BMC Bioinformatics

RESEARCH

Open Access

Neuroimaging feature extraction using a neural network classifier for imaging genetics

Cédric Beaulac[1,2*], Sidi Wu[3], Erin Gibson[1], Michelle F. Miranda[2], Jiguo Cao[3], Leno Rocha[2], Mirza Faisal Beg[1] and Farouk S. Nathoo[2]

*Correspondence:
beaulac.cedric@gmail.com

[1] School of Engineering Science, Simon Fraser University, Burnaby, Canada
[2] Department of Mathematics and Statistics, University of Victoria, Victoria, Canada
[3] Department of Statistics and Actuarial Sciences, Simon Fraser University, Burnaby, Canada

Abstract

Background: Dealing with the high dimension of both neuroimaging data and genetic data is a difficult problem in the association of genetic data to neuroimaging. In this article, we tackle the latter problem with an eye toward developing solutions that are relevant for disease prediction. Supported by a vast literature on the predictive power of neural networks, our proposed solution uses neural networks to extract from neuroimaging data features that are relevant for predicting Alzheimer's Disease (AD) for subsequent relation to genetics. The neuroimaging-genetic pipeline we propose is comprised of image processing, neuroimaging feature extraction and genetic association steps. We present a neural network classifier for extracting neuroimaging features that are related with the disease. The proposed method is data-driven and requires no expert advice or a priori selection of regions of interest. We further propose a multivariate regression with priors specified in the Bayesian framework that allows for group sparsity at multiple levels including SNPs and genes.

Results: We find the features extracted with our proposed method are better predictors of AD than features used previously in the literature suggesting that single nucleotide polymorphisms (SNPs) related to the features extracted by our proposed method are also more relevant for AD. Our neuroimaging-genetic pipeline lead to the identification of some overlapping and more importantly some different SNPs when compared to those identified with previously used features.

Conclusions: The pipeline we propose combines machine learning and statistical methods to benefit from the strong predictive performance of blackbox models to extract relevant features while preserving the interpretation provided by Bayesian models for genetic association. Finally, we argue in favour of using automatic feature extraction, such as the method we propose, in addition to ROI or voxelwise analysis to find potentially novel disease-relevant SNPs that may not be detected when using ROIs or voxels alone.

Keywords: Dimensionality reduction, Feature extraction, Neural Network Classifier, Bayesian Hierarchical Modelling, Imaging genetics

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Brain imaging genomic studies have great potential for better understanding psychopathology and neurodegenerative disorders. While high-throughput genotyping technology can determine high-density genetic markers single nucleotide polymorphisms (SNPs), neuroimaging technology provides a great level of detail of brain structure and function [1]. Various modalities of brain imaging can be used to generate meaningful biological information that can in turn be used to evaluate how genetic variation influences disease and cognition. In Alzheimer's disease (AD), structural modalities such as magnetic resonance imaging (MRI) can detect the presence of neuronal cell loss and gray matter atrophy, both indicators of neurodegeneration. Such neuroimaging phenotypes are attractive because they are closer to the biology of genetic function than clinical diagnosis [2].

Imaging genetic data analysis is a statistically challenging task due to the high dimension of both the neuroimages and genetic data. Further increasing the challenge is the fact that the data can be of multiple forms; neuroimages can be collected in multiple formats, e.g. MRI, Computerised Tomography (CT), Positron Emission Tomography (PET) using different machines and in different institutions. Consequently, it is important to find a general solution to the dimension problem that is applicable on a wide range of data structure, which is what we propose in this manuscript.

We consider studies having an emphasis on exploring the relation between genetic variation and brain imaging from structural modalities such as MRI and consider associated statistical methodology for dimension reduction and genetic variable selection. We focus our effort on the identification of SNPs that are potentially related to disease, for example, AD, with brain imaging endophenotypes which have the potential to provide additional structure related to the underlying etiology of the disease. Existing approaches for such analysis are based on considering the imaging data through a specific set of regions of interest (ROIs) (see, e.g., [3–7]) or they are based on a full voxel-wise analysis with statistical models fit at each voxel (see, e.g., [8–12]).

The first approach for statistical analysis in studies of imaging genetics developed brain-wide and genome-wide mass univariate analyses [9]. A drawback of this framework is that it ignores linkage disequilibrium and the associated multicollinearity between genetic markers as well as dependence between the components of the imaging phenotype. Hibar et al. [8] employed gene-based dimensionality reduction to avoid collinearity of SNP vectors. Vounou et al. [6] employed sparse procedures based on reduced-rank regression while Ge et al. [11] considered multi-locus interactions and developed kernel machine approaches. A review of methods is provided by Nathoo et al. [13].

Bayesian joint modelling combining imaging, genetic and disease data has been considered in [14] and [15]. The proposed joint models use logistic regression to relate disease endpoints to imaging-based features and a second regression relates imaging to genetic markers. Spike-and-slab selection is employed in both regression components of the joint model. Hierarchical models accounting for spatial dependence in the imaging phenotype using Markov random fields have been developed in [16] and [17]. Zhu et al. [7] developed a Bayesian reduced rank regression reducing the dimension of the regression coefficient matrix and incorporating a sparse latent factor representation for

the covariance matrix of the imaging data based on a gamma process prior. Kundu et al. [18] proposed a semiparametric conditional graphical model for imaging genetics within the context of functional brain connectivity where a Dirichlet process mixture is used for clustering regression coefficients into a modular structure. Azadeh et al. [19] developed a voxelwise Bayesian approach that began by partitioning the brain into ROIs and then fitting multivariate regression models to lower-dimensional projections of the voxel-specific data within each ROI separately and in parallel across ROIs.

We investigate here a new approach for extracting imaging features in either the ROI or the voxelwise setting. Statistical learning approaches for feature construction and dimension reduction have been developed based on a number of approaches such as Gaussian Mixture Models (GMM) [20] and Principal Component Analysis (PCA) [21]. In the former, Chaddad et al. use the assignment weights of GMMs as a set of features while in the latter the low-dimension projection of PCA plays the role of extracted features. The ability of neural networks (NNs) to effectively reduce the dimension of large data has been known for some time [22]. Since then, NNs have been at the foundation of multiple feature extraction models [23–25] in image analysis. The autoencoder (AE) is a commonly used NN model for feature extraction [26, 27]. It consists of two pieces, an encoder and a decoder. The former compresses the data, embedding it within a lower-dimensional representation, while the latter decompresses this representation to its original dimension. Both of these components are optimized simultaneously so as to reduce the reconstruction error. The encoder and the decoder can take various forms but we will assume both are NNs.

Predicting a diagnosis successfully using NNs is also supported by a large literature [28–34] that has demonstrated that various modern neural network architectures, such as Convolutional Neural Networks (CNNs) [35–37], weighted probabilistic neural networks [38] and ensembles of deep neural networks [36, 39] can achieve extremely high accuracy in the classification of MRI and PET scans. Shen et al. [40] present a thorough review of early applications of deep learning in medical imaging. Specifically within the context of imaging genetics, Ning et al. [41] were among the first to apply NN approaches. Their approach was to train a NN taking both imaging data and genetic markers as inputs to predict a binary disease response (AD diagnosis).

In the manuscript, we first present a novel three-step imaging genetic pipeline: image processing, feature extraction and finally genetic inference. This separates the pieces where we do not require strong interpretability such as image processing and feature extraction from the pieces where we do need interpretability, namely in genetic inference. Then, we argue in favour of using a prediction model for the feature extraction step. Finally, we implement a simple version of the proposed pipeline as a proof of concept and discuss our findings.

This separation is beneficial for multiple reasons. First, it allows us to utilize the increased prediction accuracy of blackbox models for feature extraction without suffering from their drawbacks such as the lack of interpretability of these models or their inability to provide us with rigorous confidence intervals or anything statistically equivalent. Additionally, it is easy to modify and improve the three pieces individually, making this pipeline applicable to a wide range of data structures. This is central to our contribution because what we propose is a general approach exemplified with a specific

implementation of the approach. This way, our proposed pipeline is applicable to a wide range of imaging data and can be constructed with the latest state-of-the-art models.

Consequently, the novelty of our pipeline lies in how we utilize well-established models altogether so that the resulting SNP selection has greater meaning and relevance for disease while the imaging features are nonlinear representations that are otherwise not attainable through standard voxelwise and ROI based imaging genetic analysis.

Using a classification model for feature extraction ensures that the lower-dimensional representation, the extracted features, is relevant in predicting the neurological disease of interest. A popular NN architecture for feature extraction is the AE. However, there is no way to guarantee that the lower-dimension representation is correlated with the disease of interest. Using a NNC to extract features is a way to combine the strength of AEs for producing low-dimensional representations with the high predictive accuracy of NNCs to extract features relevant to disease diagnosis. Those features are subsequently related to genetics using a Bayesian inference model accounting for grouping of regression coefficients within SNPs and within genes.

We demonstrate that it is possible to achieve higher prediction accuracy to classify disease status (AD relative to normal controls (NC)) when using NNC features compared with features used previously in the literature based on known AD ROIs. This improvement in classification accuracy could be made even larger by using more sophisticated models but this is outside of the scope of this manuscript where our focus is imaging genetics. We do not argue in favour of a specific model for image classification but rather in favour of using classification models for feature extraction. Consequently, what we propose is a general approach where the classification model can be changed depending on the task at hand.

The rest of the paper proceeds as follows. We introduce our proposed pipeline in Sect. 2. Then, in Sect. 3 we discuss our experimental testing setup and an implementation of the proposed approaches with ADNI data. Section 4 presents our findings on a case example. Finally, Sect. 5 concludes with a discussion about our experimental results, implications of the findings and possible extensions.

Proposed pipeline

Concept

Based on the premise that neuroimaging data is a better representation of the phenotype of interest than clinical diagnostics, we aim at capturing genetic variations related to the disease by directly considering the brain structure. Due to the high-dimensionality of neuroimaging, we propose NNs to extract features related to disease while simultaneously reducing data's dimensionality.

We assume that the natural generation of data follows the premise that genotype is related to brain structure that in turn is related to disease as explained in [42], sequentially in that order. Our framework thus reverses this process which, while clearly an oversimplification, provides a useful mechanism for thinking about data analysis and SNP selection.

The automated disease-relevant feature extraction is based on training a classification model on the imaging data with the disease diagnostic variable as output. Without loss of generality, we propose a NN, without specifically proposing an architecture

at this moment. The neurons of the second to last layer of this NN prediction function act as the features extracted by the model. Because the NN is optimized to predict disease diagnosis as accurately as possible using the image data, those neurons are in fact the variables constructed from the images that are the most appropriate to predict the disease and are consequently features relevant for SNP selection. An alternative, which we make comparisons to in our test analysis are features extracted from known disease regions using expert knowledge.

Formal definition

Let $v_{n,m}$ denote voxel $m \in \{1, \dots, M\}$ for subject $n \in \{1, \dots, N\}$ and $\mathbf{v}_n$ denote the complete imaging data for subject n. We identify with $\mathbf{v}_n^*$ the processed image for subject n. Here, the processed images may take on different forms but $\mathbf{v}_n^*$ is some standardized image data that the prediction model f takes as input. The processing might only involve image registration in its simplest form or it might involve the extraction of volumetric and cortical thickness statistics using FreeSurfer for instance. Then, y_n is the disease phenotype for subject n which can be binary or multi-class categorical. We further let $g_{n,s}$ denote the genetic variant $s \in \{1, \dots, S\}$ for subject n so that $\mathbf{g}_n$ is the genetic data for subject n.

Let h be the image processing function which takes the images $\mathbf{v}$ as input and outputs $\mathbf{v}^*$. We define as f the classification function which takes $\mathbf{v}^*$, the processed imaging data, as input and outputs y, the disease phenotype. We define f as a NN function composed of L layers each identified as $f_l : l \in (1, L)$, f_1 being the input layer of f_L the output layer. Each layer l may have a different number of neurons x, say K_l. In our current parametrization, the output layer is a K_L-dimensional vector, $\mathbf{o}$, where $o_{n,k} = \hat{P}(y_n = k)$, the predicted probability that subject n belongs to class k. After training the neural network f, we fit a statistical model, p, which has the genetic data $\mathbf{g}$ as explanatory variable and the neurons of the second to last layer of f, f_{L-1}, as response.

A detailed representation of the proposed pipeline is shown in Fig. 1. All components previously described are trained as follows: (i) process the raw images $\mathbf{v}$, (ii) train a prediction model, f, of choice by taking the processed images $\mathbf{v}^*$ as inputs and the diagnosis score as output and (iii) train the inference model p that predicts the features extracted from the prediction model using genetic markers as inputs. The use of a statistical model as our choice of inference model is based on the current availability of interpretable and inference-focused models in the literature.

The proposed approach can be generalized to include various prediction models such as CNNs taking images as inputs or different NN architectures with inputs being the imaging features extracted from commonly used softwares such as FreeSurfer developed by Dale, Fischl and Sereno (see [43, 44]). This setup also has the flexibility to easily handle multiple brain imaging modalities which would extract features from, for example EEG, MRI and fMRI using a modular NN with different modules corresponding to different modalities. Similarly, a wide range of inference models can be used and later combined using Bayesian model averaging techniques that account for model uncertainty at the inference stage.

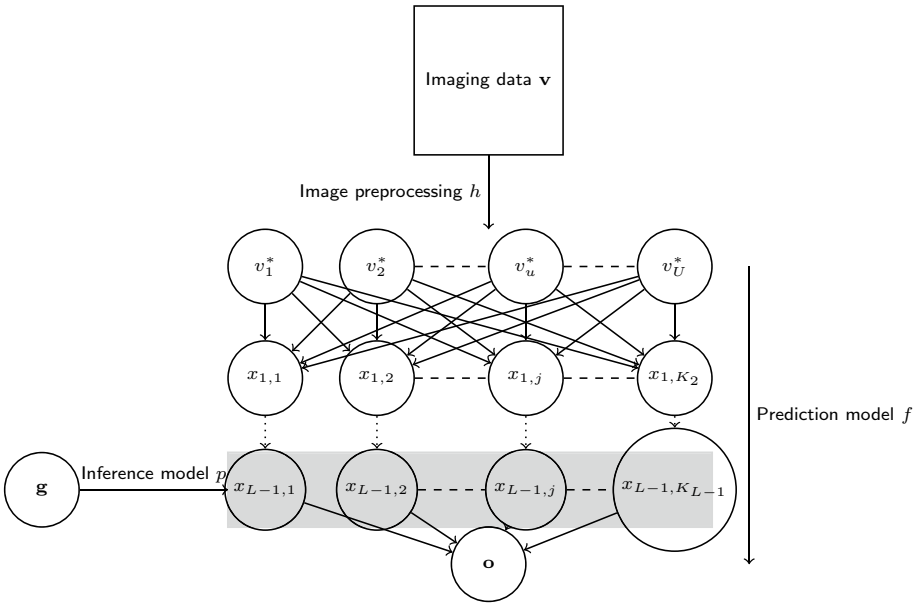

Fig. 1 A conceptual representation of the proposed methodology. In this instance, the prediction model f is depicted as a fully connected NN with L layers

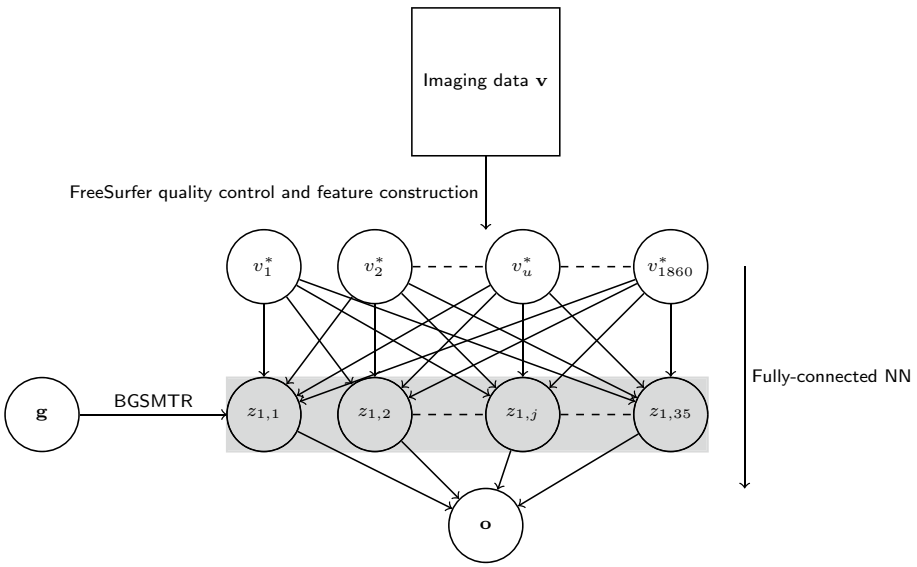

Fig. 2 Experimental implementation of the proposed pipeline

Methods

The aim of this section is to provide readers with a concrete implementation of the proposed pipeline to lay out a test application. It also provides results that highlight the benefits and drawbacks of the proposed approach. In the following test analysis we use disease (AD), MRI and genetic data from the ADNI1 study. FreeSurfer is used for image processing, a simple NN for the prediction model and a multivariate group-sparse Bayesian regression model for SNP selection. Figure 2 provides a visual representation of this simple implementation.

We compare the prediction accuracy of the 56 volumetric and cortical thickness measurements considered in [5, 45], and [46], which include locations of regions of interest such as the hippocampus, cerebellum and ventricles relevant for AD, with features automatically extracted by our proposed technique. We also compare the SNPs identified given those two sets of phenotype features. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

Cohort of subjects

The cohort of subjects we use in our test application has been previously described by Mirabnahrazam et al. [34]. Briefly, the ADNI1 database has genetic information for 818 subjects. Genotyping information of the ADNI1 subjects was downloaded in PLINK [47] format from the LONI Image Data Archive (https://ida.loni.usc.edu/). During the genotyping phase, 620,901 SNPs were obtained on the Illumina Human610-Quad BeadChip platform. Genomic quality control was conducted using the PLINK software and yielded 521,014 SNPs for 570 subjects. When excluding subjects that had no diagnosis label available, we ended up with 543 subjects for our analysis. The diagnosis values we consider for this experiment are NC, MCI and AD.

In summary, we have a cohort of 543 subjects with 145 NC, 256 MCI and 142 AD. We have T1-weighted baseline MRI scans for every subject as well as 521,014 SNPs.

Image preprocessing

The T1-weighted baseline MRI scans were downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (n=543). A detailed description of the MRI acquisition protocols can be found on the ADNI website (https://adni.loni.usc.edu/methods/documents/mri-protocols). The T1-weighted images $\mathbf{v}$ were then segmented into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) tissue compartments using Freesurfer (version 6.0), which is freely available for download (http://surfer.nmr.mgh.harvard.edu), and has been described previously [43, 44, 48]. A standardized quality control procedure was used to manually identify and correct any errors in the automated tissue segmentation in accordance with FreeSurfer's troubleshooting guidelines. Subsequently, cortical GM was parcellated into 68 regions using FreeSurfer's cortical Desikan-Killiany atlas [49] and 62 regions using Freesurfer's Desikan-Killiany-Tourville atlas [50]. Subcortical GM was parcellated into 45 regions using Freesurfer's "aseg" atlas and subcortical WM was parcellated into 70 regions using Freesurfer's "wmparc" atlas [51]. For the white matter parcellation (wmparc), optional Freesurfer parameters were used to ensure the entire white matter compartment was parcellated, (not just WM within a fixed default distance from GM), and any T1 hypotensities were labelled as white matter. This was done to ensure that the white matter parcellation included all white matter voxels and was not biased by individual T1 hypointensity

burden. For all other parcellations, the default Freesurfer options were used. From these four parcellations, a total of 1860 features were obtained. These features included:

- The volume, mean, standard deviation, min, max, and range of Freesurfer normalized T1 intensity values for the "aseg" (270 total features) and "wmparc" (420 total features) atlas parcellations.
- The number of vertices, surface area, gray matter volume, thickness (mean, standard deviation), curvature (mean, Gaussian), folding index, and curvature index for the Desikan-Killiany-Tourville (558 total features) and Desikan-Killiany (612 total features) atlas parcellations.

These 1860 features form $\mathbf{v}^*$, the processed image.

Prediction model for feature extraction

We propose a fully connected NN as a prediction model for this simple test application. The inputs of our prediction model are the entirety of the features extracted with FreeSurfer described previously, $\mathbf{v}^*$. The output is AD diagnosis, which is a categorical variable for the ADNI1 data base and finally, the second to last layer of this NN are the features we are interested in.

In this proposed approach, there is great flexibility to build the early stages of the NN. Specifically, we have control over the number of hidden layers and the non-linear activation function. Assuming the response is a K_L-class categorical variable, the output of the NN is a K_L-dimensional vector $\mathbf{o}$ where $o_{n,k} = \hat{P}(y_n = k)$ which represents the belief that subject n belongs to class k. The relation between the second to last layer and the output layer can be thought of as the one established between predictors and output in a multiclass logistic regression. To do so, we take K_L linear combinations of the K_{L-1} inputs $\mathbf{x}_{L-1}$, so that $\mathbf{o}^* = B\mathbf{x}_{L-1}$, where B is a $K_L \times K_{L-1}$ matrix of coefficients. Then, as activation function, we apply, element-wise, the softmax function to make sure the values are positive and sum to one: $o_j = \frac{\exp(o_j^*)}{\sum_{k=1}^{K} \exp(o_k^*)}$.

The model is trained in a similar fashion to a multi-case logistic regression. We minimize the negative log likelihood loss $NLLL(\mathbf{o}, \mathbf{y}) = \sum_{n=1}^{N} nlll_n$ where $nlll_n = -\sum_{k=1}^{K_L} \log(o_{n,k}) 1(y_n = k)$. This is essentially the equivalent of maximizing the log likelihood of a multinomial distribution. Thus, one could think of the features extracted $\mathbf{x}_{L-1}$ to effectively be *one logistic regression away* from the disease response, however, these features are constructed from data-driven non-linear functions built from the input.

We use the Python language and the Panda package [52] to import and manipulate the data set. The feature extraction is entirely done using Python. We use the Pytorch package [53] to define and train the NNC. Our NN is a single hidden layer NN with 35 hidden nodes trained with the Adagrad [54] optimizer. Finally, in order to train the NNC to distinguish AD from NC patients and thus to extract features related with the difference between those two groups, we only keep NC and AD during the training of the NNC, thus excluding MCI patients. In other words, we train the NNC on a cohort of 287 subjects (145 NC and 142 AD).

Most of the parameters, such as the number of hidden layers (1), the optimizer (Adagrad), the learning rate (0.01), the learning decay (0) and the number of epochs (350) were selected using cross-validation with the exception of the number of neurons in the hidden layer. We have initially set the number of neurons in the second to last layer to 56 as we wanted to design our model to extract the same number of features as in previous articles [5, 45], and [46]. However, reducing its number of neurons to 35 did not decrease the accuracy, so our final set of automatically-extracted features has 35 brain features.

Inference model

The SNPs dimension contrasts with its small fraction expected to be related to the imaging phenotypes. SNPs are connected to traits through various pathways and multiple SNPs on one gene often jointly carry out genetic functionalities. Therefore, it is desirable to develop a model to exploit the group structure of SNPs.

Wang et al. [4] developed Group-Sparse Multi-task Regression and Feature Selection (G-SMuRFS) to perform simultaneous estimation and SNP selection across phenotypes. Consider matrices as boldface uppercase letters and vectors as boldface lowercase letters. Given the SNP data of the ADNI participants as $\{\boldsymbol{g}_1, ..., \boldsymbol{g}_n\} \subseteq \mathbb{R}^S$, where n is the number of participants (sample size), S is the number of SNPs (feature dimensionality), $\boldsymbol{G} = [\boldsymbol{g}_1, ..., \boldsymbol{g}_n]$, and the imaging phenotypes as $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \subseteq \mathbb{R}^C$, C the number of imaging phenotypes, $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$, $\boldsymbol{W}$ being a $S \times C$ matrix of regression coefficients, where the entry w_{ij} of the weight matrix $\boldsymbol{W}$ measures the relative importance of the i-th SNP in predicting the response of the j-th imaging phenotype, the matrix algebraic mathematical formulation of the regression is:

$$\min_{\boldsymbol{W}} ||\boldsymbol{W}^T\boldsymbol{G} - \boldsymbol{X}||_F^2 + \gamma_1||\boldsymbol{W}||_{Gr_{2,1}} + \gamma_2||\boldsymbol{W}||_{2,1}$$

where $||.||_{Gr_{2,1}}$ is the group $l_{2,1}$-norm, devised by Wang et al. [4]. We recapitulate this norm definition: consider that the SNPs, are partitioned into Q groups $\Pi = \{\pi_q\}_{q=1}^Q$, such that, the i-th row of $\boldsymbol{W}$, $\{\boldsymbol{w^i}\}_{i=1}^{m_q} \in \pi_q$ are genetically linked, m_q being the number of SNPs in π_q. Denote $\boldsymbol{W} = [\boldsymbol{W}^1 ... \boldsymbol{W}^Q]^T$, $\boldsymbol{W^q} \in \mathbb{R}^{m_q \times c} (1 \leq q \leq Q)$, then the group $l_{2,1}$ -norm can be both defined as

$$||W||_{G_{2,1}} = \sum_{q=1}^{Q} \sqrt{\sum_{i \in \pi_q} \sum_{j=1}^{c} w_{ij}^2} = \sum_{q=1}^{Q} ||\boldsymbol{W^q}||_F$$

While producing sparse point estimates of regression coefficients, the G-SMuRFS lacked standard error computation. Kyung et al. [55] demonstrated that boot-strapping standard error computations preform poorly when the true value of the coefficient is zero, so an equivalent hierarchical Bayesian model was developed in [5]. The hierarchical model takes the form

$$\mathbf{x}_\ell | \mathbf{W}, \sigma^2 \overset{ind}{\sim} MVN_c(\boldsymbol{W}^T\mathbf{g}_\ell\,,\, \sigma^2 I_c), \ell = 1, \ldots, n,$$

with the coefficients corresponding to different genes assumed conditionally independent

$$\mathbf{W}^{(q)}|\lambda_1^2, \lambda_2^2, \sigma^2 \overset{ind}{\sim} p(\mathbf{W}^{(q)}|\lambda_1^2, \lambda_2^2, \sigma^2) \qquad q = 1, \ldots, Q,$$

and with the prior distribution for each $\mathbf{W}^{(q)}$ having a density function that is based on a product of multivariate Laplace kernels

$$p(\mathbf{W}^{(q)}|\lambda_1^2, \lambda_2^2, \sigma^2) \propto \exp\left\{-\frac{\lambda_1}{\sigma}\sqrt{\sum_{i\in\pi_q}\sum_{j=1}^{c} w_{ij}^2}\right\} \prod_{i\in\pi_q} \exp\left\{-\frac{\lambda_2}{\sigma}\sqrt{\sum_{j=1}^{c} w_{ij}^2}\right\}.$$

This product Laplace density can be expressed as a Gaussian scale mixture which allows for the implementation of Bayesian inference using a standard Gibbs sampling algorithm. The algorithm is implemented in the R package *bgsmtr*, https://cran.r-project.org/web/packages/bgsmtr/bgsmtr.pdf which is available for download on the Comprehensive R Archive Network (CRAN). The selection of tuning parameters λ_1, λ_2 in this model requires cross-validation.

This model serves as our primary inference model in this test application and we refer to this model by the name of its associated package, BGSMTR.

Results

The framework we propose is designed for the identification of SNPs related to a disease of interest. In the simple implementation provided, we aim at identifying SNPs related to AD. Based on the assumption that neuroimaging features that can accurately predict disease status are more closely related to the disease, we compare the accuracy performances of logistic regression models that take NN-extracted features as inputs to the accuracy of a model that utilizes previously expert-selected features in recent imaging genetics publications such as [5, 45], and [46]. For that purpose, we proceed with 50 repetitions of random sub-sampling validation: randomly dividing the data set into a training set and a test set. The training set contains 200 observations while the 73 other observations are assigned to the test set. Compared to k-fold cross-validation, random sub-sampling validation has the benefit of allowing us to fix the size of the training and testing set independently from the number of Monte Carlo samples.

Table 1 shows the results. The model trained using the automatically extracted features not only has a significantly higher accuracy (p-value < 0.0001) but also has a smaller performance variance across the sub-samples. The better prediction performance suggests that these features are useful for subsequent genetic analysis. More sophisticated prediction models can be investigated in future studies.

To provide an additional perspective of the NN-extracted features and to visually compare them to the features selected based on standard ROIs, we compute a 2-dimensional

Table 1 Mean and standard deviation of the accuracy of a logistic regression that separate NC from AD using two different sets of features: the ROI-based features (Expert) and the features automatically extracted by our proposed NN classifier (Automatic)

Features	Mean	Standard dev.
Expert	0.81808	0.03552
Automatic	0.91726	0.02340

embedding for both sets of features using a t-distributed stochastic neighbor embedding (t-SNE) as proposed in [56], a t-distributed variant of the original SNE proposed in [57]. Different from PCA that finds a linear representation capturing as much variability as possible, the SNEs proposed in [57] try to identify a low-dimensional representation to optimally preserve a neighborhood identity. A neighborhood-preserving embedding is especially interesting here as the features are extracted to carry information about the disease status of the patient. Figures 3 and 4 contain the embeddings of the training cohort containing strictly the NC and AD patients. A randomly selected neighborhood in Fig. 3 is more likely to have a high concentration of one class compared to a randomly selected neighborhood in Fig. 4.

To begin the genetic analysis, we follow the recommendations found in [17, 46] and adjust for subject specific factors by fitting univariate least squares linear regression for every feature (both NN-derived and ROI-based features) onto the age, gender, education level, the APOE genotype and the total intracranial volume. The residuals from each regression are then used as the adjusted imaging response in the inference model.

We then proceed with a two-step process to reduce the number of SNPs selected. First, we reduce the large number of SNPs to a smaller subset of 485 potentially related with AD SNPs [5] based on expert advice. Second, we fit univariate models between every feature and every SNP and keep the top 100 SNPs based on the resulting p-values [58, 59]. We rank the SNPs by their smallest p-value, among all models they are included in.

Table 2 contains the top 20 SNPs extracted using univariate regression as explained above. Status, novel or known, is checked against two previous publications, [4] and [5]. By comparing the SNPs associated with both sets of features, we first notice the top 3 SNPs are quite similar and that overall many SNPs belong to both groups.

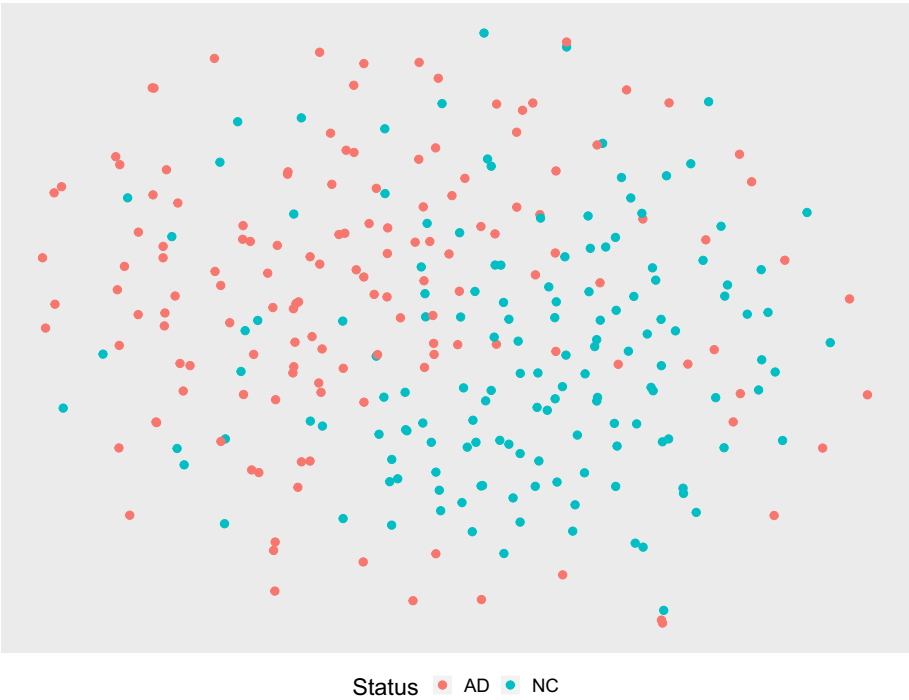

Fig. 3 2-Dimensional embedding of the NN-extracted features

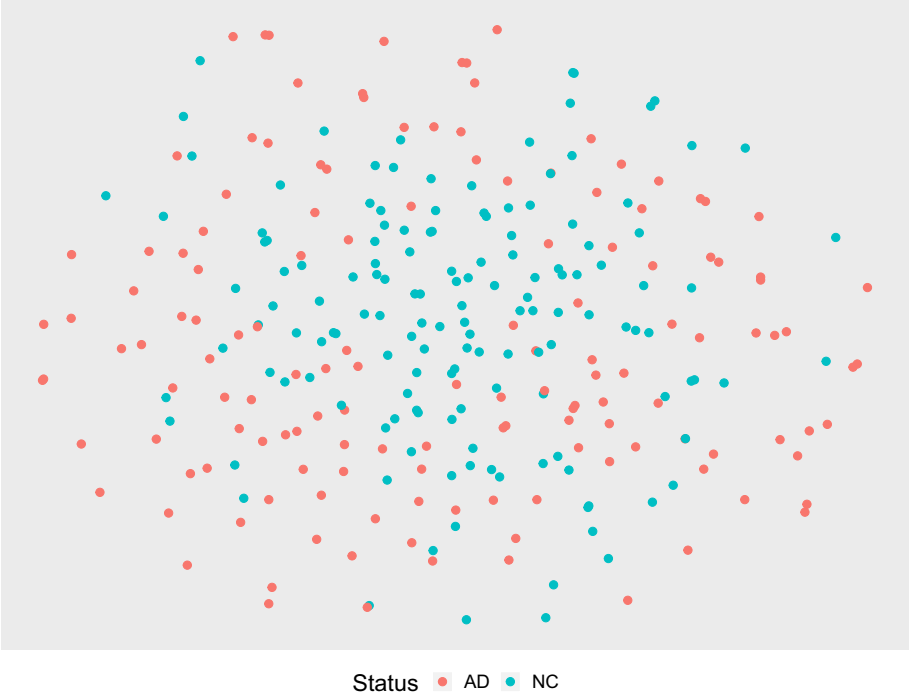

Fig. 4 2-Dimensional embedding of the expert-selected features

Table 2 Second screening results: top 20 SNPs using simple linear regression (univariate regression) with NN-extracted features and expert-extracted features, respectively

SNPs related to NN-extracted features			SNPs related to expert-extracted features		
SNP	**Gene**	**Status**	**SNP**	**Gene**	**Status**
rs12758257$^{(2)}$	ECE1		rs17399090	DAPK1	Known ([5])
rs2243581	SORCS1		**rs12758257**$^{(1)}$	ECE1	
rs213025$^{(3)}$	ECE1		**rs213025**$^{(3)}$	ECE1	
rs12756690$^{(25)}$	ECE1		**rs4935775**$^{(48)}$	SORL1	
rs6584777	SORCS1		rs2179179	NEDD9	
rs9368621	NEDD9		rs475639	PICALM	
rs213028$^{(20)}$	ECE1		rs17209374	SORCS1	
rs9461448$^{(90)}$	PGBD1		rs6905101	NEDD9	
rs11006130$^{(16)}$	TFAM	Known ([4])	rs3026841	ECE1	Known ([4])
rs3739784	DAPK1		**rs2276346**$^{(56)}$	SORL1	
rs7897726	SORCS1		**rs212531**$^{(41)}$	ECE1	
rs12001404$^{(28)}$	DAPK1		rs17367504	MTHFR	
rs3128521$^{(27)}$	DAPK1		**rs9468690**$^{(50)}$	NEDD9	
rs2450129	GAB2		rs666682	PICALM	
rs12378686$^{(43)}$	DAPK1		rs2064112	NEDD9	
rs1318241	GAB2		**rs11006130**$^{(9)}$	TFAM	
rs1114188$^{(47)}$	DAPK1		**rs11218301**$^{(29)}$	SORL1	
rs11601559	SORL1		**rs3118846**$^{(64)}$	DAPK1	Known ([5])
rs731600	GAB2		**rs3781827**$^{(45)}$	SORL1	
rs1893447	GAB2		**rs213028**$^{(7)}$	ECE1	

SNPs in bold are found related to both sets of features, the superscripted number identifies the rank of the SNPs when using the other set of features

Additionally, we notice that the genes are also quite similar between the two sets of screened SNPs. However, we also identified multiple SNPs that were not identified using the original, expert-based, features. The possibility of identifying additional SNPs based on features that are more predictive of disease is the potential added-value of the proposed approach. Thus the NN derived features can be used alongside more standard ROI-based features. These novel SNPs could simply be carrying a gene specific signature but this is also a reason why we rely on a multivariate regression model to determine the final set of SNPs.

For this reason, we follow with a subsequent multivariate regression that will better allow us to distinguish between association with the features and confounding SNPs. We use the 100 screened SNPs as predictors in our inference model, the BGSMTR model described earlier.

Table 3 contains the top 20 SNPs ranked by the posterior standard score: the posterior mean divided by the posterior standard deviation. In this table we see again a mix of novel and known SNPs and once again, the status, novel or known, is checked against two previous publications [4, 5]. Among other, identifying the association with AD through MRI features of SNPs rs1699105, rs1699105, rs2025935 and rs12209631 to name a few is consistent with previous publications [4, 5]. Since half the SNPs identified were identified in previous publications, our approach is consistent with known results and this consistency is very positive in light of the reproducibility of

Table 3 BGSMTR results: top 20 SNPs related to NN-extracted features with the highest standard score

SNP	Gene	Chromosome	Status	No. of related features
rs2243581	SORCS1	10		1
rs1699105	SORL1	11	Known ([5])	4
rs6511720	LDLR	19		15
rs6457200	NEDD9	6		8
rs11006130	TFAM	10	Known ([4])	3
rs2025935	CR1	1	Known ([4, 5])	1
rs1568400	THRA	17	Known ([5])	1
rs3785817	GRN	17		11
rs3026845	ECE1	1		1
rs12209631	NEDD9	6	Known ([5])	5
rs2418828	SORCS1	10		1
rs3118846	DAPK1	9	Known ([5])	1
rs213037	ECE1	1		1
rs1801131	MTHFR	1		3
rs9368621	NEDD9	6		1
rs3793647	DAPK1	9		2
rs17014873	BIN1	2		1
rs12758257	ECE1	1		1
rs762484	TF	1		1
rs2182335	NEDD9	6	Known ([4])	1

The last column counts the number of NN-extracted features for which a 95% credible interval excluded zero

our data-driven approach. Our approach exhibits signs of consistency and reproducibility with past experiments.

On the flip side, if we only discover known SNPs then there is little advantage to our approach. The SNP rs6511720 is ranked very high on the list and was associated with 15 features (according to 95% credible intervals with selection as in [5]). The SNPs rs6457200, rs2243581 and rs3785817 are also ranked high and/or are related with multiple features.

Discussion

The results above provide a strong argument in favour the proposed pipeline which can be used in addition to a standard voxelwise or ROI based imaging genetics analysis. The features extracted are not only better at predicting the neurological disease of interest but more importantly, these features allowed the identification of different SNPs. For instance, we identified the SNP rs6511720, being related with 15 features, and in the meanwhile this SNP was not found to be related with expert-selected features. Therefore, our proposed method could lead to the identification of novel causal SNPs. Furthermore, the extraction process is data-driven and requires no expert advice, outside of the diagnostic. Consequently, we argue in favor of using automatic feature extraction in addition to ROI or voxelwise features to find signal potentially novel SNPs that may not be detected when using ROIs or voxels alone. Our focus here is to identify SNPs related with MRI in a manner that is predictive of disease and obtain confidence intervals and posterior distributions. Integrating machine learning approaches within imaging genetics studies is of potential use as demonstrated in our analysis.

One advantage of the procedure we propose is its flexibility: we can easily improve on each of the three pieces of the pipeline separately. However, a limitation of the study of Sect. 3 is that only a single implementation was tested on a single data set. On the flip side, it opens up possible improvements for future projects. In this first implementation of our proposed pipeline, we use the well-established FreeSurfer software to obtain volumetric and cortical thickness statistics from the MRI scans. We obtain automatically extracted features in a data-driven which have higher predictive power relevant for disease. Thus, it seems reasonable to extend that principle to image processing and also try to automatically process the images in a data-driven way. For instance, a common NNC for images is the Convolutional Neural Networks (CNNs) [35–37]. Using a CNN taking as input the 3-dimensional brain scan images and training this model to predict the diagnosis would be of potentially great value for further investigation. The convolutional layers replace some of the image processing steps and the lower-level layers act as the feature extractor. However, some processing, mostly registration, would still be required. As previously demonstrated [28, 29], we expect the CNN to provide an even better prediction accuracy and thus features more closely related to AD. Another interesting approach to explore is to use an AE to reduce the dimension of the images in an unsupervised manner first. Different AEs can be trained for each brain region separately and it allows the number of features extracted per region to vary. This allows the collection of AEs to extract more features from regions with higher variability or from regions with more predictive power. Finally, a different perspective for future work would be to model and capture the complex interactions between the neuroimaging data and genetic

data using an heterogeneous information networks. Zhao et al. [60] successfully combined different data modalities for drug-disease associations using a graph representation learning model when given a biological heterogeneous information networks.

Additionally, our work demonstrates the use of different objective functions to extract features and reduce the dimension of large observations, such as neuroimages. Instead of using unsupervised models, we are able to direct the feature extraction towards a variable of interest, in our case the disease diagnostic variable which would have otherwise not been used in the analysis relating imaging to genetics. However, with gradient-based models, such as NN, we can design many other objective functions and tailor the feature extraction process for problem-specific needs. This idea can be applied in various ways when we analyse neuroimages, and we recommend considering a large collection of objective functions that are data-driven when extracting features instead of strictly relying on expert advice.

We choose to use a NN for feature extraction, this comes with strengths and weaknesses. Because our goal is to do inference at the SNP level we agreed to lose interpretability on the neuroimage feature level, this is usually considered a weakness of blackbox models such as NNs. In counterparts, this allows us to get nonlinear features that are functions of the complete processed images and the use of classification models ensure that those features are indeed most relavent to AD. The automatic feature extraction approach provides genuine added value when used alongside studies that are conducted at either the ROI or voxelwise level. It requires no external expertise for feature selection and uses disease data that are typically available but are not typically used in such analyses. The features are built considering disease prediction through nonlinear representations of neuroimaging.

Finally, the last step of our pipeline involves an inference step using a multivariate Bayesian group sparse regression. There is scope for generalizing this step to account for model uncertainty where the Bayesian model used is included within a collection of different models (e.g., [7, 14–16, 18]) and then Bayesian model averaging is used for inference at the SNP level while accounting for model uncertainty. This extension will be explored as part of future work.

Acknowledgements

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

Author contributions

CB: conceptualization, methodology, software, formal analysis, writing—original draft, writing—review & editing, visualization. SW: Software, Formal analysis, Validation, Writing—Original Draft, Writing—Review & Editing. EG: Software, Data Curation, Writing - Review & Editing. MFM: Conceptualization, Writing - Review & Editing, Supervision. JC:

Conceptualization, Writing - Review & Editing, Supervision. LR: Software, Formal analysis, Writing - Original Draft. MFB: Resources, Funding acquisition. FSN: Conceptualization, Methodology, Software, Writing - Original Draft, Writing - Review & Editing, Supervision, Funding acquisition.

Funding
The funding bodies have no specific roles. Cédric Beaulac is funded by the Canadian Statistical Sciences Institute (CANSSI). Michelle F. Miranda is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC). Jiguo Cao holds a Tier II Canada Research Chair in Data Science. Farouk S. Nathoo holds a Tier II Canada Research Chair in Biostatistics for spacial and High-Dimensional Data. Mirzal Faisal Beg and Erin Gibson are funded by Alzheimer Society Research Program (ASRP) and the National Health Institute (NIH).

Availability of data and materials
The majority of the code used in preparation of this manuscript is available on the first author's GitHub page; https://github.com/CedricBeaulac/Neuroimaging_genetics. However because the data is only available through ADNI the code on itself does not run. The ADNI data base is publicly available at https://adni.loni.usc.edu by filling the appropriate forms.

Declarations

Ethics approval and consent to participate
Ethics approval and consent to participate was collected by ADNI.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 28 September 2022 Accepted: 21 June 2023
Published online: 30 June 2023

References

1. Jinhua S, Yu X, Qiao Z, Luyun W, Ze Y, Jie Y. Predictive classification of Alzheimer's disease using brain imaging and genetic data. Sci Rep. 2022;12:2405.
2. Meyer-Lindenberg A. The future of fMRI and genetics research. NeuroImage. 2012;62(2):1286–92. https://doi.org/10.1016/j.neuroimage.2011.10.063.
3. Wang H, Nie F, Huang H, Risacher SL, Saykin AJ, Shen L, Initiative ADN. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. Bioinformatics. 2012;28(12):127–36.
4. Wang H, Nie F, Huang H, Kim S, Nho K, Risacher SL, Saykin AJ, Shen L, Initiative ADN. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. Bioinformatics. 2012;28(2):229–37.
5. Greenlaw K, Szefer E, Graham J, Lesperance M, Nathoo FS, Initiative ADN. A Bayesian group sparse multi-task regression model for imaging genetics. Bioinformatics. 2017;33(16):2513–22.
6. Vounou M, Nichols TE, Montana G, Initiative ADN. Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. Neuroimage. 2010;53(3):1147–59.
7. Zhu H, Khondker Z, Lu Z, Ibrahim JG. Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. J Am Stat Assoc. 2014;109(507):977–90.
8. Hibar DP, Stein JL, Kohannim O, Jahanshad N, Saykin AJ, Shen L, Kim S, Pankratz N, Foroud T, Huentelman MJ. Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. Neuroimage. 2011;56(4):1875–91.
9. Stein JL, Hua X, Lee S, Ho AJ, Leow AD, Toga AW, Saykin AJ, Shen L, Foroud T, Pankratz N. Voxelwise genome-wide association study (vGWAS). NeuroImage. 2010;53(3):1160–74.
10. Ge T, Feng J, Hibar DP, Thompson PM, Nichols TE. Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. Neuroimage. 2012;63(2):858–73.
11. Ge T, Nichols TE, Ghosh D, Mormino EC, Smoller JW, Sabuncu MR, Initiative ADN. A kernel machine method for detecting effects of interaction between multidimensional variable sets: an imaging genetics application. Neuroimage. 2015;109:505–14.
12. Huang M, Nichols T, Huang C, Yu Y, Lu Z, Knickmeyer RC, Feng Q, Zhu H, Initiative ADN. FVGWAS: fast voxelwise genome wide association analysis of large-scale imaging genetic data. Neuroimage. 2015;118:613–27.
13. Nathoo FS, Kong L, Zhu H, Initiative ADN. A review of statistical methods in imaging genetics. Can J Stat. 2019;47(1):108–31.
14. Batmanghelich NK, Dalca AV, Sabuncu MR, Golland P. Joint modeling of imaging and genetics. In: International Conference on Information Processing in Medical Imaging, 2013:766–777. Springer
15. Batmanghelich NK, Dalca A, Quon G, Sabuncu M, Golland P. Probabilistic modeling of imaging, genetics and diagnosis. IEEE Trans Med Imaging. 2016;35(7):1765–79.

16. Stingo FC, Guindani M, Vannucci M, Calhoun VD. An integrative Bayesian modeling approach to imaging genetics. J Am Stat Assoc. 2013;108(503):876–91.
17. Song Y, Ge S, Cao J, Wang L, Nathoo FS. A Bayesian spatial model for imaging genetics. Biometrics. 2021. https://doi.org/10.1111/biom.13460.
18. Kundu S, Kang J. Semiparametric bayes conditional graphical models for imaging genetics applications. Stat. 2016;5(1):322–37.
19. Azadeh S, Hobbs BP, Ma L, Nielsen DA, Moeller FG, Baladandayuthapani V. Integrative Bayesian analysis of neuroimaging-genetic data through hierarchical dimension reduction. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), 2016:824–828. IEEE
20. Chaddad A. Automated feature extraction in brain tumor by magnetic resonance imaging using gaussian mixture models. Int J Biomed Imaging. 2015;2015:8.
21. López M, Ramírez J, Górriz JM, Álvarez I, Salas-Gonzalez D, Segovia F, Chaves R, Padilla P, Gómez-Río M, Initiative ADN. Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease. Neurocomputing. 2011;74(8):1260–71.
22. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504–7.
23. Pradeep J, Srinivasan E, Himavathi S. Diagonal based feature extraction for handwritten character recognition system using neural network. In: 2011 3rd International Conference on Electronics Computer Technology, 2011;4:364–368. IEEE
24. Chen Y, Jiang H, Li C, Jia X, Ghamisi P. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Trans Geosci Remote Sens. 2016;54(10):6232–51.
25. El-Kenawy E-SM, Ibrahim A, Mirjalili S, Eid MM, Hussein SE. Novel feature selection and voting classifier algorithms for COVID-19 classification in CT images. IEEE Access. 2020;8:179317–35.
26. Wang W, Huang Y, Wang Y, Wang L. Generalized autoencoder: a neural network framework for dimensionality reduction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014:490–497
27. Wang Y, Yao H, Zhao S. Auto-encoder based dimensionality reduction. Neurocomputing. 2016;184:232–42.
28. Suk H-I, Lee S-W, Shen D, Initiative ADN. Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. NeuroImage. 2014;101:569–82.
29. Islam J, Zhang Y. Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. Brain Informatics. 2018;5(2):1–14.
30. Lu D, Popuri K, Ding GW, Balachandar R, Beg MF, Initiative ADN. Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimer's disease. Med Image Anal. 2018;46:26–34.
31. Lin W, Tong T, Gao Q, Guo D, Du X, Yang Y, Guo G, Xiao M, Du M, Qu X. Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. Front Neurosci. 2018;12:777.
32. Duraisamy B, Shanmugam JV, Annamalai J. Alzheimer disease detection from structural MR images using FCM based weighted probabilistic neural network. Brain Imaging Behav. 2019;13(1):87–110.
33. Jain R, Jain N, Aggarwal A, Hemanth DJ. Convolutional neural network based Alzheimer's disease classification from magnetic resonance brain images. Cognit Syst Res. 2019;57:147–59.
34. Mirabnahrazam G, Ma D, Lee S, Popuri K, Lee H, Cao J, Wang L, Galvin JE, Beg MF, Initiative ADN, et al.: Machine learning based multimodal neuroimaging genomics dementia score for predicting future conversion to alzheimer's disease. J Alzheimer's Dis (Preprint), 2022:1–21
35. LeCun Y. Generalization and network design strategies. Connect Perspect. 1989;19:143–55.
36. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
37. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016.
38. Kusy M, Kowalski PA. Weighted probabilistic neural network. Inf Sci. 2018;430:65–76.
39. Zhang C, Ma Y. Ensemble machine learning: methods and applications. Berlin: Springer; 2012.
40. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng. 2017;19:221–48.
41. Ning K, Chen B, Sun F, Hobel Z, Zhao L, Matloff W, Toga AW, Initiative ADN. Classifying alzheimer's disease with brain imaging and genetic data using a neural network framework. Neurobiol Aging. 2018;68:151–8.
42. Batmanghelich NK, Dalca A, Quon G, Sabuncu M, Golland P. Probabilistic modeling of imaging, genetics and diagnosis. IEEE Trans Med Imaging. 2016;35(7):1765–79.
43. Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis: I. Segmentation and surface reconstruction. Neuroimage. 1999;9(2):179–94.
44. Fischl B, Sereno MI, Dale AM. Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. Neuroimage. 1999;9(2):195–207.
45. Shen L, Kim S, Risacher SL, Nho K, Swaminathan S, West JD, Foroud T, Pankratz N, Moore JH, Sloan CD. 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. 2010;53(3):1051–63.
46. Szefer E, Lu D, Nathoo F, Beg MF, Graham J, Initiative ADN. Multivariate association between single-nucleotide polymorphisms in Alzgene linkage regions and structural changes in the brain: discovery, refinement and validation. Statistical Appl Genet Mol Biol. 2017;16(5–6):367–86.
47. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation plink: rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):13742–015.
48. Fischl B. FreeSurfer. Neuroimage. 2012;62(2):774–81.
49. Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage. 2006;31(3):968–80.
50. Klein A, Tourville J. 101 labeled brain images and a consistent human cortical labeling protocol. Front Neurosci. 2012;6:171.

51. Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, Van Der Kouwe A, Killiany R, Kennedy D, Klaveness S. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron. 2002;33(3):341–55.
52. Wes McKinney: Data Structures for Statistical Computing in Python. In: Stéfan van der Walt, Jarrod Millman (eds.) Proceedings of the 9th Python in Science Conference, 2010:56–61. https://doi.org/10.25080/Majora-92bf1922-00a
53. Paszke A, Gross S, Massa F, et al. An imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors, Advances in Neural Information Processing Systems (vol. 32). New York: Curran Associates Inc.; 2019. pp. 8024–8035
54. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res. 2011;12(7):2121.
55. Kyung M, Gill J, Ghosh M, Casella G. Penalized regression, standard errors, and Bayesian lassos. Bayesian Anal. 2010;5:369–411.
56. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(11)
57. Hinton GE, Roweis S. Stochastic neighbor embedding. Adv Neural Inf Process Syst 2002;15
58. Yin J, Li H. A sparse conditional gaussian graphical model for analysis of genetical genomics data. Ann Appl Stat. 2011;5(4):2630.
59. Li Y, Nan B, Zhu J. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. Biometrics. 2015;71(2):354–63.
60. Zhao B-W, Wang L, Hu P-W, Wong L, Su X-R, Wang B-Q, You Z-H, Hu L. Fusing higher and lower-order biological information for drug repositioning via graph representation learning. IEEE Trans Emerging Topics Comput. 2023. https://doi.org/10.1109/TETC.2023.3239949.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions