chowdhury, Y. Stern, L. S. Honig, K. L. Bell, J. C. Morris, M. A. Mintun, S. Schneider, D. Marson, R. Griffith, B. Badger, H. Grossman, C. Tang, J. Stern, L. deToledo-Morrell, R. C. Shah, J. Bach, R. Duara, R. Isaacson, S. Strauman, M. S. Albert, J. Pedroso, J. Toroney, H. Rusinek, M. J. de Leon, S. M. de Santi, P. M. Doraiswamy, J. R. Petrella, M. Aiello, C. M. Clark, C. Pham, J. Nunez, C. D. Smith, C. A. Given II, P. Hardy, S. T. DeKosky, M. Oakley, D. M. Simpson, M. S. Ismail, A. Porsteinsson, C. McCallum, S. C. Cramer, R. A. Mulnard, C. McAdams-Ortiz, R. Diaz-Arrastia, K. Martin-Cook, M. DeVous, A. I. Levey, J. J. Lah, J. S. Cellar, J. M. Burns, H. S. Anderson, M. M. Laubinger, G. Bartzokis, D. H. S. Silverman, P. H. Lu, R. Fletcher, F. Parfitt, H. Johnson, M. Farlow, S. Herring, A. M. Hake, C. H. van Dyck, M. G. MacAvoy, L. A. Bifano, H. Chertkow, H. Bergman, C. Hosein, S. Black, S. Graham, C. Caldwell, H. Feldman, M. Assaly, G. Y. R. Hsiung, A. Kertesz, J. Rogers, D. Trost, C. Bernick, D. Gitelman, N. Johnson, M. Mesulam, C. Sadowsky, T. Villena, S. Mesner, P. S. Aisen, K. B. Johnson, K. E. Behan, R. A. Sperling, D. M. Rentz, K. A. Johnson, A. Rosen, J. Tinklenberg, W. Ashford, M. Sabbagh, D. Connor, S. Obradov, R. Killiany, A. Norbash, T. O. Obisesan, A. Jayam-Trouth, P. Wang, A. P. Auchus, J. B. Huang, R. P. Friedland, C. DeCarli, E. Fletcher, O. Carmichael, S. Kittur, S. Mirje, S. C. Johnson, M. Borrie, T. Y. Lee, S. Asthana, C. M. Carlsson, S. G. Potkin, D. Highum, A. Preda, D. Nguyen, P. N. Tariot, B. A. Hendin, D. W. Scharre, M. Kataki, D. Q. Beversdorf, E. A. Zimmerman, D. Celmins, A. D. Brown, S. Gandy, M. E. Marenberg, B. W. Rovner, G. Pearlson, K. Blank, K. Anderson, A. J. Saykin, R. B. Santulli, N. Pare, J. D. Williamson, K. M. Sink, H. Potter, B. A. Raj, A. Giordano, B. R. Ott, C. K. Wu, R. Cohen, K. L. Wilks. A commonly carried allele of the obesity-related FTO gene is associated with reduced brain volume in the healthy elderly. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 18, pp. 8404–8409, 2010. DOI: 10.1073/pnas.0910878107.

[32] E. M. Reiman, K. W. Chen, X. F. Liu, D. Bandy, M. X. Yu, D. Lee, N. Ayutyanont, J. Keppler, S. A. Reeder, J. B. S. Langbaum, G. E. Alexander, W. E. Klunk, C. A. Mathis, J. C. Price, H. J. Aizenstein, S. T. DeKosky, R. J. Caselli. Fibrillar amyloid-β burden in cognitively normal people at 3 levels of genetic risk for Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 16, pp. 6820–6825, 2009. DOI: 10.1073/pnas.090034510.

[33] C. D. Sloan, L. Shen, J. D. West, H. A. Wishart, L. A. Flashman, L. A. Rabin, R. B. Santulli, S. J. Guerin, C. H. Rhodes, G. J. Tsongalis, T. W. McAllister, T. A. Ahles, S. L. Lee, J. H. Moore, A. J. Saykin. Genetic pathway-based hierarchical clustering analysis of older adults with cognitive complaints and amnestic mild cognitive impairment using clinical and neuroimaging phenotypes. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 153B, no. 5, pp. 1060–1069, 2010. DOI: 10.1002/ajmg.b.31078.

[34] S. Swaminathan, L. Shen, S. L. Risacher, K. K. Yoder, J. D. West, S. Kim, K. Nho, T. Foroud, M. Inlow, S. G. Potkin, M. J. Huentelman, D. W. Craig, W. J. Jagust, R. A. Koeppe, C. A. Mathis, C. R. Jack Jr, M. W. Weiner, A. J. Saykin, Alzheimer's Disease Neuroimaging Initiative. Amyloid pathway-based candidate gene analysis of [^{11}C]PiB-PET in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. *Brain Imaging and Behavior*, vol. 6, no. 1, pp. 1–15, 2012. DOI: 10.1007/s11682-011-9136-1.

[35] M. C. Chiang, M. Barysheva, K. L. McMahon, G. I. de Zubicaray, K. Johnson, G. W. Montgomery, N. G. Martin, A. W. Toga, M. J. Wright, P. Shapshak, P. M. Thompson. Gene network effects on brain microstructure and intellectual performance identified in 472 twins. *Journal of Neuroscience*, vol. 32, no. 25, pp. 8732–8745, 2012. DOI: 10.1523/JNEUROSCI.5993-11.2012.

[36] A. J. Saykin, L. Shen, T. M. Foroud, S. G. Potkin, S. Swaminathan, S. Kim, S. L. Risacher, K. Nho, M. J. Huentelman, D. W. Craig, P. M. Thompson, J. L. Stein, J. H. Moore, L. A. Farrer, R. C. Green, L. Bertram, C. R. Jack Jr, M. W. Weiner, Alzheimer's Disease Neuroimaging Initiative. Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. *Alzheimer's & Dementia*, vol. 6, no. 3, pp. 265–273, 2010. DOI: 10.1016/j.jalz.2010.03.013.

[37] S. G. Potkin, J. A. Turner, J. A. Fallon, A. Lakatos, D. B. Keator, G. Guffanti, F. Macciardi. Gene discovery through imaging genetics: Identification of two novel genes associated with schizophrenia. *Molecular Psychiatry*, vol. 14, no. 4, pp. 416–428, 2009. DOI: 10.1038/mp.2008.127.

[38] L. Shen, S. Kim, S. L. Risacher, K. Nho, S. Swaminathan, J. D. West, T. Foroud, N. Pankratz, J. H. Moore, C. D. Sloan, M. J. Huentelman, D. W. Craig, B. M. DeChairo, S. G. Potkin, C. R. Jack Jr, M. W. Weiner, A. J. Saykin, Alzheimer's Disease Neuroimaging Initiative. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. *NeuroImage*, vol. 53, no. 3, pp. 1051–1063, 2010. DOI: 10.1016/j.neuroimage.2010.01.042.

[39] J. L. Stein, X. Hua, S. Lee, A. J. Ho, A. D. Leow, A. W. Toga, A. J. Saykin, L. Shen, T. Foroud, N. Pankratz, M. J. Huentelman, D. W. Craig, J. D. Gerber, A. N. Allen, J. J. Corneveaux, B. M. DeChairo, S. G. Potkin, M. W. Weiner, P. M. Thompson, Alzheimer's Disease Neuroimaging Initiative. Voxelwise genome-wide association study (vGWAS). *NeuroImage*, vol. 53, no. 3, pp. 1160–1174, 2010. DOI: 10.1016/j.neuroimage.2010.02.032.

[40] A. Biffi, C. D. Anderson, R. S. Desikan, M. Sabuncu, L. Cortellini, N. Schmansky, D. Salat, J. Rosand, Alzheimer's Disease Neuroimaging Initiative. Genetic variation and neuroimaging measures in Alzheimer disease. *Archives of Neurology*, vol. 67, no. 6, pp. 677–685, 2010. DOI: 10.1001/archneurol.2010.108.

[41] J. S. K. Kauwe, S. Bertelsen, K. Mayo, C. Cruchaga, R. Abraham, P. Hollingworth, D. Harold, M. J. Owen, J. Williams, S. Lovestone, J. C. Morris, A. M. Goate, Alzheimer's Disease Neuroimaging Initiative. Suggestive synergy between genetic variants in TF and HFE as risk factors for Alzheimer's disease. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*,

vol. 153B, no. 4, pp. 955–959, 2010. DOI: 10.1002/ajmg.b.31053.

[42] B. C. Dickerson, D. A. Wolk. Dysexecutive versus amnestic phenotypes of very mild Alzheimer's disease are associated with distinct clinical, genetic and cortical thinning characteristics. *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 82, no. 1, pp. 45–51, 2011. DOI: 10.1136/jnnp.2009.199505.

[43] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007. DOI: 10.1086/519795.

[44] S. Gombar, H. J. Jung, F. Dong, B. Calder, G. Atzmon, N. Barzilai, X. L. Tian, J. Pothof, J. H. J. Hoeijmakers, J. Campisi, J. Vijg, Y. Suh. Comprehensive microRNA profiling in B-cells of human centenarians by massively parallel sequencing. *BMC Genomics*, vol. 13, Article number 353, 2012. DOI: 10.1186/1471-2164-13-353.

[45] Y. Benjamini, D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001. DOI: 10.1214/aos/1013699998.

[46] D. P. Hibar, J. L. Stein, O. Kohannim, N. Jahanshad, A. J. Saykin, L. Shen, S. Kim, N. Pankratz, T. Foroud, M. J. Huentelman, S. G. Potkin, C. R. Jack Jr, M. W. Weiner, A. W. Toga, P. M. Thompson, Alzheimer's Disease Neuroimaging Initiative. Voxelwise gene-wide association study (vGeneWAS): Multivariate gene-based association testing in 731 elderly subjects. *NeuroImage*, vol. 56, no. 4, pp. 1875–1891, 2011. DOI: 10.1016/j.neuroimage.2011.03.077.

[47] D. P. Hibar, J. L. Stein, O. Kohannim, N. Jahanshad, C. R. Jack, M. W. Weiner, A. W. Toga, P. M. Thompson. Principal components regression: Multivariate, gene-based tests in imaging genomics. In *Proceedings of IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, Chicago, USA, pp. 289-293, 2011. DOI: 10.1109/ISBI.2011.5872408.

[48] D. P. Hibar, O. Kohannim, J. L. Stein, M. C. Chiang, P. M. Thompson. Multilocus genetic analysis of brain images. *Frontiers in Genetics*, vol. 2, Article number 73, 2011. DOI: 10.3389/fgene.2011.00073.

[49] O. Kohannim, D. P. Hibar, J. L. Stein, N. Jahanshad, X. Hua, P. Rajagopalan, A. W. Toga, C. R. Jack Jr, M. W. Weiner, G. I. de Zubicaray, K. L. McMahon, N. K. Hansell, N. G. Martin, M. J. Wright, P. M. Thompson, Alzheimer's Disease Neuroimaging Initiative. Discovery and replication of gene influences on brain structure using LASSO regression. *Frontiers in Neuroscience*, vol. 6, Article number 115, 2012. DOI: 10.3389/fnins.2012.00115.

[50] T. Yang, J. Wang, Q. Sun, D. P. Hibar, N. Jahanshad, L. Liu, Y. L. Wang, L. Zhan, P. M. Thompson, J. P. Ye. Detecting genetic risk factors for Alzheimer's disease in whole genome sequence data via Lasso screening. In *Proceedings of the 12th IEEE International Symposium on Biomedical Imaging*, Brooklyn, USA, pp. 985-989, 2015. DOI: 10.1109/ISBI.2015.7164036.

[51] M. Silver, G. Montana, Alzheimer's Disease Neuroimaging Initiative. Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. *Statistical Applications in Genetics and Molecular Biology*, vol. 11, no. 1, Article number 7, 2012. DOI: 10.2202/1544-6115.1755.

[52] M. Yuan, Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006. DOI: 10.1111/j.1467-9868.2005.00532.x.

[53] M. Silver, P. Chen, R. Y. Li, C. Y. Cheng, T. Y. Wong, E. S. Tai, Y. Y. Teo, G. Montana. Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two Asian cohorts. *PLoS Genetics*, vol. 9, no. 11, Article number e1003939, 2013. DOI: 10.1371/journal.pgen.1003939.

[54] J. C. Barrett, B. Fry, J. Maller, M. J. Daly. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*, vol. 21, no. 2, pp. 263–265, 2005. DOI: 10.1093/bioinformatics/bth457.

[55] X. K. Hao, J. T. Yu, D. Q. Zhang. Identifying genetic associations with MRI-derived measures via tree-guided sparse learning. In *Proceedings of the 17th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Boston, USA, pp. 757-764, 2014. DOI: 10.1007/978-3-319-10470-6_94.

[56] X. K. Hao, X. H. Yao, S. L. Risacher, A. J. Saykin, J. T. Yu, H. F. Wang, L. Tan, L. Shen, D. Q. Zhang. Identifying candidate genetic associations with MRI-Derived AD-related ROI via tree-guided sparse learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 6, pp. 1986–1996, 2019. DOI: 10.1109/TCBB.2018.2833487.

[57] J. Wang, J. P. Ye. Multi-layer feature reduction for tree structured group lasso via hierarchical projection. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 1279–1287, 2015. DOI: 10.5555/2969239.2969382.

[58] X. K. Hao, J. W. Yan, X. H. Yao, S. L. Risacher, A. J. Saykin, D. Q. Zhang, L. Shen. Diagnosis-guided method for identifying multi-modality neuroimaging biomarkers associated with genetic risk factors in Alzheimer's disease. In *Proeedings of Pacific Symposium*, Kohala Coast, USA, pp. 108–119, 2016.

[59] X. K. Hao, X. H. Yao, J. W. Yan, S. L. Risacher, A. J. Saykin, D. Q. Zhang, L. Shen, Alzheimer's Disease Neuroimaging Initiative. Identifying multimodal intermediate phenotypes between genetic risk factors and disease status in Alzheimer's disease. *Neuroinformatics*, vol. 14, no. 4, pp. 439–452, 2016. DOI: 10.1007/s12021-016-9307-8.

[60] M. L. Wang, X. K. Hao, J. Huang, W. Shao, D. Q. Zhang. Discovering network phenotype between genetic risk factors and disease status via diagnosis-aligned multi-modality regression method in Alzheimer's disease. *Bioinformatics*, vol. 35, no. 11, pp. 1948–1957, 2019. DOI: 10.1093/bioinformatics/bty911.

[61] R. Tibshirani. Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical So-*

ciety: *Series B* (*Statistical Methodology*), vol. 73, no. 3, pp. 273–282, 2011. DOI: 10.1111/j.1467-9868.2011.00771.x.

[62] M. Vounou, T. E. Nichols, G. Montana, Alzheimer's Disease Neuroimaging Initiative. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *NeuroImage*, vol. 53, no. 3, pp. 1147–1159, 2010. DOI: 10.1016/j.neuroimage.2010.07.002.

[63] M. Vounou, E. Janousova, R. Wolz, J. L. Stein, P. M. Thompson, D. Rueckert, G. Montana, Alzheimer's Disease Neuroimaging Initiative. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *NeuroImage*, vol. 60, no. 1, pp. 700–716, 2012. DOI: 10.1016/j.neuroimage.2011.12.029.

[64] M. Silver, E. Janousova, X. Hua, P. M. Thompson, G. Montana, Alzheimer's Disease Neuroimaging Initiative. Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. *NeuroImage*, vol. 63, no. 3, pp. 1681–1694, 2012. DOI: 10.1016/j.neuroimage.2012.08.002.

[65] H. Wang, F. P. Nie, H. Huang, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, L. Shen, Alzheimer's Disease Neuroimaging Initiative. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the ADNI cohort. *Bioinformatics*, vol. 28, no. 2, pp. 229–237, 2012. DOI: 10.1093/bioinformatics/btr649.

[66] T. Park, G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008. DOI: 10.1198/016214508000000337.

[67] G. Casella, M. Ghosh, J. Gill, M. Kyung. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, vol. 5, no. 2, pp. 369–411, 2010. DOI: 10.1214/10-BA607.

[68] H. T. Zhu, Z. Khondker, Z. H. Lu, J. G. Ibrahim, Alzheimer's Disease Neuroimaging Initiative. Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association*, vol. 109, no. 507, pp. 977–990, 2014. DOI: 10.1080/01621459.2014.923775.

[69] Z. H. Lu, Z. Khondker, J. G. Ibrahim, Y. Wang, H. T. Zhu, Alzheimer's Disease Neuroimaging Initiative. Bayesian longitudinal low-rank regression models for imaging genetic data from longitudinal studies. *NeuroImage*, vol. 149, pp. 305–322, 2017. DOI: 10.1016/j.neuroimage.2017.01.052.

[70] H. Wang, F. P. Nie, H. Huang, J. W. Yan, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, L. Shen, Alzheimer's Disease Neuroimaging Initiative. . From phenotype to genotype: An association study of longitudinal phenotypic markers to Alzheimer's disease relevant SNPs. *Bioinformatics*, vol. 28, no. 18, pp. i619–i625, 2012. DOI: 10.1093/bioinformatics/bts411.

[71] X. Q. Wang, J. W. Yan, X. H. Yao, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, L. Shen, H. Huang. Longitudinal genotype-phenotype association study through temporal structure auto-learning predictive model. *Journal of Computational Biology*, vol. 25, no. 7, pp. 809–824, 2018. DOI: 10.1089/cmb.2018.0008.

[72] T. Zhou, K. H. Thung, M. X. Liu, D. G. Shen. Brain-wide genome-wide association study for Alzheimer's disease via joint projection learning and sparse regression model. *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 1, pp. 165–175, 2019. DOI: 10.1109/TBME.2018.2824725.

[73] X. F. Zhu, H. I. Suk, H. Huang, D. G. Shen. Structured sparse low-rank regression model for brain-wide and genome-wide associations. In *Proceedings of the 19th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Athens, Greece, pp. 344–352, 2016. DOI: 10.1007/978-3-319-46720-7_40.

[74] X. F. Zhu, H. I. Suk, H. Huang, D. G. Shen. Low-rank graph-regularized structured sparse regression for identifying genetic biomarkers. *IEEE Transactions on Big Data*, vol. 3, no. 4, pp. 405–414, 2017. DOI: 10.1109/TBDATA.2017.2735991.

[75] X. F. Zhu, W. H. Zhang, Y. Fan, Alzheimer's Disease Neuroimaging Initiative. A robust reduced rank graph regression method for neuroimaging genetic analysis. *Neuroinformatics*, vol. 16, no. 3, pp. 351–361, 2018. DOI: 10.1007/s12021-018-9382-0.

[76] H. Wang, F. P. Nie, H. Huang, S. L. Risacher, A. J. Saykin, L. Shen, Alzheimer's Disease Neuroimaging Initiative. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, vol. 28, no. 12, pp. i127–i136, 2012. DOI: 10.1093/bioinformatics/bts228.

[77] J. Y. Liu, G. Pearlson, A. Windemuth, G. Ruano G, N. L. Perrone-Bizzozero, V. Calhoun. Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Human Brain Mapping*, vol. 30, no. 1, pp. 241–255, 2009. DOI: 10.1002/hbm.20508.

[78] S. A. Meda, B. Narayanan, J. Y. Liu, N. I. Perrone-Bizzozero, M. C. Stevens, V. D. Calhoun, D. C. Glahn, L. Shen, S. L. Risacher, A. J. Saykin, G. D. Pearlson. A large scale multivariate parallel ICA method reveals novel imaging-genetic relationships for Alzheimer's disease in the ADNI cohort. *NeuroImage*, vol. 60, no. 3, pp. 1608–1621, 2012. DOI: 10.1016/j.neuroimage.2011.12.076.

[79] H. Hotelling. The most predictable criterion. *Journal of Educational Psychology*, vol. 26, no. 2, pp. 139–142, 1935. DOI: 10.1037/h0058165.

[80] N. M. Correa, Y. O. Li, T. Adali, V. D. Calhoun. Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to detection of associative networks in schizophrenia. *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 6, pp. 998–1007, 2008. DOI: 10.1109/JSTSP.2008.2008265.

[81] S. Wold, H. Martens, H. Wold. The multivariate calibration problem in chemistry solved by the PLS method: Section C generalized singular values and data analysis. In *Proceedings of a Conference Held at Pite Havsbad*, Springer, Pite Havsbad, Sweden, pp. 286–293, 1983. DOI: 10.1007/BFb0062108.

[82] A. Krishnan, L. J. Williams, A. R. McIntosh, H. Abdi. Partial least squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage*, vol. 56, no. 2, pp. 455–475, 2011. DOI: 10.1016/j.neuroimage.2010.07.034.

[83] E. Le Floch, V. Guillemot, V. Frouin, P. Pinel, C. Lalanne, L. Trinchera, A. Tenenhaus, A. Moreno, M. Zilbovicius, T. Bourgeron, S. Dehaene, B. Thirion, J. B. Poline, É. Duchesnay. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. *NeuroImage*, vol. 63, no. 1, pp. 11–24, 2012. DOI: 10.1016/j.neuroimage.2012.06.061.

[84] K. A. Lê Cao, D. Rossouw, C. Robert-Granié, P. Besse. A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, Article number 35, 2008. DOI: 10.2202/1544-6115.1390.

[85] J. W. Yan, L. Du, S. Kim, S. L. Risacher, H. Huang, J. H. Moore, A. J. Saykin, L. Shen, Alzheimer's Disease Neuroimaging Initiative. Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. *Bioinformatics*, vol. 30, no. 17, pp. i564–i571, 2014. DOI: 10.1093/bioinformatics/btu465.

[86] L. Du, H. Huang, J. W. Yan, S. Kim, S. L. Risacher, M. Inlow, J. H. Moore, A. J. Saykin, L. Shen, Alzheimer's Disease Neuroimaging Initiative. Structured sparse canonical correlation analysis for brain imaging genetics: An improved Graphnet method. *Bioinformatics*, vol. 32, no. 10, pp. 1544–1551, 2016. DOI: 10.1093/bioinformatics/btw033.

[87] D. D. Lin, V. D. Calhoun, Y. P. Wang. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. *Medical Image Analysis*, vol. 18, no. 6, pp. 891–902, 2014. DOI: 10.1016/j.media.2013.10.010.

[88] J. Fang, D. D. Lin, S. C. Schulz, Z. B. Xu, V. D. Calhoun, Y. P. Wang. Joint sparse canonical correlation analysis for detecting differential imaging genetics modules. *Bioinformatics*, vol. 32, no. 22, pp. 3480–3488, 2016. DOI: 10.1093/bioinformatics/btw485.

[89] L. Du, J. W. Yan, S. Kim, S. L. Risacher, H. Huang, M. Inlow, J. H. Moore, A. J. Saykin, L. Shen. A novel structure-aware sparse learning algorithm for brain imaging genetics. In *Proceedings of the 17th International Conference on Medical Image Computing and Computer-Assisted Intervention* Springer, Boston, USA, pp. 329–336, 2014. DOI: 10.1007/978-3-319-10443-0_42.

[90] L. Du, K. F. Liu, X. H. Yao, S. L. Risacher, J. W. Han, L. Guo, A. J. Saykin, L. Shen. Fast multi-task SCCA learning with feature selection for multi-modal brain imaging genetics. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*, Madrid, Spain, pp. 356–361, 2018. DOI: 10.1109/BIBM.2018.8621298.

[91] L. Du, K. F. Liu, X. H. Yao, S. L. Risacher, J. W. Han, A. J. Saykin, L. Guo, L. Shen. Multi-task sparse canonical correlation analysis with application to multi-modal brain imaging genetics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 1, pp. 227–239, 2021. DOI: 10.1109/TCBB.2019.2947428.

[92] L. Du, K. F. Liu, X. H. Yao, J. W. Yan, S. L. Risacher, J. W. Han, L. Guo, A. J. Saykin, L. Shen, Alzheimer's Disease Neuroimaging Initiative. Pattern discovery in brain imaging genetics via SCCA modeling with a generic non-convex penalty. *Scientific Reports*, vol. 7, no. 1, Article number 14052, 2017. DOI: 10.1038/s41598-017-13930-y.

[93] L. Du, K. F. Liu, T. Zhang, X. H. Yao, J. W. Yan, S. L. Risacher. J. W. Han, L. Guo, A. J. Saykin, L. Shen, Alzheimer's Disease Neuroimaging Initiative. A novel SCCA approach via truncated ℓ_1-norm and truncated group lasso for brain imaging genetics. *Bioinformatics*, vol. 34, no. 2, pp. 278–285, 2018. DOI: 10.1093/bioinformatics/btx594.

[94] L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, J. E. Taylor. Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage*, vol. 72, pp. 304–321, 2013. DOI: 10.1016/j.neuroimage.2012.12.062.

[95] X. K. Hao, C. X. Li, J. W. Yan, X. H. Yao, S. L. Risacher, A. J. Saykin, L. Shen, D. Q. Zhang. Identification of associations between genotypes and longitudinal phenotypes via temporally-constrained group sparse canonical correlation analysis. *Bioinformatics*, vol. 33, no. 14, pp. i341–i349, 2017. DOI: 10.1093/bioinformatics/btx245.

[96] J. Fang, C. Xu, P. Zille, D. D. Lin, H. W. Deng, V. D. Calhoun, Y. P. Wang. Fast and accurate detection of complex imaging genetics associations based on greedy projected distance correlation. *IEEE Transactions on Medical Imaging*, vol. 37, no. 4, pp. 860–870, 2018. DOI: 10.1109/TMI.2017.2783244.

[97] J. Q. Fan, Y. Feng, L. Xia. A projection-based conditional dependence measure with applications to high-dimensional undirected graphical models. *Journal of Econometrics*, vol. 218, no. 1, pp. 119–139, 2020. DOI: 10.1016/j.jeconom.2019.12.016.

[98] X. K. Hao, C. X. Li, L. Du, X. H. Yao, J. W. Yan, S. L. Risacher, A. J. Saykin, L. Shen, D. Q. Zhang, Alzheimer's Disease Neuroimaging Initiative. Mining outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease. *Scientific Reports*, vol. 7, Article number 44272, 2017. DOI: 10.1038/srep44272.

[99] W. X. Hu, A. Y. Zhang, B. Cai, V. Calhoun, Y. P. Wang. Distance canonical correlation analysis with application to an imaging-genetic study. *Journal of Medical Imaging*, vol. 6, no. 2, Article number 026501, 2019. DOI: 10.1117/1.JMI.6.2.026501.

[100] M. L. Wang, W. Shao, X. K. Hao, L. Shen, D. Q. Zhang. Identify consistent cross-modality imaging genetic patterns via discriminant sparse canonical correlation analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 4, pp. 1549–1561, 2021. DOI: 10.1109/TCBB.2019.2944825.

[101] M. L. Wang, W. Shao, X. K. Hao, D. Q. Zhang. Identify complex imaging genetic patterns via fusion self-expressive network analysis. *IEEE Transactions on Medical Imaging*, vol. 40, no. 6, pp. 1673–1686, 2021. DOI: 10.1109/TMI.2021.3063785.

[102] M. L. Wang, W. Shao, S. Huang, D. Q. Zhang. Deep self-reconstruction sparse canonical correlation analysis for brain imaging genetics. In *Proceedings of the 18th IEEE International Symposium on Biomedical Imaging*, Nice,

France, pp. 1790–1793, 2021. DOI: 10.1109/ISBI48211.2021.9434077.

[103] M. L. Wang, W. Shao, X. K. Hao, S. Huang, D. Q. Zhang. Identify connectome between genotypes and brain network phenotypes via deep self-reconstruction sparse canonical correlation analysis. *Bioinformatics*, vol. 38, no. 8, pp. 2323–2332, 2022. DOI: 10.1093/bioinformatics/btac074.

[104] A. Gossmann, P. Zille, V. Calhoun, Y. P. Wang. FDR-corrected sparse canonical correlation analysis with applications to imaging genomics. *IEEE Transactions on Medical Imaging*, vol. 37, no. 8, pp. 1761–1774, 2018. DOI: 10.1109/TMI.2018.2815583.

[105] J. Dukart, F. Sambataro, A. Bertolino, Alzheimer's Disease Neuroimaging Initiative. Accurate prediction of conversion to Alzheimer's disease using imaging, genetic, and neuropsychological biomarkers. *Journal of Alzheimer's Disease*, vol. 49, no. 4, pp. 1143–1159, 2016. DOI: 10.3233/JAD-150570.

[106] R. Filipovych, B. Gaonkar, C. Davatzikos. A composite multivariate polygenic and neuroimaging score for prediction of conversion to Alzheimer's disease. In *Proceedings of the 2nd International Workshop on Pattern Recognition in NeuroImaging*, IEEE, London, UK, pp. 105–108, 2012. DOI: 10.1109/PRNI.2012.9.

[107] Y. Fan, D. G. Shen, R. C. Gur, R. E. Gur, C. Davatzikos. COMPARE: Classification of morphological patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging*, vol. 26, no. 1, pp. 93–105, 2007. DOI: 10.1109/TMI.2006.886812.

[108] J. L. Peng, L. An, X. F. Zhu, Y. Jin, D. G. Shen. Structured sparse kernel learning for imaging genetics based Alzheimer's disease diagnosis. In *Proceedings of the 19th International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Athens, Greece, pp. 70–78, 2016. DOI: 10.1007/978-3-319-46723-8_9.

[109] A. Singanamalli, H. B. Wang, A. Madabhushi, Alzheimer's Disease Neuroimaging Initiative. Cascaded multi-view canonical correlation (CaMCCo) for early diagnosis of Alzheimer's disease via fusion of clinical, imaging and omic features. *Scientific Reports*, vol. 7, no. 1, Article number 8137, 2017. DOI: 10.1038/s41598-017-03925-0.

[110] J. W. Yan, S. L. Risacher, K. Nho, A. J. Saykin, L. Shen, Alzheimer's Disease Neuroimaging Initiative. Identification of discriminative imaging proteomics associations in Alzheimer's disease via a novel sparse correlation model. In *Proceedings of Pacific Symposium*, Big Island, USA, pp. 94–104, 2017. DOI: 10.1142/9789813207813_0010.

[111] L. Du, K. F. Liu, X. H. Yao, S. L. Risacher, L. Guo, A. J. Saykin, L. Shen. Diagnosis status guided brain imaging genetics via integrated regression and sparse canonical correlation analysis. In *Proceedings of the 16th IEEE International Symposium on Biomedical Imaging*, Venice, Italy, pp. 356–359, 2019. DOI: 10.1109/ISBI.2019.8759489.

[112] P. Zille, V. D. Calhoun, Y. P. Wang. Enforcing co-expression within a brain-imaging genomics regression framework. *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2561–2571, 2018. DOI: 10.1109/TMI.2017.2721301.

[113] X. Bi, L. Q. Yang, T. F. Li, B. S. Wang, H. T. Zhu, H. P. Zhang. Genome-wide mediation analysis of psychiatric and cognitive traits through imaging phenotypes. *Human Brain Mapping*, vol. 38, no. 8, pp. 4088–4097, 2017. DOI: 10.1002/hbm.23650.

[114] N. K. Batmanghelich, A. Dalca, G. Quon, M. Sabuncu, P. Golland. Probabilistic modeling of imaging, genetics and diagnosis. *IEEE Transactions on Medical Imaging*, vol. 35, no. 7, pp. 1765–1779, 2016. DOI: 10.1109/TMI.2016.2527784.

[115] D. Q. Zhang, Y. P. Wang, L. P. Zhou, H. Yuan, D. G. Shen, Alzheimer's Disease Neuroimaging Initiative. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011. DOI: 10.1016/j.neuroimage.2011.01.008.

[116] Y. Wang, W. Goh, L. Wong, G. Montana, Alzheimer's Disease Neuroimaging Initiative. Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes. *BMC Bioinformatics*, vol. 14, no. Suppl 16, Article number S6, 2013. DOI: 10.1186/1471-2105-14-S16-S6.

[117] X. H. Yao, J. W. Yan, S. Kim, K. Nho, S. L. Risacher, M. Inlow, J. H. Moore, A. J. Saykin, L. Shen, Alzheimer's Disease Neuroimaging Initiative. Two-dimensional enrichment analysis for mining high-level imaging genetic associations. In *Proceedings of the 8th International Conference on Brain Informatics and Health*, Springer, London, UK, pp. 115–124, 2015. DOI: 10.1007/978-3-319-23344-4_12.

[118] Q. J. M. Huys, T. V. Maia, M. J. Frank. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, vol. 19, no. 3, pp. 404–413, 2016. DOI: 10.1038/nn.4238.

[119] R. Birnbaum, D. R. Weinberger. Functional neuroimaging and schizophrenia: A view towards effective connectivity modeling and polygenic risk. *Dialogues in Clinical Neuroscience*, vol. 15, no. 3, pp. 279–289, 2013. DOI: 10.31887/DCNS.2013.15.3/rbirnbaum.

[120] D. P. Hibar, J. L. Stein, N. Jahanshad, O. Kohannim, A. W. Toga, K. L. McMahon, G. I. de Zubicaray, G. W. Montgomery, N. G. Martin, M. J. Wright, M. W. Weiner, P. M. Thompson. Exhaustive search of the SNP-SNP interactome identifies epistatic effects on brain volume in two cohorts. In *Proceedings of the 16th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Nagoya, Japan, pp. 600–607, 2013. DOI: 10.1007/978-3-642-40760-4_75.

[121] S. M. Gross, R. Tibshirani. Collaborative regression. *Biostatistics*, vol. 16, no. 2, pp. 326–338, 2015. DOI: 10.1093/biostatistics/kxu047.

[122] T. Zhou, K. H. Thung, X. Zhu, D. G. Shen. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. *Human Brain Mapping*, vol. 40, no. 3, pp. 1001–1016, 2019. DOI: 10.1002/hbm.24428.

[123] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, C. I. Sánchez. A survey on deep learning in

medical image analysis. *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. DOI: 10.1016/j.media.2017.07.005.

[124] D. Grapov, J. Fahrmann, K. Wanichthanarak, S. Khoomrung. Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. *OMICS*: *A Journal of Integrative Biology*, vol. 22, no. 10, pp. 630–636, 2018. DOI: 10.1089/omi.2018.0097.

[125] J. H. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-Gonzalez, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, Alzheimer's Disease Neuroimaging Initiative, Australian Imaging Biomarkers and Lifestyle flagship study of ageing. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, vol. 63, Article number 101694, 2020. DOI: 10.1016/j.media.2020.101694.

Mei-Ling Wang received the M. Sc. degree in information and communication engineering from Nanjing University of Information Science and Technology, China in 2016, and the Ph. D. degree in computer science and technology from Nanjing University of Aeronautics and Astronautics, China in 2020. She is currently a postdoctor with College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China.

Her research interests include machine learning and brain imaging genetics.

E-mail: mely@nuaa.edu.cn

ORCID iD: 0000-0001-6569-2798

Wei Shao received the B. Sc. and M. Sc. degrees in information and computing science from Nanjing University of Technology, China in 2009 and 2012, respectively, and the Ph. D. degree in software engineering from Nanjing University of Aeronautics and Astronautics, China in 2018. He is currently an associate professor with College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China.

His research interests include machine learning and bioinformatics.

E-mail: 527606857@qq.com

ORCID iD: 0000-0003-1476-2068

Xiao-Ke Hao received the B. Sc. and M. Sc. degrees in computer science and technology from Nanjing University of Information Science and Technology, China in 2009 and 2012, respectively, and the Ph. D. degree in computer science and technology from Nanjing University of Aeronautics and Astronautics, China in 2017. He is currently an associate professor with School of Artificial Intelligence, Hebei University of Technology, China. He has published over 20 scientific articles in refereed journals such as *IEEE Transactions on Image Processing*, *Medical Image Analysis*, *Bioinformatics*. He is a member of the Artificial Intelligence and Pattern Recognition Society of the China Computer Federation (CCF).

His research interests include machine learning, pattern recognition and medical image analysis.

E-mail: haoxiaoke@hebut.edu.cn

ORCID iD: 0000-0003-3281-3340

Dao-Qiang Zhang received the B. Sc. and Ph. D. degrees in computer science from Nanjing University of Aeronautics and Astronautics (NUAA), China in 1999 and 2004, respectively. He joined Department of Computer Science and Engineering of NUAA, as a lecturer in 2004, and is a professor at present. He has published over 200 scientific articles in refereed international journals such as *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Medical Imaging*, *IEEE Transactions on Image Processing*, *Neuroimage*, *Human Brain Mapping*, and *Medical Image Analysis*, and conference proceedings such as IJCAI, AAAI, NIPS, CVPR, MICCAI, and KDD, with 12 000+ citations by Google Scholar. He was nominated for the National Excellent Doctoral Dissertation Award of China in 2006, and won the Best Paper Award and the Best Student Award of several international conferences such as PRICAI'06, STMI'12 and BICS'16, etc. He has served as a program committee member for some international conferences like IJCAI, AAAI, NIPS, MICCAI, SDM, PRICAI, ACML, etc. He is a member of the Machine Learning Society of the Chinese Association of Artificial Intelligence (CAAI), and the Artificial Intelligence and Pattern Recognition Society of the China Computer Federation (CCF).

His research interests include machine learning, pattern recognition, data mining and medical image analysis.

E-mail: dqzhang@nuaa.edu.cn (Corresponding author)

ORCID iD: 0000-0002-5658-7643

Machine Intelligence Research

Citation: M. L. Wang, W. Shao, X. K. Hao, D. Q. Zhang. Machine learning for brain imaging genomics methods: a review. *Machine Inteligence Research*, vol.20, no.1, pp.57–78, 2023. https://doi.org/10.1007/s11633–022–1361–0

Articles may interest you

Machine learning for cataract classification/grading on ophthalmic imaging modalities: a survey. *Machine Intelligence Research*, vol.19, no.3, pp.184-208, 2022.
DOI: 10.1007/s11633-022-1329-0

Supervised and semi-supervised methods for abdominal organ segmentation: a review. *Machine Intelligence Research*, vol.18, no.6, pp.887-914, 2021.
DOI: 10.1007/s11633-021-1313-0

Fmri-based decoding of visual information from human brain activity: a brief review. *Machine Intelligence Research*, vol.18, no.2, pp.170-184, 2021.
DOI: 10.1007/s11633-020-1263-y

Neural decoding of visual information across different neural recording modalities and approaches. *Machine Intelligence Research*, vol.19, no.5, pp.350-365, 2022.
DOI: 10.1007/s11633-022-1335-2

Application of machine learning for online reputation systems. *Machine Intelligence Research*, vol.18, no.3, pp.492-502, 2021.
DOI: 10.1007/s11633-020-1275-7

Eeg-based emotion recognition using multiple kernel learning. *Machine Intelligence Research*, vol.19, no.5, pp.472-484, 2022.
DOI: 10.1007/s11633-022-1352-1

Identification and classification of driving behaviour at signalized intersections using support vector machine. *Machine Intelligence Research*, vol.18, no.3, pp.480-491, 2021.
DOI: 10.1007/s11633-021-1295-y