





## RESEARCH ARTICLE

# Sparse deep neural networks on imaging genetics for schizophrenia case–control classification

Jiayu Chen<sup>1</sup>  | Xiang Li<sup>2</sup> | Vince D. Calhoun<sup>1,2,3</sup> | Jessica A. Turner<sup>1,3</sup> | Theo G. M. van Erp<sup>4,5</sup> | Lei Wang<sup>6</sup>  | Ole A. Andreassen<sup>7</sup> | Ingrid Agartz<sup>7,8,9</sup> | Lars T. Westlye<sup>7,10</sup>  | Erik Jönsson<sup>7,9</sup> | Judith M. Ford<sup>11,12</sup> | Daniel H. Mathalon<sup>11,12</sup> | Fabio Macciardi<sup>4</sup> | Daniel S. O'Leary<sup>13</sup> | Jingyu Liu<sup>1,2</sup>  | Shihao Ji<sup>2</sup>

<sup>1</sup>Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS): (Georgia State University, Georgia Institute of Technology and Emory University), Atlanta, Georgia

<sup>2</sup>Department of Computer Science, Georgia State University, Atlanta, Georgia

<sup>3</sup>Psychology Department and Neuroscience Institute, Georgia State University, Atlanta, Georgia

<sup>4</sup>Department of Psychiatry and Human Behavior, School of Medicine, University of California, Irvine, Irvine, California

<sup>5</sup>Center for the Neurobiology of Learning and Memory, University of California, Irvine, Irvine, California

<sup>6</sup>Department of Psychiatry and Behavioral Sciences, Northwestern University, Chicago, Illinois

<sup>7</sup>Norwegian Centre for Mental Disorders Research (NORMENT), Division of Mental Health and Addiction, Institute of Clinical Medicine, University of Oslo & Oslo University Hospital, Oslo, Norway

<sup>8</sup>Department of Psychiatric Research, Diakonhjemmet Hospital, Oslo, Norway

<sup>9</sup>Department of Clinical Neuroscience, Centre for Psychiatric Research, Karolinska Institutet, Stockholm, Sweden

<sup>10</sup>Department of Psychology, University of Oslo, Oslo, Norway

<sup>11</sup>Department of Psychiatry, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, California

<sup>12</sup>Veterans Affairs San Francisco Healthcare System, San Francisco, California

<sup>13</sup>Department of Psychiatry, Carver College of Medicine, University of Iowa, Iowa City, Iowa

## Correspondence

Jiayu Chen and Jingyu Liu, 55 Park PI NE, Atlanta, GA 30303.  
Email: jchen84@gsu.edu (J. C.) and jliu75@gsu.edu (J. L.)

## Funding information

National Institutes of Health, Grant/Award Numbers: U24 RR025736-01, 1R01EB006841, 5R01MH094524, P20GM103472, P30GM122734, R01 EB020062, R01 MH084803, R01EB005846, R01MH106655, U01 MH097435, U24 RR021992; National Science Foundation, Grant/Award Numbers: 1539067, 1636893,

## Abstract

Deep learning methods hold strong promise for identifying biomarkers for clinical application. However, current approaches for psychiatric classification or prediction do not allow direct interpretation of original features. In the present study, we introduce a sparse deep neural network (DNN) approach to identify sparse and interpretable features for schizophrenia (SZ) case–control classification. An  $L_0$ -norm regularization is implemented on the input layer of the network for sparse feature selection, which can later be interpreted based on importance weights. We applied the proposed approach on a large multi-study cohort with gray matter volume (GMV) and single nucleotide polymorphism (SNP) data for SZ classification. A total of 634 individuals served as

Jiayu Chen and Xiang Li contributed equally to this study.

Jingyu Liu and Shihao Ji contributed equally as senior authors.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *Human Brain Mapping* published by Wiley Periodicals LLC.

1734853; Norges Forskningsråd, Grant/Award Number: 223273

training samples, and the classification model was evaluated for generalizability on three independent datasets of different scanning protocols ( $N = 394, 255,$  and  $160,$  respectively). We examined the classification power of pure GMV features, as well as combined GMV and SNP features. Empirical experiments demonstrated that sparse DNN slightly outperformed independent component analysis + support vector machine (ICA + SVM) framework, and more effectively fused GMV and SNP features for SZ discrimination, with an average error rate of 28.98% on external data. The importance weights suggested that the DNN model prioritized to select frontal and superior temporal gyrus for SZ classification with high sparsity, with parietal regions further included with lower sparsity, echoing previous literature. The results validate the application of the proposed approach to SZ classification, and promise extended utility on other data modalities and traits which ultimately may result in clinically useful tools.

#### KEYWORDS

deep neural network, gray matter volume, schizophrenia, single nucleotide polymorphism, sparse

## 1 | INTRODUCTION

Schizophrenia (SZ), a disabling psychiatric disorder with a lifetime prevalence  $\sim 0.8\%$ , casts a serious socioeconomic burden worldwide (McGrath, Saha, Chant, & Welham, 2008). More than a century after Kraepelin's dichotomy was formulated, precise treatment is still not available for SZ (Insel, 2014; Insel et al., 2010). Current diagnostic and treatment practice are largely based on descriptive clinical characteristics whose relationships to underlying biological processes await delineation (Cuthbert & Insel, 2013; Insel et al., 2010). This gap underlies many issues faced by clinical psychiatry, including vague boundaries between defined clinical entities, and heterogeneity within individual clinical entities. As a result, symptom presentations often do not neatly fit the categorical diagnostic system, and one diagnostic label covers biologically diverse conditions. These issues challenge treatment planning, which turns out to be largely empirical (Chen, Liu, & Calhoun, 2019; Insel & Cuthbert, 2015). It has now been widely acknowledged that objective biological markers are needed to quantify abnormalities underlying phenotypic manifestation, which allows characterizing disorders based on a multitude of dimensions and along a spectrum of functioning, so as to improve patient stratification and inform treatment planning (Casey et al., 2013; Cuthbert, 2014).

Hopes have been invested in machine learning approaches as a solution to this challenge, given the complexity of SZ. Patients with SZ present widespread structural and functional brain abnormalities, including gray matter loss in the frontal, temporal and parietal cortices and subcortical structures (Ivleva et al., 2013; van Erp et al., 2016; van Erp et al., 2018), reduced fractional anisotropy in most major white matter fasciculi (Kelly et al., 2018), as well as abnormal resting state functional connectivity in default mode, somatomotor, visual, auditory, executive control and attention networks (Garrity et al., 2007; Skatun et al., 2017; Woodward, Rogers, & Heckers, 2011). In parallel, genome wide association studies

(GWASs) of SZ lend support for a polygenic architecture, where the disease risk is attributable to many genetic variants with modest effect sizes (Ripke et al., 2014). These findings have boosted the efforts to model SZ in a multivariate framework, which is expected to not only delineate the relationships between individual biomarkers and SZ, but also to provide a generalizable mathematical model that can be used to predict risk.

One straightforward approach is to feed voxelwise neurobiological features (e.g., gray matter density) into support vector machine (SVM). With this strategy, Nieuwenhuis et al. obtained a classification accuracy of  $\sim 70\%$  which was confirmed in independent data with a sample size of a few hundred (Nieuwenhuis et al., 2012). Whether more sophisticated feature selection can be combined with classifiers to yield improved discrimination has also been explored. For instance, resting state connectivity between networks extracted by independent component analysis (ICA), followed by  $K$  nearest neighbors, yielded an accuracy of 96% in a data set consisting of 28 controls and 28 patients, which were randomly partitioned to serve as training and testing samples (Arbabshirani, Kiehl, Pearson, & Calhoun, 2013). In addition, fusion of multiple modalities that may carry complementary information of the brain holds promise for further improvement. In a work by Liang et al., combining gray and white matter features resulted in an average classification accuracy of  $\sim 76\%$  in 48 controls and 54 patients with first episode SZ, in a 10-fold cross validation set up (Liang et al., 2019). In contrast to neurobiological features, genetic variables, such as single nucleotide polymorphisms (SNPs), in general suffer modest effect sizes (Ripke et al., 2014) and could hardly be directly trained for classification. A more commonly used feature for risk discrimination is polygenic risk score (PGRS), which reflects the cumulative risk of multiple variants, and proves to be a generalizable and promising marker for disease discrimination and patient stratification (Frank et al., 2015; Vassos et al., 2017), with complementary value for group classification beyond brain magnetic resonance imaging (MRI) and cognitive data (Doan et al., 2017).

More recently, the advancement of deep learning methods has opened a new perspective on elucidating biological underpinnings of SZ. Deep Neural Networks (DNNs) are known to excel in handling high-dimensional data and automatically identifying high-level latent features, which promotes them as promising tools for better understanding of complex traits such as SZ. In a pioneer study, Plis et al. demonstrated the application of restricted Boltzmann machine-based deep belief network to structural MRI data. A classification accuracy of  $\sim 90\%$  was obtained with a 10-fold cross validation in 181 controls and 198 patients with SZ (Plis et al., 2014). A deep discriminant autoencoder network has been proposed and applied to functional connectivity features, and yielded a leave-site-out classification accuracy of  $\sim 81\%$  in 377 controls and 357 patients with SZ (Zeng et al., 2018). A comparable leave-site-out accuracy of  $\sim 80\%$  was observed in 542 controls and 558 patients with SZ, when a multi-scale recurrent neural network was applied to time courses of functional MRI data (Yan et al., 2019). However, these approaches do not provide importance weights of original biological features indicating their relative contribution to classification, making interpretation less straightforward.

As commonly implemented, DNNs are black-boxes with hundreds of layers of convolution, non-linearities, and gates, optimized solely for competitive performance. While the value of DNN may be backed up with a claimed high accuracy on benchmarks, it would be desired to be able to verify, interpret, and understand the reasoning of the system. This is particularly essential for the psychiatric community, for the purpose of deconstructing complex disorders and facilitating improved treatment. In the current work, we introduce a sparse DNN model which allows identifying sparse and interpretable features for SZ discrimination. The sparsity is achieved with an  $L_0$ -norm regularization on the input layer of the network for feature selection. Under the  $L_0$ -norm sparsity constraint, the model is trained to select the most important features while retaining the high SZ classification accuracy. We applied the sparse DNN approach on a multi-site gray matter volume (GMV) and SNP data set for SZ discrimination. In brief, a total of 634 individuals (346 controls and 288 patients with SZ) served as the training set, which was internally partitioned for hyperparameter tuning. The resulting classification model was then evaluated for generalizability on three independent data sets ( $N = 394, 255, \text{ and } 160$ , respectively). We examined the classification power of pure GMV features, as well as whether combining GMV with SNP features would benefit classification. The performance of the proposed approach was compared with that yielded by ICA + linear SVM. Empirical experiments demonstrate that the selected brain regions by sparse DNN are interpretable and echo many previous neuroscience studies.

## 2 | MATERIALS AND METHODS

### 2.1 | Participants

A total of 1,443 individuals aggregated from multiple studies, including MCIC, COBRE, FBIRN, NU, BSNIP, TOP, and HUBIN, were employed for the current study. The institutional review board at each

site approved the study and all participants provided written informed consent. Diagnosis of SZ was confirmed using the Structured Clinical Interview for Diagnosis for DSM-IV or DSM-IV-TR. Table 1 provides the primary demographic information of individual study. More details regarding scanning information are listed in Table S1, which also provides a summary of previous publications with description of recruitment. The training sample consisted of 288 cases and 346 controls from MCIC, COBRE, FBIRN, and NU. Meanwhile, three independent data sets, BSNIP ( $N = 394$ ), TOP ( $N = 255$ ) and HUBIN ( $N = 160$ ) were used for validation.

### 2.2 | Structural MRI data

Whole-brain  $T_1$ -weighted images were collected with 1.5T and 3T scanners of various models, as summarized in Table 1 and Table S1. The images of the training set were preprocessed using a standard Statistical Parametric Mapping 12 (SPM12, <http://www.fil.ion.ucl.ac.uk/spm>) voxel based morphometry pipeline (Ashburner & Friston, 2005; Gupta et al., 2015; Lin et al., 2017; Segall et al., 2009), a unified model where image registration, bias correction and tissue classification are integrated. The resulting modulated images were resliced to  $1.5 \text{ mm} \times 1.5 \text{ mm} \times 1.5 \text{ mm}$  and smoothed by 6 mm full width at half-maximum Gaussian kernel. A mask (average GMV  $> 0.2$ ) was applied to include 429,655 voxels. We further investigated correlations between individual images and the average GMV image across all the subjects. Subjects with correlations  $< 3SD$  were considered as outliers and excluded from subsequent analyses (Chen, Calhoun, et al., 2019). Finally, voxelwise regression was conducted to eliminate the effects from age, sex, and dummy-coded site covariates (Gupta et al., 2015). While all the scanning parameters (Table S1) would yield 93 dummy variables in the training data, we chose to correct scanning effects by "site" to avoid eliminating too much information due to unknown collinearity. It should be noted that many approaches have been proposed for eliminating site effects, including multisite harmonization that aims to align data distributions (Wrobel et al., 2020). However, in this work we wanted to examine the generalizability of the classification models, including the vulnerability to site effects. Consequently, we conducted a simple correction of site effects using linear regression. The validation images were preprocessed separately, using the same pipeline.

### 2.3 | SNP data

The SNP data were collected and processed as described in our previous work (Chen, Calhoun, et al., 2019). DNA samples drawn from blood or saliva were genotyped with different platforms (see Table S1). No significant difference was observed in genotyping call rates between blood and saliva samples. A standard pre-imputation quality control (QC) (Chen et al., 2013) was performed using PLINK (Purcell et al., 2007). In the imputation, SHAPEIT was used for pre-phasing (Delaneau, Marchini, & Zagury, 2012), IMPUTE2 for

**TABLE 1** Subject demographic information

Cohort	N	Sex (M/F)	Age (mean ± SD)	Age (min – max)	Diagnosis (HC/SZ)
<b>Training</b>					
MCIC + COBRE + FBIRN + NU	634	459/175	35.44 ± 12.12	16–65	346/288
<b>Validation</b>					
TOP	255	144/111	33.75 ± 8.99	17–62	154/101
HUBIN	160	108/52	41.69 ± 8.56	19–56	76/84
BSNIP	394	221/173	36.44 ± 12.47	16–64	208/186

imputation (Marchini & Howie, 2010), and the 1,000 Genomes data as the reference panel (Altshuler et al., 2012). Only markers with INFO score > 0.3 were retained. Polygenic risk scores (PGRS) for SZ were then computed using PRSice (Euesden, Lewis, & O'Reilly, 2015), which was a sum of genetic profiles weighted by the odds ratios reported in the PGC SZ GWAS, reflecting the cumulative risk for SZ of a set of SNPs (Ripke et al., 2014). Specifically, the genotype data were pruned at  $r^2 < 0.1$  (Chen et al., 2018). Then a full model PGRS was computed on 61,253 SNPs retained after pruning.

## 2.4 | Sparse DNN

Figure 1 shows the overall architecture of our method, which contains three stages. First, the GMV voxels are partitioned into a set of groups (or brain regions) with a pre-defined radius. Then a sparse DNN model is deployed for feature (brain region) selection, followed by augmenting the selected sparse regions of GMV with the SNP data for classifier retraining. In the sequel, we will introduce each of these steps in more details.

Given a GMV dataset  $D = \{(x_i, y_i), i = 1, 2, \dots, N\}$ , where  $x_i$  denotes the  $i$ -th subject's GMV image and  $y_i$  denotes the corresponding label: case or control, we train a neural network  $h(x; \theta)$ , parameterized by  $\theta$ , to fit to the dataset  $D$  with the goal of achieving good generalization to unseen test data. For a GMV image  $x \in R^{M \times 1}$ , we use  $x^j$  to represent the  $j$ -th voxel of image  $x$ , where  $j = 1, 2, \dots, M$  and  $M = 429,655$  in our study.

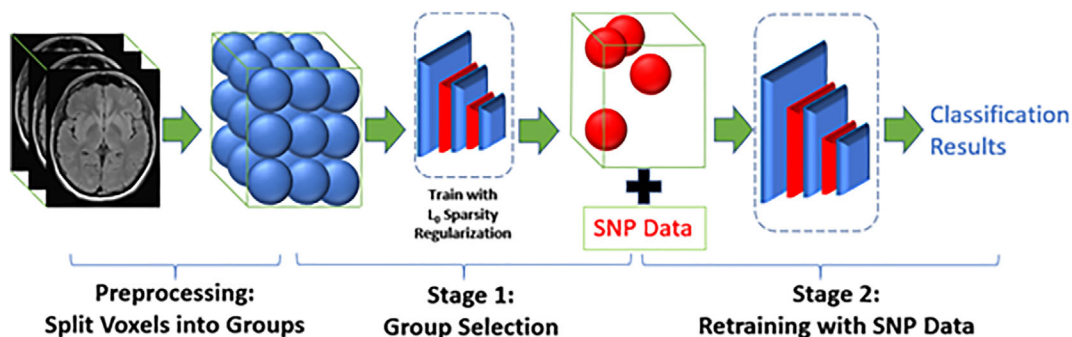
As the number of voxels  $M$  is much larger than the number of functional regions of human brain (e.g., typically around 100 as

defined by various brain atlases), we first partition the brain voxels into a set of small regions, which are defined to have the same radius in an exclusive manner. In brief, partition of the brain enumerates all the  $M$  voxels in an iterative way. In each iteration, we first select an unassigned voxel as a root to start a new region. We then compute the Euclidean distance between the root voxel and all the unassigned voxels, of which those voxels with distance smaller than  $R$  are assigned into this region. We iterate this process until all the voxels are assigned to one of the regions. Consequently, there is no overlap between defined regions. We denote the  $k$ -th region  $G_k$ . After this preprocessing step, we identify  $K$  regions, from which we aim to identify important regions for SZ discrimination.

Stage 1 of our algorithm is to prune insignificant regions from  $K$  pre-defined regions. We formulate our region selection algorithm by considering a regularized empirical risk minimization procedure with an  $L_0$ -norm regularization. Specifically, we attach a binary random variable  $z^k \in \{0, 1\}$  to all the voxels in region  $G_k$ :

$$\tilde{x} = x \odot Az, \quad z \in \{0, 1\}^K, \quad (1)$$

where  $z \in R^{K \times 1}$  denotes a binary mask for brain image  $x \in R^{M \times 1}$ ,  $\odot$  is an element-wise product, and  $A \in R^{M \times K}$  is an affiliation matrix we construct from the preprocessing step above, with element  $A_{j,k} = 1$  if voxel  $x^j$  is in region  $G_k$ , and 0 otherwise. For all the voxels in a region  $G_k$ , they share the same binary mask  $z^k$ , and  $k \in \{1, 2, \dots, K\}$ . This means if  $z^k$  is 0, all the voxels in region  $G_k$  will have a value of 0, otherwise the value of  $x^j$  is retained. In the sequel, we will discuss our method that can learn  $z$  from training set  $D$ , and we wish  $z^k$  takes

**FIGURE 1** Overall architecture of our method

value of 1 if  $G_k$  is an important region and 0 otherwise. In other words,  $\mathbf{z}$  is a measure of feature (region) importance that we wish to learn from data.

We regard  $\mathbf{z}$  as the feature importance weight for the prediction of DNN model  $h(\mathbf{x}; \boldsymbol{\theta})$  and learn  $\mathbf{z}$  by minimizing the following  $L_0$ -norm regularized loss function:

$$\begin{aligned} R(\boldsymbol{\theta}, \mathbf{z}) &= \frac{1}{N} \sum_{i=1}^N \mathcal{L}(h(\mathbf{x}_i \odot \mathbf{A}\mathbf{z}; \boldsymbol{\theta}), y_i) + \lambda \|\mathbf{z}\|_0 \\ &= \frac{1}{N} \sum_{i=1}^N \mathcal{L}(h(\mathbf{x}_i \odot \mathbf{A}\mathbf{z}; \boldsymbol{\theta}), y_i) + \lambda \sum_{k=1}^K \mathbf{1}_{\{z^k \neq 0\}}, \end{aligned} \quad (2)$$

where  $\mathcal{L}(\cdot)$  denotes the data loss over training data  $D$ , such as the cross-entropy loss for classification,  $\|\mathbf{z}\|_0$  is the  $L_0$ -norm that measures number of nonzero elements in  $\mathbf{z}$ ,  $\lambda$  is a regularization hyperparameter that balances between data loss and feature sparsity, and  $\mathbf{1}_{\{c\}}$  is an indicator function that is 1 if the condition  $c$  is satisfied, and 0 otherwise. Thus in Equation (2) we have two sets of parameters  $(\boldsymbol{\theta}, \mathbf{z})$ , where  $\boldsymbol{\theta}$  denotes the weights of neural network and  $\mathbf{z}$  denotes the importance weights attached to all small regions (i.e.,  $\mathbf{x} \odot \mathbf{A}\mathbf{z}$ ) for feature selection. We optimize  $\boldsymbol{\theta}, \mathbf{z}$  jointly to minimize the classification loss on training data. Due to the  $L_0$  regularization, the solution of  $\mathbf{z}$  is a sparse vector, which performs region/feature selection. The importance weights as captured by  $\mathbf{z}$  essentially measure the contribution of each region to the final classification loss as shown in Equation (2), thus can be used for interpretation. To optimize Equation (2), however, we note that both the first term and the second term of Equation (2) are not differentiable w.r.t.  $\mathbf{z}$ . Therefore, further approximations need to be considered.

We can approximate this optimization problem via an inequality from stochastic variational optimization (Bird, Kunze, & Barber, 2018). Specifically, given any function  $\mathcal{F}(\mathbf{z})$  and any distribution  $q(\mathbf{z})$ , the following inequality holds

$$\min_{\mathbf{z}} \mathcal{F}(\mathbf{z}) \leq \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} [\mathcal{F}(\mathbf{z})], \quad (3)$$

that is, the minimum of a function is upper bounded by the expectation of the function. With this result, we can derive an upper bound of Equation (2) as follows.

Since  $z^k, \forall k \in \{1, \dots, K\}$  is a binary random variable, we assume  $z^k$  is subject to a Bernoulli distribution with parameter  $\pi^k \in [0, 1]$ , that is,  $z^k \sim \text{Ber}(z; \pi^k)$ . Thus, we can upper bound  $\min_{\mathbf{z}} R(\boldsymbol{\theta}, \mathbf{z})$  by the expectation

$$\bar{R}(\boldsymbol{\theta}, \boldsymbol{\pi}) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\pi})} [\mathcal{L}(h(\mathbf{x}_i \odot \mathbf{A}\mathbf{z}; \boldsymbol{\theta}), y_i)] + \lambda \sum_{k=1}^K \pi^k, \quad (4)$$

Now the second term of the Equation (4) is differentiable w.r.t. the new model parameters  $\boldsymbol{\pi}$ . However, the first term is still problematic since the expectation over a large number of binary random variables  $\mathbf{z} \in \{0, 1\}^K$  is intractable, so is its gradient. To solve this problem, we adopt the hard-concrete estimator (Louizos, Welling, & Kingma, 2017). Specifically, the hard-concrete gradient estimator employs a reparameterization trick to approximate the original

optimization problem of Equation (4) by a close surrogate loss function

$$\begin{aligned} \hat{R}(\boldsymbol{\theta}, \log \boldsymbol{\alpha}) &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\mathbf{u} \sim \mathcal{U}(0,1)} [\mathcal{L}(h(\mathbf{x}_i \odot g(\mathbf{A}\mathbf{f}(\log \boldsymbol{\alpha}, \mathbf{u})); \boldsymbol{\theta}), y_i)] \\ &+ \lambda \sum_{k=1}^K \sigma \left( \log \alpha_k - \beta \log \frac{-\gamma}{\zeta} \right) = \mathcal{L}_D(\boldsymbol{\theta}, \log \boldsymbol{\alpha}) + \lambda \mathcal{L}_C(\log \boldsymbol{\alpha}), \end{aligned} \quad (5)$$

with

$$f(\log \alpha_k, u_k) = \sigma \left( \frac{\log u_k - \log(1-u_k) + \log \alpha_k}{\beta} \right) (\zeta - \gamma) + \gamma, \quad (6)$$

and

$$g(\cdot) = \min(1, \max(0, \cdot)) \quad (7)$$

where  $\sigma(t) = 1/(1 + \exp(-t))$  is the sigmoid function,  $\mathcal{L}_D$  measures how well the classifier fits to training data  $D$ ,  $\mathcal{L}_C$  measures the expected number of non-zeros in  $\mathbf{z}$ , and  $\beta = \frac{2}{3}$ ,  $\gamma = -0.1$ , and  $\zeta = 1.1$  are the typical parameters of the hard-concrete distribution. Function  $g(\cdot)$  is a hard-sigmoid function that bounds the stretched concrete distribution between 0 and 1. With this reparameterization, the surrogate loss function Equation (5) is differentiable w.r.t. its parameters.

After training, we learn  $\log \boldsymbol{\alpha}$  from the dataset  $D$ . At test time, we employ the following estimator to generate a sparse mask or feature importance weight:

$$\hat{\mathbf{z}} = \min \left( \mathbf{1}, \max \left( \mathbf{0}, \sigma \left( \frac{\log \boldsymbol{\alpha}}{\beta} \right) (\zeta - \gamma) + \gamma \right) \right), \quad (8)$$

which is the sample mean of  $\mathbf{z}$  under the hard-concrete distribution  $q(\mathbf{z} | \log \boldsymbol{\alpha})$ .

After we train the sparse DNN with the  $L_0$ -norm regularization, we get the trained neural network parameters  $\boldsymbol{\theta}$  and sparse mask  $\hat{\mathbf{z}} \in [0, 1]^K$  over all  $K$  regions, with element  $\hat{z}^k$  a continuous variable that represents the importance of region  $G_k$ . Because of the sparsity inducing property of the  $L_0$ -norm, many elements of learned  $\hat{\mathbf{z}}$  are pushed to zero, which are considered as unimportant regions and thus pruned from the model. The level of sparsity can be modulated by hyperparameter  $\lambda$ : the larger  $\lambda$  is, the sparser regions is identified, and *vis-a-versa*.

In Stage 2 of our algorithm, we can further improve the accuracy of the classifier by finetuning the DNN with the selected  $L$  regions from Stage 1 but without the  $L_0$ -norm regularization. To examine whether incorporating genetic features can improve the classification accuracy, we also concatenate the PGRS feature to the selected voxels as the input of the DNN classifier to finetune the classifier.

In our study, the training data consisted of 634 individuals (346 controls and 288 cases), which were equally partitioned into three subsets (each containing 33% of the samples). A nested three-fold cross validation was then implemented to identify the discriminating genetic and brain MRI features and construct a classification

model for SZ. The region radius  $R$  we used was 12 mm and the brain was partitioned into spherical regions sequentially (in an ascending order) based on voxel indices, that is, in each round we selected the first (as indicated by voxel index) unassigned voxel as the seed to generate a parcel. This led to a total of 1,111 regions with an average size of 387 voxels. In Stage 1 group selection and Stage 2 retraining, we used a DNN classifier with 2 fully connected layers of 200 and 16 neurons, respectively, and the rectified linear unit (ReLU) activation function. We performed grid search to find the best hyperparameters for our sparse DNN model. In Stage 1 group selection, we used the SGD optimizer with learning rates of 0.005 and 1 for model parameter  $\theta$  and  $\log\alpha$ , respectively. In Stage 2 retraining classifier, we used the Adam optimizer with learning rate of 0.005 for  $\theta$  and a weight decay of  $1e-5$ . After the sparse DNN was trained on the GMV features, the regions with nonzero  $\hat{z}$ s were considered as important regions for the SZ classification. The selected regions across three-fold cross validation were highlighted for model interpretation. In particular, we tuned hyperparameter  $\lambda$  to compare the classification performances with different levels of sparsity. In Stage 2 retraining, the selected voxel regions were fed into the classifier and could concatenate the PGRS feature to improve the classification accuracy. The model established in the training data was further evaluated on three external data sets: BSNIP, TOP, and HUBIN.

## 2.5 | ICA + linear SVM and elastic net regularization

To compare with sparse DNN, we also conducted classification using linear SVM with components extracted by ICA as input. ICA decomposes data into a linear combination of underlying components among which independence is maximized (Amari, 1998; Bell & Sejnowski, 1995). When applied to sMRI data, ICA essentially identifies maximally independent components, each including a weighted pattern of voxels with covarying gray matter patterns across samples (Xu et al., 2008). ICA has been widely used in the neuroimaging field, yielding meaningful and generalizable brain networks which are not well captured by anatomical atlas (Beckmann, DeLuca, Devlin, & Smith, 2005; Calhoun, Adali, Pearson, & Pekar, 2001; Gupta et al., 2015). In the current work, following the training and testing of the sparse DNN, we applied ICA on the GMV data for 67% of the training samples. The resulting components were then fed into linear SVM to obtain a classification model. This model was then assessed on the remaining 33% of the training samples for accuracy. Since the number of ICA components was a hyperparameter to be tuned, we repeated the above process with different component numbers. Echoing the sparse DNN experiments, we compared models with a low vs. high number of GMV components as predictors in terms of the classification performance. When genetic feature was further incorporated, PGRS was treated as an additional predictor, which was sent into linear SVM along with the GMV components. Note that genetic data were available only for TOP and HUBIN, such that only these two data sets were examined for imaging genetic based classification.

In addition, we conducted permutation tests to estimate the null error rates where classification models derived from permuted diagnosis labels were applied to the three validation data sets. The null error rates reflected the chance to make a correct guess regarding diagnostic identities and were used to contrast with the original results to validate the classification models presented in the current work.

Besides ICA + SVM, we also examined the performance of elastic-net regularization in classification, see Supporting Information for details.

## 2.6 | Voxelwise group difference

Finally, we conducted a voxelwise analysis of case-control differences in GMV using two-sample  $t$ -test. Voxels showing significant group differences were identified controlling for false discovery rate ( $q < 0.05$ ) (Benjamini & Hochberg, 1995). The inferred directions of changes in GMV were then compared with those inferred from DNN to assess the interpretability of the DNN features.

## 3 | RESULTS

The performance is summarized in Table 2. While we adjusted  $\lambda$  in Equation (2) to obtain models with different levels of sparsity (i.e., leading to 5–30 important regions in sparse DNN), we noted that the classification accuracy dropped significantly with sparsity lower than 5 regions, and relatively saturated around sparsity of 20 regions. Consequently, we reported results of 5 and 20 regions respectively, to show how sparsity impacted performance. And similarly, for ICA + SVM, we reported results of 5 and 20 components. It can be seen in Table 2 that, for both ICA and DNN approaches, lower error rates were achieved when 20 rather than only 5 brain regions/components served as predictors. When fewer brain regions were used to train the model, the mean error rate across three independent data sets was 34.64% for DNN and 35.38% for ICA, which were comparable, though in specific data sets discrepancies could be noted. When the classification model was allowed to incorporate more brain regions/components, the mean error rate across three data sets decreased to 31.02% for DNN models and 31.44% for ICA models. Specifically, the error rates were comparable between ICA and DNN in HUBIN, while DNN outperformed ICA in TOP (error rate improved by 3.66%, nine more subjects classified accurately) and ICA excelled in BSNIP (error rate improved by 1.78%, seven more subjects classified accurately).

When PGRS was further incorporated for classification, the DNN approach also showed slightly lower mean error rates across two validation data sets compared to ICA, as shown in in Table 2 (32.10% vs. 32.81% for 5 regions, and 28.99% vs. 32.06% for 20 regions). It was also noted that DNN yielded consistent improvement in accuracy across all the data sets compared to when only GMV features were used, either with 5 or 20 regions as predictors, where the decrease in

**TABLE 2** Summary of classification error rates

	sMRI			sMRI + PRS	
	TOP (255)	HUBIN (160)	BSNIP (394)	TOP (255)	HUBIN (160)
<b>DNN (5 regions)</b>					
EER1	35.69	33.08	32.99	32.94	28.13
EER2	34.90	36.25	34.26	33.33	33.13
EER3	34.90	36.25	33.50	32.55	32.5
EER mean	35.16	35.19	33.58	32.94	31.25
<b>ICA + SVM (5 ICs)</b>					
EER1	36.86	31.88	36.29	30.20	35.00
EER2	37.25	34.38	37.31	30.59	35.63
EER3	34.90	32.50	37.06	29.80	35.63
EER mean	36.34	32.92	36.89	30.20	35.42
<b>Permutation (5 ICs)</b>					
EER1	39.61	52.50	47.21	39.61	52.50
EER2	39.61	52.50	47.21	39.61	52.50
EER3	39.61	52.50	47.21	39.61	52.50
EER mean	39.61	52.50	47.21	39.61	52.50
<b>DNN (20 regions)</b>					
EER1	30.59	28.13	30.96	30.65	26.27
EER2	30.98	32.50	31.73	27.75	27.25
EER3	33.33	28.75	32.23	32.26	28.24
EER mean	31.63	29.79	31.64	30.22	27.75
<b>ICA + SVM (20 ICs)</b>					
EER1	33.33	27.50	29.95	32.94	29.38
EER2	39.22	31.25	31.73	35.29	33.75
EER3	33.33	28.75	27.92	30.98	30.00
EER mean	35.29	29.17	29.86	33.07	31.04
<b>Permutation (20 ICs)</b>					
EER1	50.59	48.75	51.52	50.98	50.00
EER2	45.10	56.88	44.67	45.10	55.00
EER3	44.31	49.38	46.95	44.31	47.50
EER mean	46.67	51.67	47.72	46.80	50.83

error rate ranged from 1.41% to 3.94%. In contrast, when ICA components were combined with PGRS for classification, the error rate did not always decrease. Among all the tests, the lowest error rate (27.75%) was observed in HUBIN, when the DNN classification model used 20 brain regions plus the PGRS.

The null error rates as shown in Table 2 were averaged across 100 permutation runs, which were consistently higher than the true error rates. With a low sparsity of 5 regions, SVM failed to converge to a decision boundary for either GMV features or GMV + PGRS features, such that all the subjects were classified as controls. Consequently, the null error rates simply reflected the ratio of controls in the TOP, HUBIN and BSNIP data. With a high sparsity of 20 regions, SVM was able to converge on permuted diagnosis labels. The resulting average null error rates ranged between 46.67% and 51.67%, which were substantially higher than the true error rates obtained from the original data.

The sizes of identified DL regions ranged from 277 to 814 voxels, with a mean of 488 voxels. In contrast, the independent components, even when thresholded at  $|z| > 3$  for top voxels, showed sizes ranging from 555 to 11,668 voxels, with a mean of 6,409 voxels. With high level of sparsity (5 regions), the DNN model used a total of 2,379 voxels (0.6% of the selected gray matter voxels) for classification while ~25,000 top voxels were used by ICA. The number of covered voxels increased to ~8,500 for DNN and ~123,000 for ICA (top voxels) when 20 regions/components were to be selected as predictors. The brain regions identified by DNN are summarized in Table 3 (5 regions) and Table 4 (20 regions), and Figures 2 and 3 show the spatial maps of individual regions. Note that only the regions identified in all three folds are listed. When 5 regions were to be selected as predictors, the three folds consistently identified the same 5 regions, spanning inferior, middle, and superior frontal gyrus, superior temporal gyrus, as well as cerebellum. The weights, or direction of

effects, were all positive, indicating higher GMV in controls compared to cases. When 20 regions were to be selected, variations were noted across folds, such that 13 brain regions were consistently identified. Compared to those covered by 5 regions, cuneus, precuneus, medial frontal gyrus, and paracentral lobule were further determined to be informative and included for classification. Most of the regions showed positive weights with higher GMV observed in controls compared to cases. Meanwhile, negative weights were observed for region 27 and 45.

For comparison, Figure 4 presents the spatial map of the voxels showing significant case-control differences in the voxelwise analysis, with the score reflecting the two-sample *t*-test *p*-value ( $-10\log(p)$ ). Most of the brain regions identified by sparse DNN, either with 5-region or 20-region sparsity, showed large overlap with the voxelwise analysis, as summarized in Table 5. With a high-sparsity setting, 4 out of 5 identified DNN regions strongly overlapped with the voxelwise results (overlap ratio > 0.8), while the remaining region showed an overlap ratio of 0.62. With a low-sparsity setting, 9 out of 13 identified DNN regions showed an overlap ratio > 0.5, with the remaining 4 regions showing an overlap ratio of 0.18, 0.34, 0.43 and 0.38, respectively. And the directions of effects inferred from the interpretable DNN model were overall highly consistent with those inferred from the original features. The only exception was region 27 identified in the 20-region model, which showed a negative weight in DNN while the voxelwise analysis indicated higher GMV in controls than cases.

## 4 | DISCUSSION

An interpretable sparse DNN approach was proposed for application to medical data analysis and its capability was examined on a large and heterogeneous SZ data set. The results confirmed that the proposed approach yielded reasonable classification accuracies, could identify meaningful brain regions, and the interpretation of these brain regions was consistent with that directly inferred from original features. Particularly, the proposed model appeared to more effectively fuse imaging and genetic features for classification compared to ICA + SVM, holding potential for data fusion.

The DNN models reliably generalized to data collected at different sites, with reasonable classification accuracies compared to ICA + SVM. Permutation tests yielded substantially higher error rates,

suggesting that the observed level of accuracy was not likely achieved by chance. And the generalizability indicates that the classification models are not vulnerable to scanning protocol, recruiting criteria, ethnicity influence, medication history, and so forth. Regarding performance, both DNN and ICA + SVM approaches presented higher accuracies when more brain regions/components served as predictors, with error rates being 31.03% and 31.86%, respectively. While many machine learning methods have been proposed for the same purpose, a direct comparison is still missing, partly due to the unavailability of a benchmark data set. Sample heterogeneity, data collection, and preprocessing could all affect classification, which makes it difficult to compare accuracies achieved from different data sets. As pointed out by Cai et al. (2020), high classification accuracies are more likely to show in single-site data, which need to be interpreted cautiously. Echoing this, more recent work has made efforts to assess proposed methods in large multi-study cohorts. For instance, the work by Yan et al. (2019) and Zeng et al. (2018) proposed and applied deep learning methods to large multi-study functional MRI data and obtained classification accuracies of 80% and 81%, respectively, for SZ with a leave-site-out set up. Another work by Cai et al. (2020) tested combinations of ICA-based functional connectivity features with various classical classification models on a large multi-study data set and obtained an accuracy of 70% for between-site generalizability. The error rate obtained in our work is comparable to the literature with performance evaluated under a similar scenario, indicating complex heterogeneity of patients with SZ. Increasing sample size of the training data and incorporating other data modalities promise further improvement.

The proposed approach highlights a sparsity constraint, allowing the DNN model to achieve comparable performances with ICA while leveraging 10 time less voxels for classification. A trade-off is also noted between explained variance and interpretability of identified features. In general, a low level of sparsity allows more features to be admitted into the classification model, which however results in more variance across samples. As shown in the current work, when a higher level of sparsity was enforced, the same 5 regions were identified across 3 folds. In contrast, with a lower sparsity, 13 out of 20 regions were consistently identified, although the latter explained more variance and yielded higher classification accuracies. It should be pointed out that, increasing the predictors from 5 to 20 regions resulted in a decrease of ~4% in error rate, which was indeed not profound considering that the 20 regions incorporated three times more voxels for

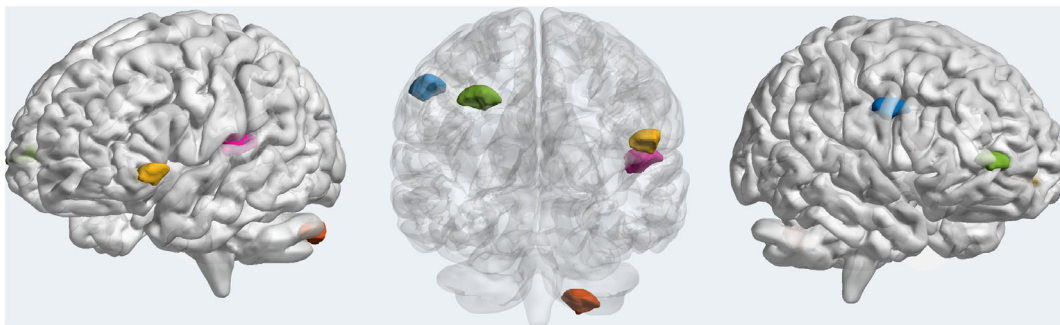
**TABLE 3** Summary of the five important brain regions identified by sparse DNN

Region	Area	Brodmann area	Volume (cc)	MNI (x, y, z)	Direction of effects
DL87	Uvula (cerebellum)	N/A	0.7/0.0	(-18, -81, -33)/(0, 0, 0)	+
DL382	Inferior frontal gyrus	47	1.9/0.0	(-54, 30, 0)/(0, 0, 0)	+
DL493	Superior frontal gyrus	10	0.0/1.2	(0, 0, 0)/(27, 60, 9)	+
	Middle frontal gyrus	10	0.0/0.9	(0, 0, 0)/(34.5, 57, 9)	+
DL555	Superior temporal gyrus	13, 22, 41	1.0/0.0	(-45, -30, 15)/(0, 0, 0)	+
DL775	Inferior frontal gyrus	9	0.0/1.0	(0, 0, 0)/(57, 12, 36)	+



**TABLE 4** Summary of the 13 important brain regions identified by sparse DNN

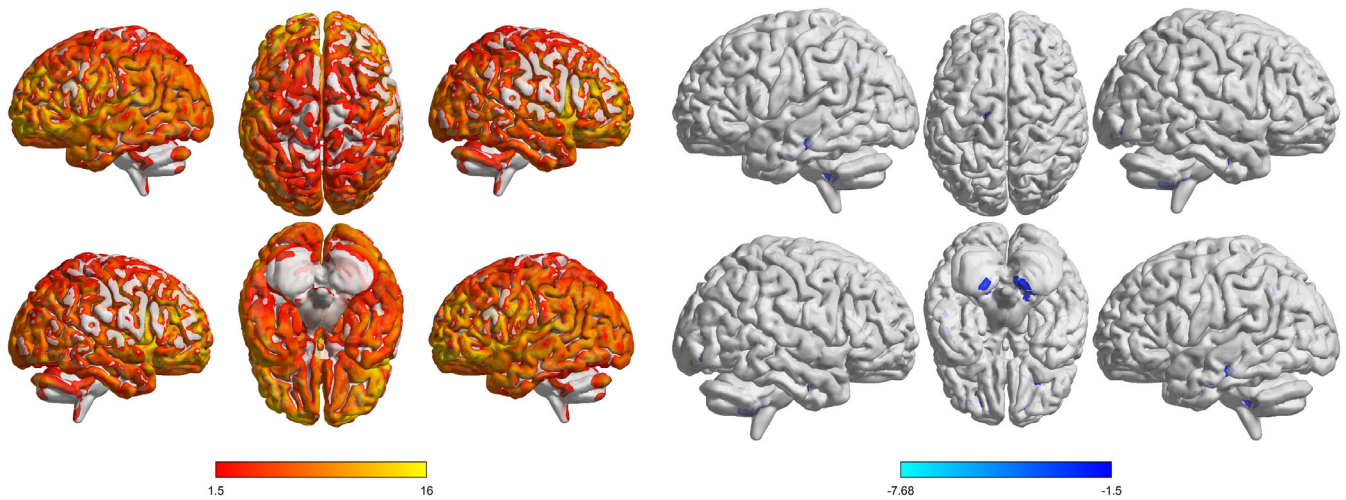
Region	Area	Brodmann area	Volume (cc)	MNI (x, y, z)	Direction of effects
DL2	Inferior semi-lunar lobule	N/A	0.1/0.0	(-7.5, -60, -54)/(0, 0, 0)	+
DL27	Cerebellar tonsil	N/A	1.4/0.0	(-15, -55.5, -43.5)/(0, 0, 0)	-
DL45	Cerebellar tonsil	N/A	0.7/0.0	(-12, -55.5, -40.5)/(0, 0, 0)	-
DL172	Superior temporal gyrus	38	0.0/1.0	(0, 0, 0)/(48, 22.5, -19.5)	+
DL260	Middle frontal gyrus	11	0.9/0.0	(-37.5, 40.5, -10.5)/(0, 0, 0)	+
DL509	Inferior frontal gyrus	13, 47	1.3/0.0	(-42, 25.5, 10.5)/(0, 0, 0)	+
DL599	Cuneus	18, 19	0.0/1.0	(0, 0, 0)/(18, -88.5, 19.5)	+
DL691	Middle frontal gyrus	10, 46	1.2/0.0	(-34.5, 46.5, 27)/(0, 0, 0)	+
DL805	Middle frontal gyrus	9	2.0/0.0	(-45, 28.5, 39)/(0, 0, 0)	+
DL846	Precuneus	7, 19	0.0/1.0	(0, 0, 0)/(30, -66, 42)	+
DL1008	Medial frontal gyrus	6	0.0/1.3	(0, 0, 0)/(7.5, -4.5, 63)	+
DL1017	Paracentral lobule	4, 5, 6	0.2/1.7	(-1.5, -40.5, 61.5)/(4.5, -37.5, 64.5)	+
DL1039	Middle frontal gyrus	6	1.3/0.0	(-21, 9, 67.5)/(0, 0, 0)	+

**FIGURE 2** Spatial maps of the five schizophrenia-discriminating regions identified by sparse DNN**FIGURE 3** Spatial maps of the 13 schizophrenia-discriminating regions identified by sparse DNN

classification. In other words, although GMV abnormalities are widely distributed across the brain in SZ, the majority of the variance can be captured by the identified five distinct regions which only covered ~0.6% of the selected gray matter voxels. The samples missed in the classification, or missing variance, likely call for a larger training data set to allow better capturing heterogeneity, as well as for information

from other data modalities, rather than simply adding more features from the sMRI modality.

SZ is a complex disorder, where genetic and environmental factors interact with each other to affect brain structure and function which ultimately manifest into clinical symptoms. With so many factors involved in the pathology of SZ, it is expected that multiple data



**FIGURE 4** Spatial maps of the voxels showing significant case-control differences in the voxelwise analysis. The positive/negative scores reflect higher/lower GMV in controls compared to cases

**TABLE 5** Overlap between important regions of sparse DNN and voxelwise analysis

Region	Region size (# of voxels)	Overlap (# of voxels)	Overlap ratio
<b>5-region</b>			
DL87	352	323	0.92
DL382	475	392	0.83
DL493	625	390	0.62
DL555	523	502	0.96
DL775	404	365	0.90
<b>20-region</b>			
DL2	814	735	0.90
DL27	565	99	0.18
DL45	277	95	0.34
DL172	315	293	0.93
DL260	559	367	0.66
DL509	494	262	0.53
DL599	372	186	0.50
DL691	514	409	0.80
DL805	569	245	0.43
DL846	441	250	0.57
DL1008	462	176	0.38
DL1017	568	363	0.64
DL1039	492	421	0.86

modalities need to be integrated to fully characterize the disorder. This also applies to classification, which should capitalize on data fusion to extract complementary information from different modalities. The proposed model holds promise for this purpose. In all the tested scenarios, the DNN approach effectively fused GMV and PGRS features to yield improved classification accuracies, indicating that the model reliably extracted SZ-related variance in PGRS that was not captured by GMV. In contrast, no consistent improvement was noted for ICA + SVM when PGRS and brain components were directly fed

into linear SVM for classification training, which is consistent with a previous study (Doan et al., 2017). The results appear to lend support that nonlinear models excel in delineating the relationships across different modalities in hidden layers and robustly capturing complementary variance that is related to the trait of interest.

The brain regions identified by sparse DNN are overall group-discriminating as indicated by the overlap with voxelwise analysis, and well documented in SZ studies. With high sparsity, 5 brain regions were consistently identified across 3 folds, as listed in Table 3,

highlighting frontal gyrus, superior temporal gyrus, and cerebellum. All the five regions presented positive weights, indicating higher GMV in controls compared to patients, which was consistent with the results of two-sample *t*-tests on original GMV features. SZ-related gray matter reduction has been widely observed in temporal and frontal regions. A longitudinal study by Thompson et al. revealed accelerated gray matter loss in early-onset SZ, with earliest deficits found in parietal regions and progressing anteriorly into temporal and prefrontal regions over 5 years (Thompson et al., 2001). The identified frontal and temporal brain regions have also been identified for SZ-related reduction in a comprehensive study on gray matter volume in psychosis using the BSNIP cohort (Ivleva et al., 2013), as well as for showing the most significant cortical thinning in patients with SZ in the ENIGMA study (van Erp et al., 2018). The role of cerebellum in SZ has been revised in recent years, where accumulating evidence suggests that cerebellum is also involved in cognitive functions and cerebellar abnormalities are noted in SZ (Andreasen & Pierson, 2008; Moberget et al., 2018). Gray matter loss around the identified cerebellar region has also been reported previously (Farrow, Whitford, Williams, Gomes, & Harris, 2005). While the three aforementioned multi-site classification studies (i.e., Cai et al., 2020; Yan et al., 2019; Zeng et al., 2018) all used functional MRI features, no direct comparison could be made regarding identified biomarkers. However, consistency was noted for disrupted brain regions between the current work and Cai et al. (2020), including superior and middle frontal gyrus, superior temporal gyrus, cuneus, precuneus, as well as paracentral lobule. Overall, the results validate the proposed method for identifying interpretable biological features relevant to selected traits.

With low sparsity, 13 brain regions were consistently identified by DNN across 3 folds, as listed in Table 4. In addition to frontal, temporal and cerebellar regions discussed above, parietal regions including cuneus, precuneus and paracentral lobule were highlighted. As implicated in Thompson et al, while temporal and prefrontal gray matter loss were characteristic of adult SZ, parietal regions were noted for earliest gray matter loss which was faster in younger patients with SZ (Thompson et al., 2001). The identified parietal regions also echoed the BSNIP findings to show higher GMV in controls compared to patients (Ivleva et al., 2013). Overall, it is reasonable that DNN prioritized to select temporal and frontal regions for classification when high sparsity was enforced, which aligns with the notion that gray matter loss in these regions characterizes adult SZ. In the meantime, when a lower sparsity was enforced, parietal abnormalities were the first priority to be added as additional predictors which offered complementary variance. Among the 13 regions, region 27 (cerebellar tonsil) was the only feature whose DNN weights did not coincide with the inference drawn from original GMV features. It was noted that the voxels in region 27 showed modest case-control differences compared to voxels in other identified brain regions (Figure 4). We suspect the selection of region 27 by DNN might be driven by some hidden properties rather than group differences, which explains the inconsistency in interpretation between DNN and two-sample *t*-tests.

One observation is that the 5 regions selected with low sparsity were not a subset of those 13 regions selected with high sparsity.

Sparse DNN is designed to optimize classification accuracy for specified hyperparameters. It is possible that the best performance is yielded by different sets of brain regions at different levels of sparsity. Meanwhile, it is noted that Brodmann Area 9, 10, 13, and 47 highlighted under 5-region sparsity were also identified as important regions with 20-region sparsity, showing a certain level of overlap in corresponding anatomical regions. Overall, we expect that the data-driven learning process may pick up slightly different small brain regions with different hyperparameters, which however has no dramatic impact on interpretation.

One limitation of the current work is that we did not extensively investigate the impact of brain partition. We did explore other partition strategies, including using a descending or random rather than ascending order for seed voxel selection as used in the main analysis. The resulting classification accuracies were comparable or slightly lower than those observed with the ascending partition, indicating that selection order has no dramatic impact on performance. Meanwhile, we did not investigate how the radius of brain regions would affect the performance. We assumed the brain regions to be spherical, which may not align with the optimal partition. It deserves further exploration whether the performance may benefit from brain atlas-based partition (such as Yeo atlas (Yeo et al., 2011)). This topic will be investigated in the future work. Besides, likely due to the limited sample size, the DNN performance saturated at 2 hidden layers. It remains a question how the performance would scale with increasing sample size. This awaits investigation when more data become available. Furthermore, while the DNN approach holds promise for data fusion, its capability of integrating multiple high-dimensional imaging modalities was not examined in the current work, given that incorporating another modality would further reduce the sample size. This will also be part of our future work. Finally, it is not clear whether the proposed algorithm would yield a comparable performance in first episode patients, which might be more challenging given the effects of illness chronicity and the medication exposure. Unfortunately, currently we only have data of no more than 20 first episode patients which are not sufficiently powered for a comprehensive evaluation of any algorithm. We will research for available data resources to answer this question in our future work.

In summary, to the best of our knowledge, this is the first study of DNN application to sMRI and genetic features for SZ case-control classification with generalizability assessed in a large and multi-study cohort. An interpretable sparse DNN approach was first proposed to allow identifying, refining and interpreting features used in classification. The results indicate that the new approach yielded reasonable classification performances, highly sparse and interpretable classification features, as well as potential for data fusion. Collectively, the current work validates the application of the proposed approach to SZ classification. We hope our future work could be extended to prediction of clinical outcome, for example, treatment response, which can be either binary or continuous variables. Particularly, we would like to examine whether any of the identified group-discriminating brain regions contribute to stratifying patients in terms of treatment response, which may not only provide more insights into underlying neurobiology of SZ, but also facilitate precision medicine.

## ACKNOWLEDGMENTS

This project was funded by the National Institutes of Health (P20GM103472, P30GM122734, R01EB005846, 1R01EB006841, R01MH106655, 5R01MH094524, U24 RR021992, U24 RR025736-01, U01 MH097435, R01 MH084803, R01 EB020062), National Science Foundation (1539067, 1636893, 1734853), Research Council of Norway (RCN#223273), K. G. Jebsen Stiftelsen and South-East Norway Health Authority.

## CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

## DATA AVAILABILITY STATEMENT

The MCIC and COBRE data are available through COINS (<https://coins.mrn.org>). The NU imaging data can be accessed through SchizConnect (<http://schizconnect.org/>) and the BSNIP imaging data through NIMH Data Archive (<https://nda.nih.gov/>). Request of access to other data should be addressed to the individual principal investigator. The code of the proposed algorithm is publicly available through GitHub ([https://github.com/lxuniverse/sparse\\_fmri](https://github.com/lxuniverse/sparse_fmri)).

## ORCID

Jiayu Chen  <https://orcid.org/0000-0001-5059-372X>

Lei Wang  <https://orcid.org/0000-0003-3870-3388>

Lars T. Westlye  <https://orcid.org/0000-0001-8644-956X>

Jingyu Liu  <https://orcid.org/0000-0002-1724-7523>

## REFERENCES

- Althuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., ... Consortium, 1000 Genomes Project, (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491, 56–65.
- Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10, 251–276.
- Andreasen, N. C., & Pierson, R. (2008). The role of the cerebellum in schizophrenia. *Biological Psychiatry*, 64, 81–88.
- Arbabshirani, M. R., Kiehl, K. A., Pearlson, G. D., & Calhoun, V. D. (2013). Classification of schizophrenia patients based on resting-state functional network connectivity. *Frontiers in Neuroscience-Switz*, 7.
- Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, 26, 839–851.
- Beckmann, C. F., DeLuca, M., Devlin, J. T., & Smith, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 360, 1001–1013.
- Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7, 1129–1159.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate—A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B Statistical Methodology*, 57, 289–300.
- Bird, T., Kunze, J., & Barber, D. (2018). Stochastic variational optimization. *arXiv preprint arXiv:1809.04855*, Vol.
- Cai, X. L., Xie, D. J., Madsen, K. H., Wang, Y. M., Bögemann, S. A., Cheung, E. F. C., ... Chan, R. C. K. (2020). Generalizability of machine learning for classification of schizophrenia based on resting-state functional MRI data. *Human Brain Mapping*, 41, 172–184.
- Calhoun, V. D., Adali, T., Pearlson, G. D., & Pekar, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, 14, 140–151.
- Casey, B. J., Craddock, N., Cuthbert, B. N., Hyman, S. E., Lee, F. S., & Ressler, K. J. (2013). DSM-5 and RDoC: Progress in psychiatry research? *Nature Reviews. Neuroscience*, 14, 810–814.
- Chen, J., Calhoun, V. D., Lin, D., Perrone-Bizzozero, N. I., Bustillo, J. R., Pearlson, G. D., ... Liu, J. (2019). Shared genetic risk of schizophrenia and gray matter reduction in 6p22.1. *Schizophrenia Bulletin*, 45, 222–232.
- Chen, J., Calhoun, V. D., Pearlson, G. D., Perrone-Bizzozero, N., Sui, J., Turner, J. A., ... Liu, J. (2013). Guided exploration of genomic risk for gray matter abnormalities in schizophrenia using parallel independent component analysis with reference. *NeuroImage*, 83, 384–396.
- Chen, J., Liu, J., & Calhoun, V. D. (2019). Translational potential of neuroimaging genomic analyses to diagnosis and treatment in mental disorders. *Proceedings of the IEEE*, 107, 912–927.
- Cuthbert, B. N. (2014). The RDoC framework: Facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry*, 13, 28–35.
- Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine*, 11, 126.
- Delaneau, O., Marchini, J., & Zagury, J. F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9, 179–181.
- Doan, N. T., Kaufmann, T., Bettella, F., Jørgensen, K. N., Brandt, C. L., Moberget, T., ... Westlye, L. T. (2017). Distinct multivariate brain morphological patterns and their added predictive value with cognitive and polygenic risk scores in mental disorders. *NeuroImage: Clinical*, 15, 719–731.
- Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: Polygenic Risk Score software. *Bioinformatics*, 31(9), 1466–1468.
- Farrow, T. F. D., Whitford, T. J., Williams, L. M., Gomes, L., & Harris, A. W. F. (2005). Diagnosis-related regional gray matter loss over two years in first episode schizophrenia and bipolar disorder. *Biological Psychiatry*, 58, 713–723.
- Frank, J., Lang, M., Witt, S. H., Strohmaier, J., Rujescu, D., Cichon, S., ... Rietschel, M. (2015). Identification of increased genetic risk scores for schizophrenia in treatment-resistant patients (vol 20, pg 150, 2015). *Molecular Psychiatry*, 20, 913–913.
- Garrity, A. G., Pearlson, G. D., McKiernan, K., Lloyd, D., Kiehl, K. A., & Calhoun, V. D. (2007). Aberrant “default mode” functional connectivity in schizophrenia. *American Journal of Psychiatry*, 164, 450–457.
- Gupta, C. N., Calhoun, V. D., Rachakonda, S., Chen, J., Patel, V., Liu, J., ... Turner, J. A. (2015). Patterns of gray matter abnormalities in schizophrenia based on an international mega-analysis. *Schizophrenia Bulletin*, 41, 1133–1142.
- Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., ... Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167, 748–751.
- Insel, T. R. (2014). The NIMH research domain criteria (RDoC) project: Precision medicine for psychiatry. *American Journal of Psychiatry*, 171, 395–397.
- Insel, T. R., & Cuthbert, B. N. (2015). Brain disorders? Precisely. *Science*, 348, 499–500.
- Ivleva, E. I., Bidesi, A. S., Keshavan, M. S., Pearlson, G. D., Meda, S. A., Dodig, D., ... Tamminga, C. A. (2013). Gray matter volume as an intermediate phenotype for psychosis: Bipolar-schizophrenia network on intermediate phenotypes (B-SNIP). *The American Journal of Psychiatry*, 170, 1285–1296.
- Kelly, S., Jahanshad, N., Zalesky, A., Kochunov, P., Agartz, I., Alloza, C., ... Donohoe, G. (2018). Widespread white matter microstructural differences in schizophrenia across 4322 individuals: Results from the ENIGMA schizophrenia DTI working group. *Molecular Psychiatry*, 23, 1261–1269.

- Liang, S., Li, Y., Zhang, Z., Kong, X., Wang, Q., Deng, W., ... Li, T. (2019). Classification of first-episode Schizophrenia Using multimodal brain features: A combined structural and diffusion imaging study. *Schizophrenia Bulletin*, *45*(3), 591–599.
- Lin, D., Chen, J., Ehrlich, S., Bustillo, J. R., Perrone-Bizzozero, N., Walton, E., ... Liu, J. (2018). Cross-tissue exploration of genetic and epigenetic effects on brain gray matter in Schizophrenia. *Schizophrenia Bulletin*, *44*(2), 443–452.
- Louizos, C., Welling, M., & Kingma, D. P. (2017). Learning sparse neural networks through  $L_0$  regularization, *arXiv preprint arXiv: 1712.01312*, Vol.
- Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews. Genetics*, *11*, 499–511.
- McGrath, J., Saha, S., Chant, D., & Welham, J. (2008). Schizophrenia: A concise overview of incidence, prevalence, and mortality. *Epidemiologic Reviews*, *30*, 67–76.
- Moberget, T., Doan, N. T., Alnæs, D., Kaufmann, T., Córdova-Palomera, A., Lagerberg, T. V., ... Westlye, L. T. (2018). Cerebellar volume and cerebellocerebral structural covariance in schizophrenia: a multisite mega-analysis of 983 patients and 1349 healthy controls. *Molecular Psychiatry*, *23*(6), 1512–1520.
- Nieuwenhuis, M., van Haren, N. E. M., Hulshoff Pol, H. E., Cahn, W., Kahn, R. S., & Schnack, H. G. (2012). Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. *NeuroImage*, *61*, 606–612.
- Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., ... Calhoun, V. D. (2014). Deep learning for neuroimaging: A validation study. *Frontiers in Neuroscience-Switz*, *8*.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., ... Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559–575.
- Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K. H., Holmans, P. A., & Consortium, Psychiat Genomics. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*, 421.
- Segall, J. M., Turner, J. A., van Erp, T. G. M., White, T., Bockholt, H. J., Gollub, R. L., ... Calhoun, V. D. (2009). Voxel-based morphometric multisite collaborative study on schizophrenia. *Schizophrenia Bulletin*, *35*, 82–95.
- Skåtun, K. C., Kaufmann, T., Doan, N. T., Alnæs, D., Córdova-Palomera, A., Jönsson, ... Westlye, L. T. (2017). Consistent functional connectivity alterations in schizophrenia spectrum disorder: A multisite study. *Schizophrenia Bulletin*, *43*(4), 914–924.
- Thompson, P. M., Vidal, C., Giedd, J. N., Gochman, P., Blumenthal, J., Nicolson, R., ... Rapoport, J. L. (2001). Mapping adolescent brain change reveals dynamic wave of accelerated gray matter loss in very early-onset schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 11650–11655.
- Thomas Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., ... Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165.
- van Erp, T. G. M., Walton, E., Hibar, D. P., Schmaal, L., Jiang, W., Glahn, D. C., ... Orhan, F. (2018). Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the enhancing neuro imaging genetics through meta analysis (ENIGMA) consortium. *Biological Psychiatry*, *84*, 644–654.
- van Erp, T. G. M., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., ... ENIGMA, Schizophrenia Working. (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular Psychiatry*, *21*(4), 547–553.
- Vassos, E., di Forti, M., Coleman, J., Iyegbe, C., Prata, D., Euesden, J., ... Breen, G. (2017). An examination of polygenic score risk prediction in individuals with first-episode psychosis. *Biological Psychiatry*, *81*, 470–477.
- Woodward, N. D., Rogers, B., & Heckers, S. (2011). Functional resting-state networks are differentially affected in schizophrenia. *Schizophrenia Research*, *130*, 86–93.
- Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., ... Working, M. D. D. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, *50*(5), 668–681.
- Wrobel, J., Martin, M. L., Bakshi, R., Calabresi, P. A., Elliot, M., Roalf, D., ... Goldsmith, J. (2020). Intensity warping for multisite MRI harmonization. *NeuroImage*, *223*.
- Xu, L., Groth, K. M., Pearlson, G., Schretlen, D. J., & Calhoun, V. D. (2009). Source-based morphometry: The use of independent component analysis to identify gray matter differences with application to schizophrenia. *Human Brain Mapping*, *30*(3), 711–724.
- Yan, W., Calhoun, V., Song, M., Cui, Yu., Yan, H., Liu, S., ... Sui, J. (2019). Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site fMRI data. *EBioMedicine*, *47*, 543–552.
- Zeng, L. L., Wang, H., Hu, P., Yang, B., Pu, W., Shen, H., ... Hu, D. (2018). Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. *eBioMedicine*, *30*, 74–85.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Chen J, Li X, Calhoun VD, et al. Sparse deep neural networks on imaging genetics for schizophrenia case-control classification. *Hum Brain Mapp*. 2021;42: 2556–2568. <https://doi.org/10.1002/hbm.25387>