



OPEN

Detecting schizophrenia with 3D structural brain MRI using deep learning

Junhao Zhang¹, Vishwanatha M. Rao¹, Ye Tian¹, Yanting Yang¹, Nicolas Acosta¹, Zihan Wan², Pin-Yu Lee¹, Chloe Zhang⁸, Lawrence S. Kegeles^{3,5}, Scott A. Small^{6,7} & Jia Guo^{3,4}✉

Schizophrenia is a chronic neuropsychiatric disorder that causes distinct structural alterations within the brain. We hypothesize that deep learning applied to a structural neuroimaging dataset could detect disease-related alteration and improve classification and diagnostic accuracy. We tested this hypothesis using a single, widely available, and conventional T1-weighted MRI scan, from which we extracted the 3D whole-brain structure using standard post-processing methods. A deep learning model was then developed, optimized, and evaluated on three open datasets with T1-weighted MRI scans of patients with schizophrenia. Our proposed model outperformed the benchmark model, which was also trained with structural MR images using a 3D CNN architecture. Our model is capable of almost perfectly (area under the ROC curve = 0.987) distinguishing schizophrenia patients from healthy controls on unseen structural MRI scans. Regional analysis localized subcortical regions and ventricles as the most predictive brain regions. Subcortical structures serve a pivotal role in cognitive, affective, and social functions in humans, and structural abnormalities of these regions have been associated with schizophrenia. Our finding corroborates that schizophrenia is associated with widespread alterations in subcortical brain structure and the subcortical structural information provides prominent features in diagnostic classification. Together, these results further demonstrate the potential of deep learning to improve schizophrenia diagnosis and identify its structural neuroimaging signatures from a single, standard T1-weighted brain MRI.

Schizophrenia is a progressive neuropsychiatric disorder that is characterized by structural changes within the brain. Recent findings from a large meta-analysis suggest that schizophrenia is associated with gray matter reductions across multiple subcortical regions including the hippocampus, amygdala, caudate, and thalamus, with structural changes in shape within those regions supporting changes in functional brain networks¹. In addition to the altered shape of such brain structures, schizophrenia is also associated with significantly greater mean volume variability of the temporal cortex, thalamus, putamen, and third ventricle². Other studies also affirm the enlargement of ventricles in schizophrenia^{3,4}. While gray matter reductions are most consistently reported in the subcortical regions, reductions have also been identified in areas such as the prefrontal, temporal, cingulate, and cerebellar cortices^{5,6}. Loss of gray matter volume has been shown to not only mark the onset of schizophrenia but also progress alongside the illness⁷.

Despite these documented changes, accurate and rapid detection of schizophrenia remains a pressing problem; previous studies are limited to only characterizing structural abnormalities at a group level, with no concrete method to make individual diagnoses at a subject level. Additionally, the diagnosis of schizophrenia based on DSM-5 criteria is costly both in terms of time and resources, without ensuring objectivity. Therefore, it is imperative to develop an objective screening tool to diagnose schizophrenia and potentially improve patient prognosis by allowing for earlier intervention.

Various attempts have been proposed to take advantage of the structural alterations present in schizophrenia for classification using neuroimaging data. Machine learning algorithms have historically presented the ability

¹Department of Biomedical Engineering, Columbia University, New York, NY, USA. ²Department of Applied Mathematics, Columbia University, New York, NY, USA. ³Department of Psychiatry, Columbia University, New York, NY, USA. ⁴The Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA. ⁵Department of Radiology, Columbia University, New York, NY, USA. ⁶Department of Neurology, Radiology, and Psychiatry, Columbia University, New York, NY, USA. ⁷The Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, NY, USA. ⁸Jericho High School, Jericho, NY, USA. ✉email: jg3400@columbia.edu

to classify psychiatric disorders in this manner^{8,9}. In particular, the support vector machine (SVM), a supervised learning algorithm able to capture non-linear patterns in high-dimensional data, has been most prevalent in schizophrenia classification. Other popular machine learning algorithms for schizophrenia classification include multivariate pattern analysis, linear discriminant analysis, and random forest^{8,10}. While standard machine learning approaches have demonstrated compelling results, their performance highly depends on the validity of manually extracted features⁸. Such features are traditionally extracted based on a combination of previously known disease characteristics and automatic feature selection algorithms¹¹. These features may not completely encode the subtle neurological differences associated with schizophrenia; alternatively, they may encode too much unnecessary information requiring additional feature reduction¹².

Deep learning has recently emerged as a new approach demonstrating superior performance over standard machine learning algorithms to classify neurological diseases using structural MRI data. Specifically, Convolutional Neural Networks (CNNs) can learn and encode the significant features necessary for classification and have become popular in medical image analysis^{13–15}. This property makes CNNs uniquely suited to tasks like schizophrenia classification, where the specific features selected can dramatically impact model performance. Some studies have already demonstrated the utility of CNNs for schizophrenia classification. Oh J et al.¹⁶ achieved an impressive state-of-the-art performance (area under the ROC curve = 0.96) using 3D CNN for schizophrenia classification based on structural MRI data and was thus compared to as the benchmark model. Nevertheless, they struggled to generalize well on an unseen private dataset. Their inconsistent performance may be attributed to the dataset and patient variability as well as certain pre-processing choices, such as the inclusion of whole head as opposed to whole-brain MRI data and severe downsampling. Moreover, their region of interest analysis was limited and did not investigate brain structures in depth to inform specific changes in structural features associated with schizophrenia. Hu et al. combined structural and diffusion MRI scans for schizophrenia classification and found that 3D CNN models could outperform 2D pre-trained CNN models as well as multiple standard machine learning algorithms like SVM. Despite this, their best 3D model only reached the area under the ROC curve of 0.84¹⁷. As a consequence, though deep learning has advanced neuroimaging-based schizophrenia classification, the preprocessing and acquisition of large datasets coupled with the achievement of high model performance and generalization remains a great challenge.

In this study, we not only address the limitations in schizophrenia classification with T1-weighted (T1W) MRI data but also take advantage of class activation maps (CAM) in a deep learning network to visualize informative regions with disease vulnerability. Our main contributions include the following: firstly, we develop a 3D CNN using structural MRI scans to yield a performance better than the benchmark model¹⁶ for schizophrenia classification; and secondly, we apply gradient class activation maps to localize the brain regions related to schizophrenia identification. By visualizing feature activations, we provide further evidence that the structure of subcortical regions and ventricular areas^{1,2} are affected in schizophrenia.

Methods

Study design. For our experiments, firstly, we implemented the schizophrenia classification task with the benchmark model using 887 structural whole-head (WH) T1W scans, following the same pre-processing and parameter settings as the implementation in the CNN benchmark¹⁶. Secondly, a modified 3D VGG¹⁸ with squeeze excitation (SE)¹⁹ and batch normalization (BN)²⁰ model (SE-VGG-11BN) was used to perform the schizophrenia vs. cognitive normal binary classification task with the input of 887 T1W structural whole-brain (WB) scans. Of the 887 scans, 437 were controls and 450 were schizophrenia patients.

Data selection. The neuroimaging data used in this study from patients with schizophrenia and normal subjects were downloaded from the SchizConnect database (<http://schizconnect.org/>). Data from three studies, BrainGluSchi²¹, COBRE²², and NMorphCH²³, were collected and organized in this public database. Images not applicable for training the deep network (e.g. those with excessive motion or noise or an image error) were excluded by visual inspection. In our experiment, the scans among all 3 studies were acquired from the same clinical MRI scanner model (SIEMENS Trio) using a standard 3D MPRAGE sequence with isotropic 1 mm resolution at 3 T field strength. The data from these studies were high in quality and resolution, and the data acquisition time was relatively recent, varying from 2008 to 2013. In summary, the data in these studies were abundant and appropriate for model training. More detailed information about this data is illustrated in Fig. 1A.

Data pre-processing. Optimizing a deep learning model using data in this specific space requires the algorithm to learn discriminative patterns when the samples are in large numbers and include all of the expected variations. By pre-processing our images, we could alleviate some of the confounding factors, enabling the model to handle the entire image at once and automatically determine the most important task-related pattern in the data.

In our data preprocessing pipeline, firstly, the raw whole-head scans from three studies were registered to the MNI152 unbiased template by robust affine registration^{24,25}, which is denoted by step one. Following the whole-head scans' registration, skull-stripping was applied on the whole-head scans using the Brain Extraction Tool²⁶ to obtain whole-brain (WB) MRI T1W scans, denoted by step two. After that, we affine-registered these whole-brain MRI T1W scans to the MNI152 unbiased template, denoted by step three. The details of these steps are illustrated in Fig. 1B.

Through affine registration, the MRI T1W scans kept similar structures in roughly the same spatial location using one template as the gold standard. We thereby reduced the variance in brain features, such as the brain volume, while still preserving differences in local anatomy, which may presumably reflect schizophrenia-related effects on brain structures. This operation could thus enable the model to focus on the decision-making patterns underlying the data.

A

Scan Parameters	BrainGluSchi	COBRE	NMorphCH
Scanner	SIEMENS TrioTim	SIEMENS TrioTim	SIEMENS TrioTim
Field strength	3T	3T	3T
Sequence	MPRAGE	MPRAGE	MPRAGE
Voxel size (mm)	1.0x1.0x1.0	1.0x1.0x1.0	1.0x1.0x1.0
TR/TE (msec)	2530/1.64	2530/1.64	2400/3.16
Acquisition year	2010 - 2013	2009 - 2013	2008 - 2013
Number of control scans (Train/Validation/Test)	89 (73/8/8)	237 (199/22/16)	111 (89/10/12)
Number of patient scans (Train/Validation/Test)	86 (70/7/9)	243 (207/23/13)	121 (98/6/17)
Female %	20.3%	24.0%	40.0%
Age range (year)	16-66	18-66	19-46
Age mean \pm SD	37.6 \pm 13.3	38.3 \pm 12.6	32.2 \pm 7.6

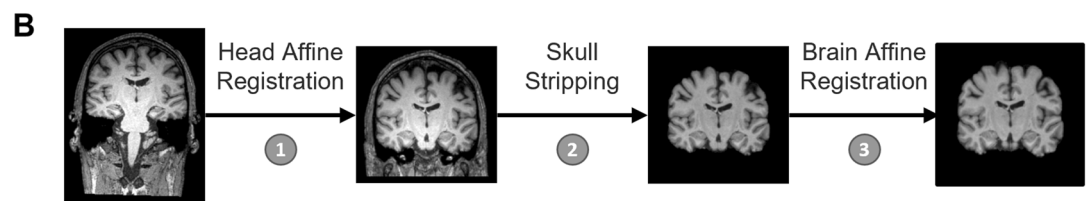


Figure 1. Sample characteristics, distribution of public schizophrenia MRI datasets and the preprocessing pipeline. **(A)** Acquisition parameters of the T1W MRI scans and the patient demographic information of each dataset. In the BrainGluSchi, COBRE, and NMorphCH datasets, normal scans consisted of whole head structural T1W MR images obtained from healthy control subjects and schizophrenia scans consisted of whole head structural T1W MR images obtained from schizophrenia and schizoaffective disorder patients. **(B)** Data preprocessing pipeline to generate the input. For each structural MRI, we process the T1W 3D volume through a standardized pipeline consisting of three steps: (1) whole head T1W affine registration to the MNI152 template space, (2) skull stripping, and (3) whole brain affine registration to the MNI152 template space. The preprocessing of structural T1W MR data is necessary to remove unwanted artifacts and transform the data into a standard format before training the deep learning models.

After visual inspection of the preprocessed scans and removal of low-quality scans to avoid their potential negative effects on the classification task, the prepared data with 887 WB MRI T1W scans were selected and randomly assigned to 10 subsets. Each subset contained a similar number of samples. Randomization was performed on the subject level to prevent data leakage. To train and evaluate the model, eight out of ten subsets were randomly selected to make up the training set. Of the other two subsets, one was used as the validation set while the other was used as the test set. Consequently, the dataset was partitioned into the train/validation/test dataset by a ratio of approximately 8:1:1 at the subject level. The gender and age distribution in each subset were similar. The details of the gender and age distribution are shown in Supplementary Fig. 1.

Down-sampling ($\times 2$) was applied to the input data with a matrix size of $192 \times 192 \times 192$ to help preserve the image information while extending the possible training batch size. This operation aimed to achieve a balance between resolution and batch size. For the model, the input is the processed 3D whole-brain T1W MRI scan while the output is a continuous-valued number representing the predicted schizophrenia likelihood.

Model architecture and implementation. For the schizophrenia classification tasks with one single input modality, the architecture 3D “VGG-11 with batch normalization” (Fig. 2) adapted from “VGG-19BN”²⁷ was developed in the PyTorch platform. VGG models are standard deep CNN architectures with 5 convolutional blocks proposed by the Visual Geometry Group, Oxford University¹⁸. They represent a group of robust architectures for object detection and image classification. This modified 3D VGG model with batch normalization and squeeze-and-excitation block (SE-VGG-11BN) was composed of two basic components: a feature extractor and a classifier. In the feature extraction portion, there was one down-sampling operation followed by five 3D

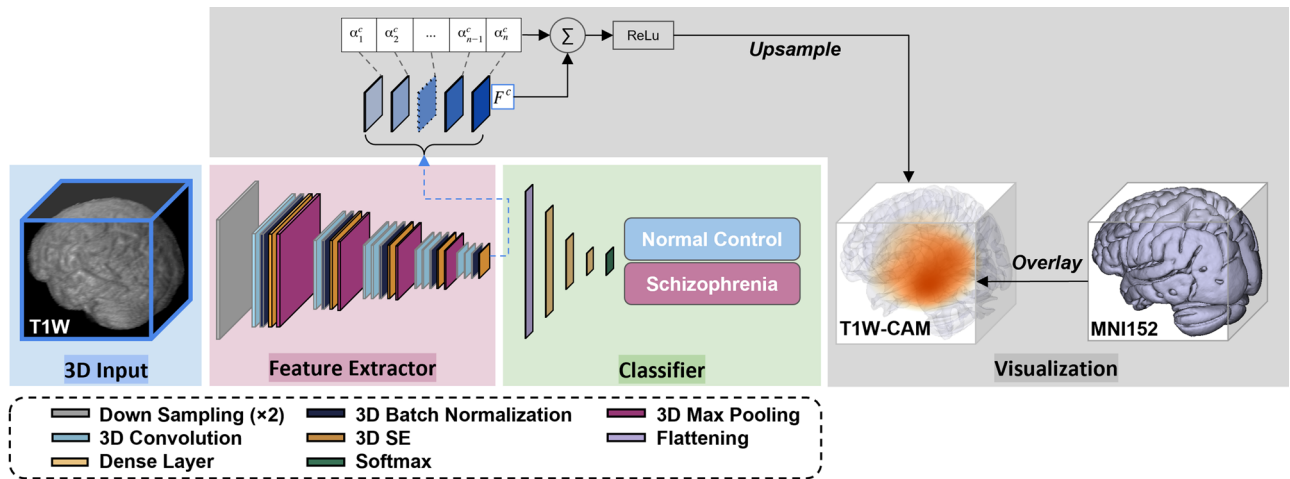


Figure 2. The flow of the classification and visualization in the 3D SE-VGG-11BN CNN model. This model consists of a modified 3D VGG-11 network with squeeze-and-excitation (SE) block and batch-normalization (BN) using T1W MRI as the model input. The class of one given T1W scan is predicted by two steps in the model: (1) extracting hierarchical features, and (2) classifying these features. In the feature extractor portion, the data is firstly under-sampled $\times 2$ and goes through several convolution blocks consisting of 3D convolution, 3D batch normalization, 3D max pooling, and 3D SE operation. The classifier consisting of three dense layers with dropout regularization yields the final prediction result. The classifier consisting of three dense layers with dropout regularization yields the final prediction result. In the feature extractor part, feature maps generated by filters at the last convolution layer are shown. These feature maps are used for visualization through the generation of the class activation map by weighting them with channel-wise average gradients.

convolution blocks, with each block containing 3D convolution, 3D batch normalization (BN), 3D squeeze-and-excitation (SE) operation, and 3D max-pooling.

Details of the operations involved in the convolutional block are illustrated as follows. The kernel size is $3 \times 3 \times 3$ and the padding and stride number is 1 in the 3D convolution. 3D batch normalization (BN) follows the convolution operation and normalizes inputs to layers in a neural network for each mini-batch. By rescaling and recentering, BN reduces the internal covariate shifts, enabling a higher learning rate. One notable difference between our model and the common VGG model lies in the introduction of the squeeze-and-excitation operation, which is a channel-wised attention mechanism used to improve the representational power of a CNN network. It adds weights factors to channels and accordingly recalibrates them to enhance significant features while ignoring the irrelevant features at almost no additional computational cost in the existing architecture. The channel-to-channel ratio is the only hyperparameter, which was tuned in the range from 8 to 32 (including 8, 12, 16, ... and up to 32), and set at 16 in the 3D SE operation. In the max-pooling, the kernel size and stride are 2×2 and 2, respectively. One slight difference from previous 3D convolution blocks is that the max-pooling in the last convolution block is abandoned since we needed a larger receptive field to generate the class activation map. In the classifier portion, three dense and two dropout layers are used to constitute the linear mapping. We specifically choose the VGG 11 model with squeeze excitation blocks because this variant performed the best for schizophrenia classification among various other CNN architectures and configurations. All the activation functions in feature extraction and classifier are rectified linear units (ReLU)²⁸ except for the penultimate and final activations, which are sigmoid and softmax functions respectively. The details of the proposed model are illustrated in Fig. 3.

In the training phase, the initial learning rate was set to $1e-4$ (tuned in the range from $1e-3$ to $1e-6$) and the batch size was 5. The setting of batch size was chosen considering convergence speed and the memory limit. The loss function was the cross-entropy loss, and the Adam²⁹ method was used to optimize the model parameters. An early stopping strategy was introduced to the training phase to avoid over-fitting. The number of epochs was set to 300.

Data augmentation could help improve the model performance by making the model more agnostic to subject-level variation. As a result, a data augmentation strategy was used in our model training as well. Basic transformers from the TorchIO library were imported to transform the raw data and thus increase the number of training datasets. The data in the training set would go through random blurring with a probability of 0.1 and random noise addition with a probability of 0.6. After these two transformations, the data would go through one of the following with a probability of 0.2: random affine transformation with scaling factor = 1 or random elastic deformation. Finally, the image would undergo random bias field distortion with a probability of 0.1 followed by random motion distortion with a probability of 0.05. Examples of each of these transformations applied on T1W WB MRI scans are shown in Fig. 4.

Performance evaluation of the model. To evaluate the descriptiveness of the predicted schizophrenia likelihood, we conducted receiver-operating characteristics (ROC) studies to analyze the concordance between

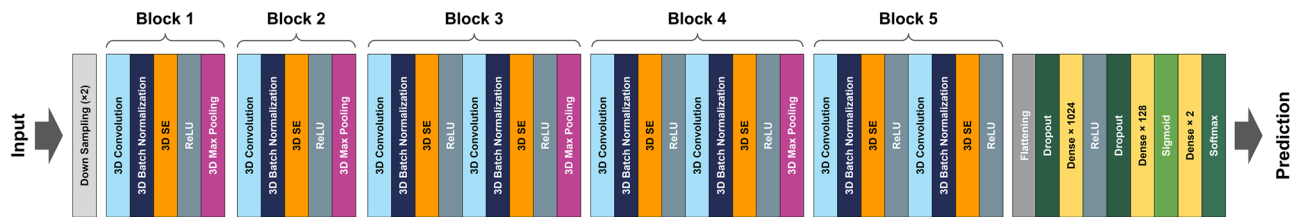


Figure 3. Overall network architecture of the proposed approach. The inputs are 3D brain volumes after being down-sampled by a factor of 2. This modified 3D VGG model is composed of two basic components: a feature extractor and a classifier. The feature extraction portion consists of five independent 3D convolution blocks. In each of the first two blocks, one stack of 3D convolution, 3D batch normalization (BN), 3D squeeze-and-excitation (SE) layers, and ReLU activation is followed by a 3D max-pooling operation. Each of the last 3 blocks contains two repetitions of 3D convolution, 3D batch normalization, 3D squeeze-and-excitation operation, and a ReLU activation. Block 3 and 4 end with a 3D max-pooling operation but the 3D max-pooling operation of block 5 is abandoned to preserve enough size of the feature maps to generate the class activation map. The classifier is comprised of three dense layers, with the first two layers preceded by a dropout layer. The first dense layer is followed by the ReLU activation function whereas the second dense layer is followed by the sigmoid activation function, and the final dense layer is followed by the softmax function.

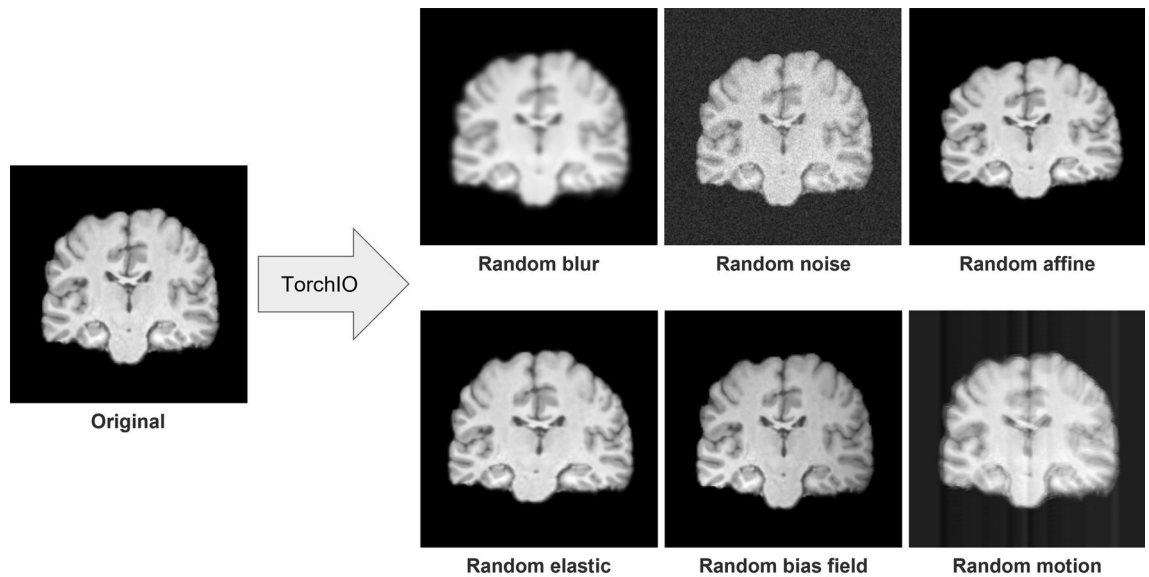


Figure 4. An example of 3D T1-weighted MRI data augmentation results in the coronal view. From left to right: the original single coronal slice; the single slice after randomly blurring; after adding Gaussian noise; after applying random affine transformation and resampling; after applying random dense and elastic deformation; after applying random bias field distortion; and after adding random motion artifact, respectively.

the model-generated classification and the ground truth labels. The ROC curves, one for each trained classification model, represent the classification performances at each potential numerical threshold to binarize the predicted schizophrenia likelihood score. The sensitivity and specificity (the sum of which peaks at the operating point), as well as the area under the ROC curve (AUC), demonstrate the effectiveness of the classification method. The significance of the difference among these ROC curves was calculated using DeLong's test³⁰.

Generalization evaluation of the model. To demonstrate the generalization of the models, data from COBRE and NMorphCH studies were selected to train the model, and data from BrainGluSchi with a nearly similar acquisition configuration was used for evaluating the capability of model generalization. The same training strategies and hyperparameter settings were maintained in the experiment.

Explainability of the model with grad-CAM. To validate the models, a gradient class activation map (Grad-CAM) was introduced to our experiment to check whether the model focuses on task-related patterns instead of some irrelevant information in the data. After excluding the possibility of the model focusing on meaningless regions in the data by applying a rough brain mask, we further investigated the brain regions that had the most contributions to the schizophrenia classification task by visualizing the class activation maps (CAM)³¹. By visualizing feature activations, we could identify which regions of the input images contribute to

the classification results. We used the WB T1W scans from all the subjects with schizophrenia to generate an averaged CAM for the schizophrenia class. We had a great interest in whether the brain regions the classifier found the most relevant to the schizophrenia class were physiologically meaningful. The weighted feature maps only activate features that have a positive influence on the prediction after applying ReLU nonlinearity operations. The heatmap of the last convolution layer highlights the most important region for classifying the sample, whereas the maps for shallow layers localize more fine-grained features.

Results

While training, we found that SE-VGG-11BN converged faster than the benchmark model on the training set and performed better than the benchmark model on the validation set. After training both our model and the benchmark model, we tested them on the same stand-alone set of scans, 51 with schizophrenia and 49 without schizophrenia. The SE-VGG-11BN model using structural T1 WB scans exhibited better performance than the benchmark model across all metrics (0.921 accuracy, 0.949 sensitivity, and 0.946 specificity). The quantitative performance metrics are summarized in Fig. 5B and C. When inspecting the ROC curves (Fig. 5A), we found that the SE-VGG-11BN model with the input of structural T1 WB scans achieved 0.987 AUC, which outperforms the benchmark model that achieved 0.938 AUC. The p -value of the ROC test (DeLong's test) indicated our model is significantly better than the benchmark model using structural T1 WH scans at a level of 0.05. We adjusted the model architecture, and the classification performance is reported below: VGG-11BN model achieved 0.96 accuracy, 0.974 sensitivity, 0.944 specificity, and AUC score of 0.964; the VGG-13BN model achieved 0.92 accuracy, 0.974 sensitivity, 0.889 specificity and AUC score of 0.963; the SE-VGG-13BN model achieved 0.933 accuracy, 0.949 sensitivity, 0.944 specificity and AUC score of 0.982. Apart from comparing our method with the benchmark model, we also compare with the 3D Dense-Net³² and 3D Res-Net³³ models which are used for schizophrenia-like disorder classification like depression. The Dense-Net achieved 0.853 accuracy, 100.0% sensitivity, 0.778 specificity, and an AUC score of 0.947. The Res-Net achieved 0.813 accuracy, 0.923 sensitivity, 0.75 specificity, and an AUC score of 0.857%. Campese et al.³⁴ compared 3D CNN, 2D CNN and SVM on structural MRI data, with 3D CNN outperforming the other two models.

SE-VGG-11BN also demonstrated improved generalization performance. When the BrainGluSchi dataset was used only for testing and the COBRE and NMorphCH datasets were used for training and validation, we observed significantly superior AUC performance of SE-VGG-11BN (0.913) over the benchmark model (0.810) (Fig. 5D). These results validate the generality of our model and highlight its reliability in predicting schizophrenia on unseen and heterogeneous structural MRI data.

To investigate the most pertinent spatial features contributing to the classification ability of the proposed deep learning algorithm, we further analyzed regional information in the anatomical structures of the structural MRI data. We illustrate the class activation map of the SE-VGG-11BN for schizophrenia patients in Fig. 6, localizing discriminative regions for schizophrenia classification in the sagittal, axial and coronal views. The class activation map indicates large "activation" in the subcortical and ventricular regions, suggesting the importance of these regions in differentiating schizophrenia for our proposed model.

Discussion

This study investigated the performance of 3D VGG-based models on the classification of schizophrenia patients using structural MRI scans. The proposed model (SE-VGG11-BN) showcased superior performance and generality compared to the benchmark model in terms of sensitivity, specificity, accuracy, and AUC in both of our experiments. Furthermore, the proposed model, SE-VGG11-BN, was interpreted with gradient class activation maps to visualize the brain regions critical for classification. The important regions for classification involved subcortical and ventricular areas; these were in line with the findings in the previous literature.

The superior performance of the proposed model against the benchmark model. SE-VGG-11BN exhibited better performance than the benchmark model. Several factors may have contributed to this result. Firstly, in contrast to the benchmark model, the proposed model contains squeeze-and-excitation (SE) blocks, which can capture important patterns across all channels after each convolutional operation. Secondly, the input of the proposed model was only down-sampled by a factor of two as opposed to the benchmark model, which used a larger factor of eight. Severely down-sampling the data likely negatively impacted model performance as lower-resolution inputs may have lost important information relevant to schizophrenia classification. Thirdly, we applied skull-stripping as part of our data preprocessing pipeline, given that the skull holds limited clinical correspondence to schizophrenia. The benchmark model used T1W WH scans, which may have unnecessarily confused the model with irrelevant features from the skull.

Interpretation of the proposed model's grad-CAM. The class activation map of the proposed model (SE-VGG-11BN) reveals that the subcortical regions and ventricular areas provide the most discriminative structural information for schizophrenia classification. This result is consistent with two recent meta-analyses considering changes in regional brain structure volume² and shape¹ associated with schizophrenia. The first study examining the heterogeneity and homogeneity of regional brain structure in schizophrenia found that mean volumes were significantly reduced for the temporal lobe, frontal lobe, anterior cingulate cortex, and the subcortical regions including the thalamus, hippocampus, and amygdala; whereas, mean volumes of the lateral and third ventricles were significantly increased in patients². The second meta-analysis investigating changes in subcortical brain shape associated with schizophrenia studied T1-weighted structural MRI scans from 2833 individuals with schizophrenia and 3929 healthy control participants contributed by 21 worldwide research groups participating in the ENIGMA Schizophrenia Working Group¹. This study revealed more-concave-than-convex shape

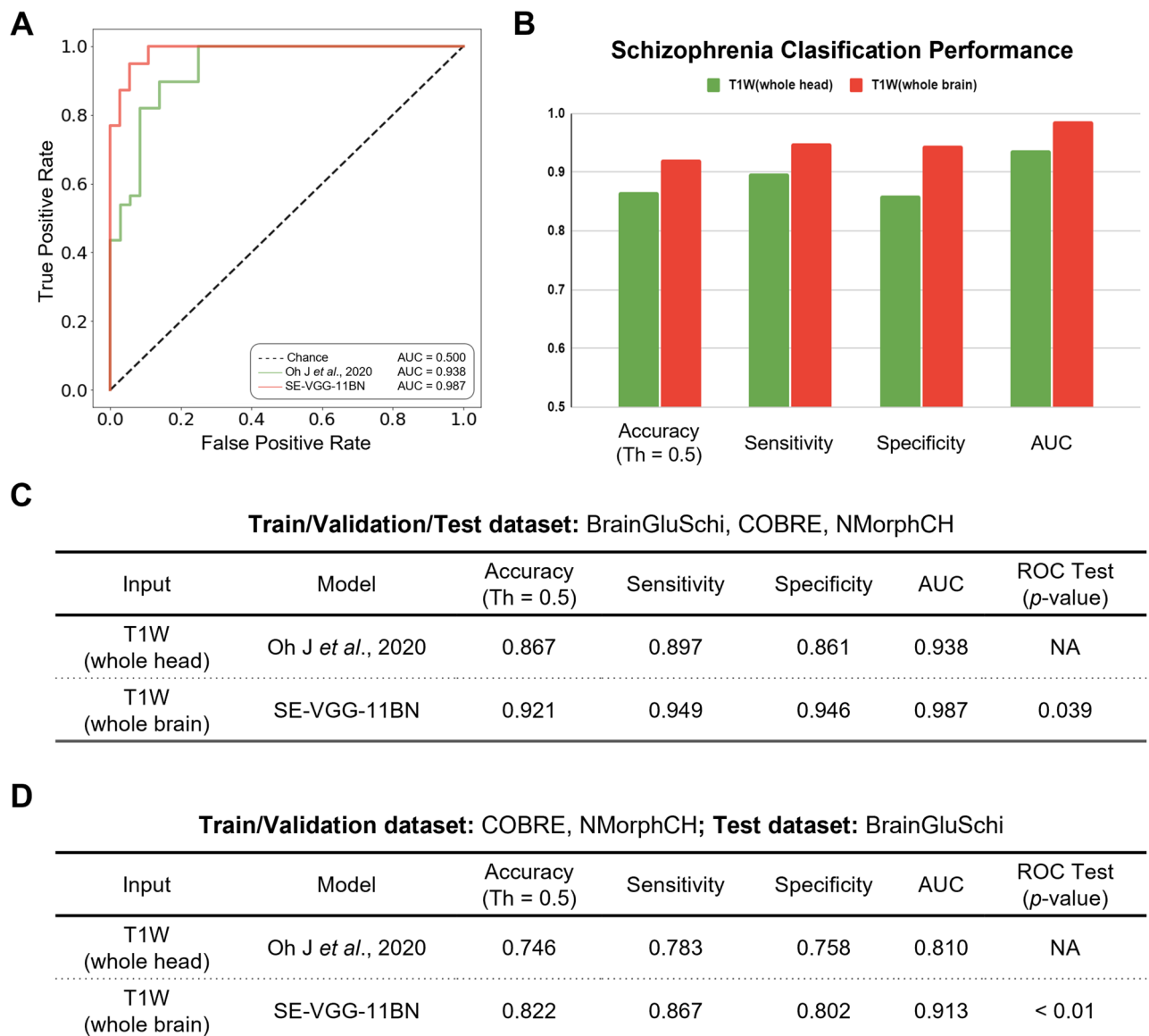


Figure 5. Quantitative performance comparisons of our model and the benchmark model. **(A)** Receiver operating characteristics (ROC) curves for schizophrenia classification on the dataset. The green line represents the ROC curve of the benchmark model with the input of T1W WH scans. The red line represents the ROC curve of the SE-VGG-11BN with the input of T1W WB scans. **(B)** Bar plot of the classification performance of these models in terms of the accuracy (at the default operating threshold of 0.5, Th = 0.5), sensitivity, specificity, and the area under the ROC curve (AUC). **(C)** Table quantitatively summarizes the performance of these models. The p -value of the ROC test (DeLong's test) indicated our model is significantly better than the benchmark model at a level of 0.05. **(D)** Generalizability of the two models trained by COBRE and NMorphCH datasets on unseen BrainGluSchi test dataset. Considering components A through D, SE-VGG-11BN exhibits a significantly better performance than the benchmark model.

differences in the hippocampus, amygdala, accumbens, and thalamus in individuals with schizophrenia compared with control participants, more-convex-than-concave shape differences in the putamen and pallidum, and both concave and convex shape differences in the caudate. Patterns of exaggerated asymmetry were observed across the hippocampus, amygdala, and thalamus in individuals with schizophrenia compared to control participants, while diminished asymmetry encompassed the ventral striatum and ventral and dorsal thalamus. Notably, the hippocampus, a region found to be remarkably related to schizophrenia progression^{35–39}, is also included in the activation regions. Findings from our deep learning-based study suggest that common mechanisms may contribute to volume and shape variability across multiple subcortical regions and ventricular areas, which may enhance our understanding of the nature of network disorganization in schizophrenia.

Limitations and future work. There are certain limitations associated with the application of 3D CNN to schizophrenia classification. Firstly, the high computational cost during training caused by data with high

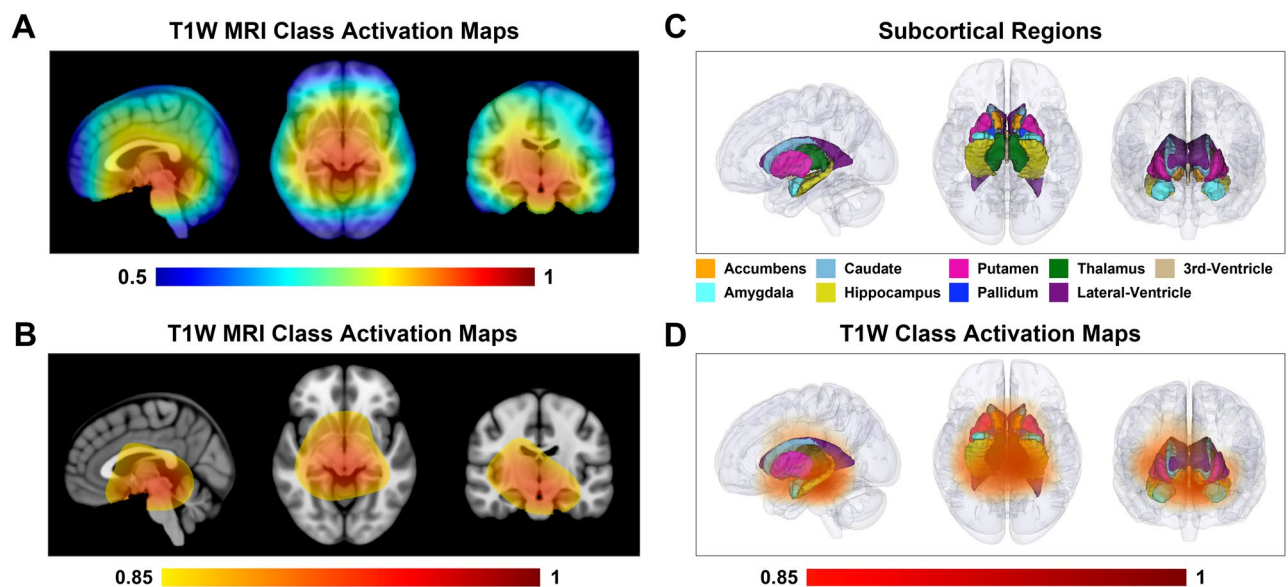


Figure 6. Class activation map on T1-weighted brain MRI images in schizophrenia classification. **(A)** The class activation map (CAM) derived from feature maps in the last convolution layer from schizophrenia patients localizes the discriminative regions for schizophrenia classification in the sagittal, axial and coronal views. The color bar ranges from 0.5 to 1. The higher the value, the more important role the region plays in schizophrenia classification. **(B)** The CAM is displayed in the sagittal, axial, and coronal views with a threshold of 0.85. The color bar ranges from 0.85 to 1. The thresholded CAM primarily lies in the subcortical regions and ventricular areas. **(C)** Subcortical regions and ventricular areas are visualized in 3D in the sagittal, axial, and coronal views. **(D)** The 3D volume rendering of the thresholded CAM demonstrates the location of the most discriminative regions in the sagittal, axial, and coronal views. The color bar ranges from 0.85 to 1. The thresholded CAM covers the subcortical regions and the ventricular areas.

dimensionality and large numbers of trainable parameters in the model may constrain the development of 3D CNN models. In this study, we down-sampled the raw data to reduce the GPU memory workload and preserved as many details as possible, simultaneously using a small down-sampling factor of two. The complexity of the models is also largely limited by the GPU memory requirement. Secondly, the sample size used in this study is relatively modest, especially for the 3D CNN network training. This most likely results in less efficient feature extraction and lower model generalizability. Introducing other high-quality labeled datasets coupled with data augmentation and effective image synthesis of new data could help model feature extraction and generalization via the introduction of more inter-subject anatomical variability and data quality deviation across different sites. Thirdly, registration error and down-sampling may ignore certain subtle anatomical differences and low-level contextual features potentially relevant to schizophrenia classification. Lastly, the training strategy of the model could be further improved. For instance, the proposed CNN model is trained from scratch, but applying and fine-tuning a pre-trained model on our data could further improve model performance. There is evidence suggesting that this approach may enhance performance by reducing the cost of a more computationally complex training stage⁴⁰.

A desirable future application of deep learning includes addressing the clinically more pressing question of discriminating schizophrenia from other psychiatric disorders, such as major depression and bipolar disorder. The partial overlap of genetic and similar symptoms characteristic between schizophrenia and other major psychiatric disorders makes this task very challenging, even for clinicians. In fact, meta-analyses of transcriptomic studies covering five major psychiatric disorders found an overlap in polygenic traits and global gene expression patterns⁴¹. Moreover, the symptoms of schizophrenia also overlap with other psychiatric disorders such as major depressive disorder, schizoaffective disorder, and post-traumatic disorder^{42,43}. Though our proposed approach could differentiate schizophrenia patients from healthy controls using T1W structural MRI data, there is currently no objective method able to classify schizophrenia from other similar neuropsychiatric disorders. Further research in this area may elucidate the mechanisms and features underlying the brain structural alterations in different psychiatric disorders. Lastly, while this model could be adapted for real-world clinical usage, it was trained on images of well-established patients who had many years of prior illness. Given that the effects of chronic illness could have been present within the T1 data, this model would have to be re-fit with T1 data from first-episode psychotic patients for more accurate clinical usage.

Data availability

The T1W MRI scans used in this project are available from the SchizConnect database, <http://www.schizconnect.org>. The code used in this project is proprietary. The preprocessing script and the deep learning model are available upon request to the corresponding author. The code for this project is © 2022 The Trustees of Columbia

University in the City of New York. This work may be reproduced and distributed for academic non-commercial purposes only.

Received: 2 October 2022; Accepted: 25 August 2023

Published online: 02 September 2023

References

- Gutman, B. A. *et al.* A meta-analysis of deep brain structural shape and asymmetry abnormalities in 2833 individuals with schizophrenia compared with 3929 healthy volunteers via the ENIGMA consortium. *Hum. Brain Mapp.* **43**(1), 352–372 (2022).
- Brugger, S. P. & Howes, O. D. Heterogeneity and homogeneity of regional brain structure in schizophrenia: A meta-analysis. *JAMA Psychiatry* **74**(11), 1104–1111 (2017).
- Shenton, M. E., Dickey, C. C., Frumin, M. & McCarley, R. W. A review of MRI findings in schizophrenia. *Schizophr. Res.* **49**(1–2), 1–52 (2001).
- Gaser, C., Nenadic, I., Buchsbaum, B. R., Hazlett, E. A. & Buchsbaum, M. S. Ventricular enlargement in schizophrenia related to volume reduction of the thalamus, striatum, and superior temporal cortex. *Am. J. Psychiatry* **161**(1), 154–156 (2004).
- Koutsouleris, N. *et al.* Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. *Schizophr. Bull.* **41**(2), 471–482 (2015).
- Patel, K. R., Cherian, J., Gohil, K. & Atkinson, D. Schizophrenia: Overview and treatment options. *Pharm. Ther.* **39**(9), 638 (2014).
- Sun, D. *et al.* Progressive brain structural changes mapped as psychosis develops in ‘at risk’ individuals. *Schizophr. Res.* **108**(1–3), 85–92 (2009).
- Arbabshirani, M. R., Plis, S., Sui, J. & Calhoun, V. D. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage* **145**, 137–165 (2017).
- Davatzikos, C. Machine learning in neuroimaging: Progress and challenges. *Neuroimage* **197**, 652 (2019).
- de Filippis, R. *et al.* Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: A systematic review. *Neuropsychiatr. Dis. Treat.* **15**, 1605 (2019).
- Rashid, B. & Calhoun, V. Towards a brain-based predictive model of mental illness. *Hum. Brain Mapp.* **41**(12), 3468–3535 (2020).
- Winterburn, J. L. *et al.* Can we accurately classify schizophrenia patients from healthy controls using magnetic resonance imaging and machine learning? A multi-method and multi-dataset study. *Schizophr. Res.* **214**, 3–10 (2019).
- Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
- Shen, D., Wu, G. & Suk, H. I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **19**, 221 (2017).
- Li, R. *et al.* Deep learning based imaging data completion for improved brain disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (eds Golland, P. *et al.*) 305–312 (Springer, 2014).
- Oh, J., Oh, B. L., Lee, K. U., Chae, J. H. & Yun, K. Identifying schizophrenia using structural MRI with a deep learning algorithm. *Front. Psychiatry* **11**, 16 (2020).
- Hu, M. *et al.* Structural and diffusion MRI based schizophrenia classification using 2D pretrained and 3D naive convolutional neural networks. *Schizophr. Res.* <https://doi.org/10.1016/j.schres.2021.06.011> (2021).
- Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Machine Learning (ICML, 2015)*.
- Hu, J., Shen, L., Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141 (2018).
- Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 448–456 (PMLR, 2015).
- Bustillo, J. R. *et al.* Glutamatergic and neuronal dysfunction in gray and white matter: A spectroscopic imaging study in a large schizophrenia sample. *Schizophr. Bull.* **43**(3), 611–619 (2017).
- Chyzyk, D., Savio, A. & Graña, M. Computer aided diagnosis of schizophrenia on resting state fMRI data by ensembles of ELM. *Neural Netw.* **68**, 23–33 (2015).
- Alpert, K., Kogan, A., Parrish, T., Marcus, D. & Wang, L. The northwestern university neuroimaging data archive (NUNDA). *Neuroimage* **124**, 1131–1136 (2016).
- Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* **5**(2), 143–156 (2001).
- Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**(2), 825–841 (2002).
- Smith, S. M. Fast robust automated brain extraction. *Hum. Brain Mapp.* **17**(3), 143–155 (2002).
- Simon, M., Rodner, E., Denzler, J. Imagenet pre-trained models with batch normalization. Preprint at <https://arxiv.org/abs/1612.01452> (2016).
- Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *ICML (2010)*.
- Kingma, D. P., Ba, J. L. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations*.
- DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **1**, 837–845 (1988).
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2921–2929 (2016).
- Wang, Yu., Gong, N. & Changyang, Fu. Major depression disorder diagnosis and analysis based on structural magnetic resonance imaging and deep learning. *J. Integr. Neurosci.* **20**(4), 977–984 (2021).
- Hong, J. *et al.* 3D FRN-ResNet: An automated major depressive disorder structural magnetic resonance imaging data identification framework. *Front. Aging Neurosci.* **14**, 912283 (2022).
- Campese, S., Lauriola, I., Scarpazza, C., Sartori, G. & Aiolfi, F. Recent advances in big data and deep learning. In *Psychiatric Disorders Classification with 3D Convolutional Neural Networks* (eds Oneto, L. *et al.*) 48–57 (Springer, 2020).
- Gaisler-Salomon, I., Schobel, S. A., Small, S. A. & Rayport, S. How high-resolution basal-state functional imaging can guide the development of new pharmacotherapies for schizophrenia. *Schizophr. Bull.* **35**(6), 1037–1044 (2009).
- Heckers, S. & Konradi, C. Hippocampal pathology in schizophrenia. In *Behavioral Neurobiology of Schizophrenia and its Treatment* (ed. Swerdlow, N. R.) 529–53 (Springer Berlin, Heidelberg, 2010).
- Lieberman, J. A. *et al.* Hippocampal dysfunction in the pathophysiology of schizophrenia: A selective review and hypothesis for early detection and intervention. *Mol. Psychiatry* **23**(8), 1764–1772 (2018).
- Schobel, S. A. *et al.* Imaging patients with psychosis and a mouse model establishes a spreading pattern of hippocampal dysfunction and implicates glutamate as a driver. *Neuron* **78**(1), 81–93 (2013).
- Provenzano, F. A. *et al.* Hippocampal pathology in clinical high-risk patients and the onset of schizophrenia. *Biol. Psychiatry* **87**(3), 234–242 (2020).
- Jarrett, K., Kavukcuoglu, K., Ranzato, M. A. & LeCun, Y. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision* 2146–2153 (IEEE, 2009).

41. Gandal, M. J. *et al.* Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**(6376), 693–697 (2018).
42. Renard, S. B. *et al.* Unique and overlapping symptoms in schizophrenia spectrum and dissociative disorders in relation to models of psychopathology: A systematic review. *Schizophr. Bull.* **43**(1), 108–121 (2017).
43. Stein, F. *et al.* Factor analyses of multidimensional symptoms in a large group of patients with major depressive disorder, bipolar disorder, schizoaffective disorder and schizophrenia. *Schizophr. Res.* **218**, 38–47 (2020).

Acknowledgements

The results published here are based on public schizophrenia neuroimaging data obtained from the SchizConnect virtual database. The data mediator was developed at the Information Sciences Institute of the University of Southern California (ISI). The BrainGluSchi is a case-control study at the University of New Mexico Hospitals to examine metabolites and neuronal dysfunction in schizophrenia patients and healthy control subjects over a broad age range. The Center for Biomedical Research Excellence (COBRE) in Brain Function and Mental Illness provides structural and functional MRI data for exploration of schizophrenia mechanisms by integrating multiple neuroimaging methods with psychiatric, neuropsychological and genetic testing. The NMorphCH is a longitudinal study examining the clinical, cognitive and MRI data from schizophrenia and control subjects at baseline and after two years. The NMorphCH study was conducted at Northwestern University. The authors appreciate Stephen Rayport and Douglas Rothman for their constructive comments and suggestions.

Author contributions

J.G. conceived, designed, and supervised all studies constituting this article and verified all statistical results. J.Z. collected, organized, and preprocessed the data, designed and optimized the VGG model, and performed statistical analyses. J.Z., V.R., Y.T., Y.Y., C.Z. and J.G. wrote the manuscript. J.Z., Y.T. and Y.Y. created and updated all display items (figures and tables). N.A., Z.W. and P.Y.L. participated in the data processing. L.K. and S.S. kindly provided insights and suggestions for the research. All authors reviewed, commented, and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-41359-z>.

Correspondence and requests for materials should be addressed to J.G.

Reprints and permissions information is available at www.nature.com/reprints.

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



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 Author(s) 2023