frontiers in Neuroscience

ORIGINAL RESEARCH
published: 07 July 2021
doi: 10.3389/fnins.2021.652987

Three-Dimensional Convolutional Autoencoder Extracts Features of Structural Brain Images With a "Diagnostic Label-Free" Approach: Application to Schizophrenia Datasets

Hiroyuki Yamaguchi[1,2]*, Yuki Hashimoto*[1]*, Genichi Sugihara*[3]*, Jun Miyata*[4]*, Toshiya Murai*[4]*, Hidehiko Takahashi*[3]*, Manabu Honda*[1]*, Akitoyo Hishimoto*[2] *and Yuichi Yamashita*[1]*

[1] Department of Information Medicine, National Center of Neurology and Psychiatry, National Institute of Neuroscience, Tokyo, Japan, *[2] Department of Psychiatry, School of Medicine, Yokohama City University, Yokohama, Japan,* *[3] Department of Psychiatry and Behavioral Sciences, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan,* *[4] Department of Psychiatry, Graduate School of Medicine, Kyoto University, Kyoto, Japan*

OPEN ACCESS

Edited by:
Albert Yang,
National Yang-Ming University, Taiwan

Reviewed by:
Adil Bashir,
Auburn University, United States
Shih-Jen Tsai,
Taipei Veterans General Hospital, Taiwan

***Correspondence:**
Yuichi Yamashita
yamay@ncnp.go.jp

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *13 January 2021*
Accepted: *07 June 2021*
Published: *07 July 2021*

Citation:
Yamaguchi H, Hashimoto Y, Sugihara G, Miyata J, Murai T, Takahashi H, Honda M, Hishimoto A and Yamashita Y (2021) Three-Dimensional Convolutional Autoencoder Extracts Features of Structural Brain Images With a "Diagnostic Label-Free" Approach: Application to Schizophrenia Datasets. Front. Neurosci. 15:652987. doi: 10.3389/fnins.2021.652987

There has been increasing interest in performing psychiatric brain imaging studies using deep learning. However, most studies in this field disregard three-dimensional (3D) spatial information and targeted disease discrimination, without considering the genetic and clinical heterogeneity of psychiatric disorders. The purpose of this study was to investigate the efficacy of a 3D convolutional autoencoder (3D-CAE) for extracting features related to psychiatric disorders without diagnostic labels. The network was trained using a Kyoto University dataset including 82 patients with schizophrenia (SZ) and 90 healthy subjects (HS) and was evaluated using Center for Biomedical Research Excellence (COBRE) datasets, including 71 SZ patients and 71 HS. We created 16 3D-CAE models with different channels and convolutions to explore the effective range of hyperparameters for psychiatric brain imaging. The number of blocks containing two convolutional layers and one pooling layer was set, ranging from 1 block to 4 blocks. The number of channels in the extraction layer varied from 1, 4, 16, and 32 channels. The proposed 3D-CAEs were successfully reproduced into 3D structural magnetic resonance imaging (MRI) scans with sufficiently low errors. In addition, the features extracted using 3D-CAE retained the relation to clinical information. We explored the appropriate hyperparameter range of 3D-CAE, and it was suggested that a model with 3 blocks may be related to extracting features for predicting the dose of medication and symptom severity in schizophrenia.

Keywords: deep learning, machine learning, neuroimaging, schizophrenia, structural MRI, convolutional autoencoder, diagnostic label

INTRODUCTION

Deep learning (DL) has dramatically improved technology in speech recognition, image recognition, and many other fields (LeCun et al., 2015). Medical imaging can benefit greatly from recent progress in image classification and object detection using this cutting-edge technology (Esteva et al., 2019). In particular, as the global burden of psychiatric disorders increases (Olesen et al., 2012; Whiteford et al., 2013), psychiatric brain imaging studies using DL are anticipated to bring many benefits to society (Vieira et al., 2017). There are two major concerns about applying DL to psychiatric brain imaging: (1) treatment of the high dimensionality of data, and (2) the heterogeneity of psychiatric disorders (Feczko et al., 2019).

The dimensionality of raw magnetic resonance imaging (MRI) data is very high (often running into the millions), and large computer resources are required to analyze them. To reduce computational demands, in most neuroimaging studies, several feature extraction methods have been used. Region of interest (ROIs), one of the most popular feature extraction methods, has contributed to detecting various structural and functional abnormalities in the brains of patients with psychiatric disorders (Fornito et al., 2012; Fusar-Poli et al., 2012; Linden, 2012; Ratnanather et al., 2013). ROIs (often dozens or hundreds) are usually set based on neuroscience knowledge (Tzourio-Mazoyer et al., 2002). For example, average gray matter volumes or cortical thicknesses at specific ROIs are extracted as feature, and then the relationship between the feature and disease clinical information is analyzed (Desikan et al., 2006; Poldrack, 2007; Nelson et al., 2017). Even in the studies using DL, ROI-based features are often used as input (Vieira et al., 2017; Heinsfeld et al., 2018; Pinaya et al., 2019). In addition, many DL studies avoid using three-dimensional (3D) images directly, but instead, DL networks are trained using two-dimensional slices (Sarraf et al., 2017; Vieira et al., 2017; Aghdam et al., 2019). A limitation of these studies is that they ignore the 3D spatial information contained within the original MRI scans.

In recent years, with improvements in computer performance and refinement of computational techniques, studies have investigated how to treat 3D MRI scans as inputs to DL. For example, Wang et al. (2018) successfully discriminated Alzheimer's dementia from healthy subjects using 3D MRI data as input to DL. Similar attempts have been made for discriminating psychiatric disorders, including schizophrenia (Qureshi et al., 2019) and developmental disorders (Wang et al., 2019). Although these studies demonstrated that DL could apply to the analysis of 3D MRI data, discrimination-based approaches may be challenging due to the heterogeneity of psychiatric disorders.

Heterogeneity is one of the main challenges that current psychiatric research faces (Feczko et al., 2019). The current symptom-based definitions of psychiatric disorders, standardized in the Diagnostic and Statistical Manual of the American Psychiatric Association (DSM) (American Psychiatric Association., 2013) and the International Classification of Diseases (ICD) (World Health Organization., 1992), have been highlighted as lacking predictive and clinical validity due to genetic and clinical heterogeneity (Owen, 2014). For example, in schizophrenia, a recent study found evidence for significant overlapping of the relatively common risk variants tagged in genome-wide association studies (GWAS) between several psychiatric disorders, and there may also be lower genetic correlation within disorders (Lee et al., 2014). In addition, even in patients given the same diagnosis of schizophrenia, the severity of symptoms, response to medication, and prognosis often vary widely among patients (van Os and Kapur, 2009; Owen et al., 2016). Therefore, in psychiatric disorders research, a simple competition for discrimination accuracy based on the current disorder categories may be insufficient to elucidate on pathophysiology, although most current studies using DL are attempting to discriminate disease in healthy subjects (Plis et al., 2014; Vieira et al., 2017; Gao et al., 2021; Quaak et al., 2021).

One possible alternative direction for using DL techniques in psychiatric neuroimaging studies may be diagnostic label-free feature extraction. In the current study, we focus on an autoencoder (AE) as a DL algorithm that allows feature extraction without labels (Hinton, 2006). AE is supervised learning in a deep neural network having an output layer with the same data as the input layer. Since the input is as supervision, no labels are needed, unlike in general supervised learning.

Indeed, there are some studies that have used AE-based feature extraction for psychiatric neuroimaging. For example, Pinaya et al. (2019) extracted features from structural MRI scans using AE, i.e., without using diagnostic labels. The authors successfully predicted the age and gender of participants, and discriminated patients with autism spectrum disorders (ASD) and schizophrenia from healthy subjects. However, these studies used ROI-based features such as cortical thickness and functional connectivity as inputs to the AE. As such, the use of 3D brain images for inputs to the AE remains challenging, with a few exceptions. For example, Martinez-Murcia et al. (2020) extracted features from 3D brain MRI data of patients with Alzheimer's dementia using a 3D convolutional autoencoder (3D-CAE). They demonstrated that the extracted feature was useful for predicting age and Mini-Mental State Examination (MMSE) scores. This supports the efficacy of labeling free features based on 3D-CAE with MRI. However, particularly when investigating psychiatric disorders, the appropriate architecture of 3D-CAE has not been fully investigated.

The purpose of this study was to investigate an efficient 3D-CAE-based feature extraction for the neuroimaging of psychiatric disorders. More specifically, in the current study, we used datasets that included patients with schizophrenia, which has frequently been reported to be heterogeneous in previous neuroimaging studies (Sugihara et al., 2017). The key points of our study are: (1) to use 3D MRI data while preserving spatial information, and (2) diagnostic label-free feature extraction using 3D-CAE. For this purpose, we explored appropriate network structures of 3D-CAE by developing models with different network structures and comparing the predictive performance of clinical information by these extracted features.

MATERIALS AND METHODS

Experimental Overview

Figure 1 illustrates an experimental overview of our study. We used two datasets, including participants diagnosed with schizophrenia as well as healthy subjects: a dataset collected at Kyoto University (Kyoto dataset) and a public dataset, The Center for Biomedical Research Excellence (COBRE[1]) dataset. (1) Gray matter was first extracted from the structural MRI data as preprocessing. (2) We then trained 3D-CAE to extract a latent feature representation from structural MRI using the Kyoto dataset. Sixteen 3D-CAEs with varying network structures were prepared for investigation of the optimal network depth and complexity. (3) Subsequently, the COBRE dataset was used to evaluate the applicability to another dataset. (4) Finally, we evaluated whether the extracted feature retained clinical information by linear regression of the clinical information using the COBRE dataset.

Convolutional Autoencoder Training

An autoencoder is a kind of DL consisting of the encoder and the decoder. The encoder learns latent representations and reduces the dimension of the input. The decoder learns to reproduce the input as close as possible to the original using the latent representations. 3D-CAE extends this architecture by using convolutional layers that can extract features directly from 3D images (Guo et al., 2017; Nishio et al., 2017; Oh et al., 2019). The CAE has two main hyper parameters: the number of convolutional layers and the number of channels, which are the target of the current study.

The convolutional layers apply a filter to input to create feature maps that summarize the feature detected in the input. The feature maps are created for the number of channels. Since the convolutional layer generates feature maps while capturing the spatial information of the matrix, convolutional neural networks are beneficial to learning features of images. As the number of channels increases, the complexity of a model increases, but the number of dimensions of latent feature increase and requires a huge amount of computational power. Also, as the number of convolutions increases, the effective receptive field increases, thus allowing global and abstract feature to be extracted. The effective receptive field is a region of the original image that can potentially influence the activation of neurons (Le and Borji, 2017; Luo et al., 2017). If the effective receptive field is small, the feature will contain only local information of the brain, and if it is large, it will contain information on the whole brain.

[1] http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html

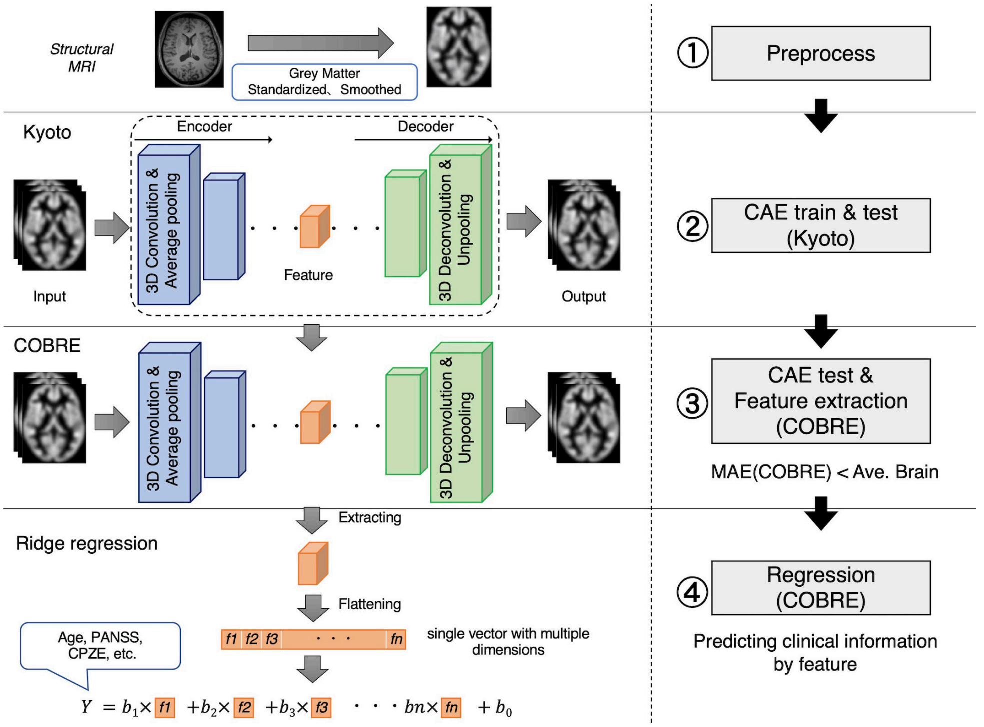

Structural MRI
Grey Matter Standardized、Smoothed
Kyoto
Encoder
Decoder
Input
3D Convolution & Average pooling
Feature
3D Deconvolution & Unpooling
Output
COBRE
3D Convolution & Average pooling
3D Deconvolution & Unpooling
Ridge regression
Extracting
Flattening
Age, PANSS, CPZE, etc.
f1 f2 f3 · · · fn
single vector with multiple dimensions
$Y = b_1 \times f1 + b_2 \times f2 + b_3 \times f3 \cdot\cdot\cdot bn \times fn + b_0$
① Preprocess
② CAE train & test (Kyoto)
③ CAE test & Feature extraction (COBRE)
MAE(COBRE) < Ave. Brain
④ Regression (COBRE)
Predicting clinical information by feature

FIGURE 1 | Experimental overview. 1. Preprocessing: The gray matter was extracted from the structural MRI and standardized and smoothed using SPM. 2. CAE training: A schematic diagram is shown. 3D images of the Kyoto dataset were input, feature was extracted, and the original image was reproduced. 3. Feature extraction: the model trained using the Kyoto dataset was adopted to the COBRE dataset without updating the weights. 4. Linear regression: Feature are extracted and flattened. Each extracted feature vectors were an explanatory variable, and demographic and clinical information were objective variables. Regression errors were evaluated to investigate whether the extracted features retain the information to predict demographic and clinical information. 3D, three-dimensional; CAE, convolutional autoencoder; COBRE, Center for Biomedical Research Excellence; CPZE, dose of antipsychotic medication; MRI, magnetic resonance imaging; SPM, Statistical Parametric Mapping.

In this study, these two hyperparameters were explored to investigate whether the total dimensions of the extracted feature and the size of the effective receptive field affected the relation of the feature to clinical information. As shown in **Figure 2**, the set of two convolution/deconvolution layers, and one pooling/unpooling layer was defined as a convolution/deconvolution "block." In this experiment, the number of blocks was set, ranging from 1 block to 4 blocks. In 4 blocks, the effective receptive field is the whole brain; in 3 blocks, it is about 30% of the brain (multiple lobes), in 2 blocks, it is 5% of the brain (multiple regions), and in 1 block it is 0.1% of the brain (1 region). The number of channels in the extraction layer was varied with 1, 4, 16, and 32 channels, but the number of channels for other layers were fixed at 32. The number of channels was considered limited to 32 due to the limitation of the current experiment's computational power. As a result, we created sixteen 3D-CAE models (4 block conditions × 4 channel conditions) to explore the effective range of hyperparameters for psychiatric brain imaging.

Other hyperparameters were fixed and common among models. The encoder was composed of convolution layers (a kernel size of 3 × 3 × 3 and a stride of 1) with rectified linear unit (ReLU) activations and average pooling layers (a kernel size of 2 × 2 × 2 and a stride of 2). The decoder was composed of convolution layers (a kernel size of 3 × 3 × 3 and a stride of 1) with ReLU activations and unpooling layers (a kernel size of 2 × 2 × 2 and a stride of 2). The loss function, consisting of the mean absolute error (MAE) between the input images and the reproduced images, was defined as follows:

$$Loss = \frac{1}{n} \sum \left| X_{input} - X_{reconstructed} \right| \tag{1}$$

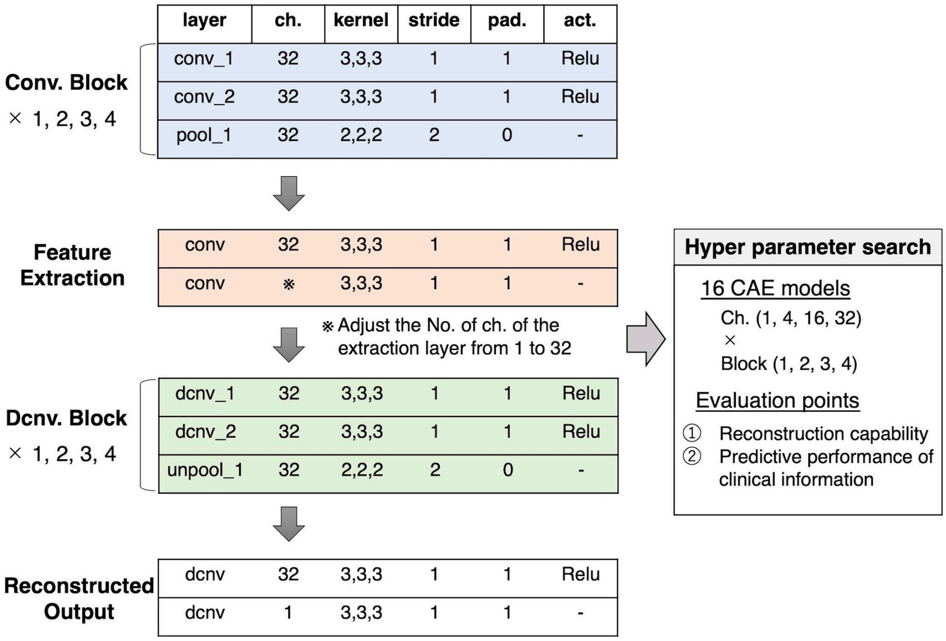
As an optimizer, we used a gradient-based method with adaptative learning rates called Adam (Kingma and Ba, 2015) (alpha = 0.0001, beta1 = 0.9, beta2 = 0.999) using mini-batches with a size of eight samples. The training process was performed with a maximum of 50,000 training iterations. We conducted the experiments in Python 3.6[2] using the Chainer v.5.4.0 library (Tokui et al., 2015).

We used a reference of training performances of 3D-CAEs, referred to as the "average brain," with which the model was assumed to output the average intensities of the training dataset regardless of the inputs. The average brain is one of the most trivial solutions where the network outputs an image without learning any information about individual differences of the inputs. The average brain was used as a reference point to indicate that the model at least reproduced individual differences. The signal intensities of voxel *i* of the average brain was determined as follows:

$$x_{ave\ i} = \frac{\sum_{s=0}^{n} x_{s,i}}{n} \tag{2}$$

where *s* is a sample from the training dataset and *n* is the number of samples.

[2] https://www.python.org/



Conv. Block
× 1, 2, 3, 4

layer	ch.	kernel	stride	pad.	act.
conv_1	32	3,3,3	1	1	Relu
conv_2	32	3,3,3	1	1	Relu
pool_1	32	2,2,2	2	0	-

Feature Extraction

layer	ch.	kernel	stride	pad.	act.
conv	32	3,3,3	1	1	Relu
conv	※	3,3,3	1	1	-

※ Adjust the No. of ch. of the extraction layer from 1 to 32

Dcnv. Block
× 1, 2, 3, 4

layer	ch.	kernel	stride	pad.	act.
dcnv_1	32	3,3,3	1	1	Relu
dcnv_2	32	3,3,3	1	1	Relu
unpool_1	32	2,2,2	2	0	-

Reconstructed Output

layer	ch.	kernel	stride	pad.	act.
dcnv	32	3,3,3	1	1	Relu
dcnv	1	3,3,3	1	1	-

Hyper parameter search

16 CAE models
Ch. (1, 4, 16, 32)
×
Block (1, 2, 3, 4)

Evaluation points
① Reconstruction capability
② Predictive performance of clinical information

FIGURE 2 | Our proposed 3D-CAE architecture. One convolution/deconvolution block was defined as repeating two convolution/deconvolution layers and one pooling/unpooling layer. The number of blocks was set from 1 to 4. The number of channels in the extraction layer was set from 1 to 32. Sixteen patterns of models with different numbers of blocks and channels were developed. In order to explore the effective number of channels and blocks, the reproduction capability and relation to clinical information were evaluated. act., activation function; 3D-CAE, three-dimensional convolutional autoencoder; ch., channel; Conv., convolution; Dcnv, deconvolution; pad., padding; pool, pooling; Relu, Rectified Linear Unit; unpool, unpooling.

Regression Analysis With Demographic and Clinical Information

Using trained 3D-CAE, latent feature vector could be extracted, and then the feature vector was flattened. The number of dimensions of that feature vector ranged from millions to hundreds, depending on the model. The relationship between the extracted feature and the clinical information was examined using regression analysis, based on the assumption that if the extracted feature is “informative,” it could help predict schizophrenia patients’ clinical information. Therefore, we confirmed this by comparing the prediction performance of 3D-CAE-based features and conventional ROI-based features. The linear regression analysis was performed with clinical and demographic information as the objective variables and the feature vectors as the explanatory variables (see the lower part of **Figure 1**). Demographic and clinical information included age, scores of positive and negative symptoms (PANSS), the dose of antipsychotic medications [chlorpromazine equivalent (CPZE)], Wechsler Adult Intelligence Scale (WAIS), duration of illness, age at onset, and diagnosis. For the regression analysis, in order to reduce the effects of correlated variables we adopted ridge regression, one of regularized linear regression methods. In the regression analysis, we executed a fivefold cross-validation process whereby the COBRE dataset was randomly divided into five groups of samples (folds), and then samples from fourfolds were used for training the regression model, and the other fold was used for the test of the regression model. The fivefold cross-validation was repeated ten times. The performance of the regression model was evaluated using the root mean square error (RMSE). The diagnosis was evaluated using accuracy.

Differences in the performances of regression models were evaluated using the two-way (number of channels × number of blocks) analysis of variance (ANOVA). Subsequently, Tukey’s multiple comparison test was performed for each group as a *post hoc* analysis. The level of significance was set to 0.05.

The 3D-CAE models were also compared with the ROI method. In the ROI method, using the automated anatomical labeling (AAL) template (Tzourio-Mazoyer et al., 2002), the GM was divided into 116 ROIs. The average intensities of each ROI were used as the ROI-based feature for regression analysis. The Student’s *t*-test was performed to compare the proposed 3D-CAE model with the ROI method. The level of significance was set to 0.05.

By calculating the gradient of the neural network at the input T1-weighted image for each subject, it is possible to visualize which regions of the input have higher weights. In this study, we attempted to visualize the regions that contribute to predicting clinical information by calculating the gradient of a composite function of feature extraction and clinical information regression functions. The calculation of a saliency map for input image *x*, *M(x)*, was defined as follows.

$$M(x) = \partial R(S(x))/\partial x \tag{3}$$

Where, *S()* was a feature extraction function based on the 3D-CAE, and *R()* was a function predicting clinical information using linear regression. To refine the visualization, the gradients’ calculation was repeated by adding Gaussian noise to the original image, similar to the technique used in SmoothGrad (Smilkov et al., 2017). The maps were then averaged by overall samples and divided by the standard deviation to obtain a *t*-value, and the values were finally converted to absolute values to yield a 3D saliency map.

Kyoto Dataset Description

A total of 172 subjects were investigated in this study, including 82 patients with schizophrenia and 90 healthy subjects. Patients were recruited from hospitals in Kyoto, Japan, and diagnosed by psychiatrists using the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV) (American Psychiatric Association., 1994) criteria for schizophrenia, confirmed with the patient edition of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID) (First et al., 1997). No patients had any comorbid DSM-IV Axis I disorder. The clinical symptoms of all patients were estimated using the Positive and Negative Syndrome Scale (PANSS) (Kay et al., 1987). Healthy subjects were screened with the non-patient edition of the SCID, confirming no history of psychiatric disorders. Exclusion criteria for all individuals included a history of head trauma, neurological illness, serious medical or surgical illness, or substance abuse. Note that participants were already diagnosed in order to expedite the data collection, but the diagnostic labels were not used to train the networks.

All participants were scanned with a 3.0-Tesla Siemens Trio scanner (Siemens Healthineers, Erlangen, Germany). The scanning parameters of the T1-weighted 3D magnetization-prepared rapid gradient-echo (3D-MPRAGE) sequences were as follows: echo time (TE) = 4.38 ms; repetition time (TR) = 2,000 ms; inversion time (TI) = 990 ms; field of view (FOV) = 225 mm × 240 mm; acquisition matrix size = 240 × 256 × 208; resolution = 0.9375 × 0.9375 × 1.0 mm^3.

COBRE Dataset Description

In this study, the COBRE dataset, which is a public dataset, was acquired as a dataset with different scanning sites and parameters to the Kyoto University dataset. All the subjects were diagnosed and screened with the SCID. The clinical symptoms of all patients were estimated using the PANSS. Exclusion criteria for individuals included a history of head trauma, neurological illness, serious medical or surgical illness, or substance abuse. We included a total of 142 subjects from this database in our study, including 71 patients with schizophrenia and 71 healthy subjects.

MRI data were acquired using a 3.0-Tesla Siemens Tim Trio scanner (Siemens Healthineers, Erlangen, Germany). The scanning parameters of the T1-weighted 3D-MPRAGE sequences were as follows: TE = 1.64 ms; TR = 2,530 ms; TI = 900 ms; FOV = 256 mm × 256 mm; acquisition matrix size = 256 × 256 × 176; resolution = 1.0 × 1.0 × 1.0 mm^3.

Demographic and clinical characteristics of Kyoto and COBRE datasets are provided in **Supplementary Table 1**. There was no significant difference between the two datasets with the exception of the sex ratio.

Division of Train, Validation, and Test

The 3D-CAE was trained using the Kyoto dataset. The dataset was randomly partitioned into training data, validation data, and test data (138 subjects, 16 subjects, and 18 subjects, respectively). Training data, validation data, and test data were used for the training of the 3D-CAE, the validation of the model during training, and the final evaluation of generalizability within the datasets independent of the training and validation data, respectively. The COBRE dataset (142 subjects) was also used to evaluate the applicability of the network to another dataset.

The regression analysis was carried out using the COBRE dataset. The bias between MRI scanning sites might have affected the distribution of features extracted by 3D-CAE; thus, affecting the prediction error of the regression. Therefore, to avoid the scanning site effect, we used a single dataset for the regression. Then the fivefold cross-validation technique was applied. Namely, the COBRE dataset samples (142 subjects) were randomly divided into five subgroups (four groups for training and one group for validation) and cross-validated by changing the combinations of groups. This fivefold cross-validation process was repeated ten times. Note that only patients with schizophrenia had clinical information available for analysis, and regressions based on the clinical information were performed using data from patients with schizophrenia (71 subjects). The details for the division of data are shown in **Table 1**.

MRI Preprocessing

The preprocessing was conducted using Statistical Parametric Mapping (SPM12, Wellcome Department of Cognitive Neurology, London, United Kingdom[3]) with the Diffeomorphic Anatomical Registration Exponentiated Lie Algebra (DARTEL) registration algorithm (Ashburner, 2007). All of the T1 whole-brain structural MRI scans were segmented into gray matter (GM), white matter, and cerebrospinal fluid. Individual GM images were normalized to the standard Montreal Neurological Institute (MNI) template with a 1.5 × 1.5 × 1.5 mm^3 voxel size and modulated for GM volumes. All normalized GM images were smoothed with a Gaussian kernel of 8 mm full width at half maximum (FWHM). Subsequently, each image was cropped to remove the background as much as possible. The GM area was extracted from original images using a binary mask, created using SPM12. As a result, the size of input images to the 3D-CAE was 121 × 145 × 121 voxels.

Subsequently, the range of signal intensities in each image was normalized with a mean of 0 and a standard deviation of 1. The standardized value of voxel i in the sample s, $x'_{s,i}$, was calculated as follows:

$$x'_{s,i} = \begin{cases} \frac{x_{s,i} - \mu_s}{\sigma_s} & (i \in \mathrm{GM}) \\ 0 & (otherwise) \end{cases} \tag{4}$$

where $x_{s,\ i}$ is the original value of intensity. μ_s and σ_s were average and standard deviation of all voxels contained in the GM area of sample s, respectively.

[3] https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

TABLE 1 | Division of dataset.

		Kyoto	COBRE
3D-CAE			
(recon. error)	Train	✓	
	Validation	✓	
	Test	✓	✓
Regression			
(pred. error)	Train		✓
	Validation		✓

The Kyoto dataset was used to develop the 3D-CAE model and was divided into train, validation and test dataset. The COBRE dataset was prepared for regression. At regression, fivefold cross-validation was performed.

3D-CAE, three-dimensional convolutional autoencoder; COBRE, Center for Biomedical Research Excellence.

RESULTS

Technical Evaluations: Reproduction Capability Performance

Figure 3A shows a representative example of learning curves for the 3D-CAE with 16 channels and 3 blocks. Progressive decreases were shown not only with "train loss" (red line), but also "validation loss" (orange line) and "test loss" (green line); this indicated that the 3D-CAE successfully learned without overfitting. The level of MAEs were well below the level of the "average brain" (dashed line) (see section "Materials And Methods" for details). In addition, the curve for "COBRE loss" (blue line) showed a similar trend. This indicated that the 3D-CAE could be applied to MRI data from another site with different scanning parameters. Similar trends of learning curves were observed for the other fifteen 3D-CAEs with different hyperparameter settings.

Figure 3B summarized the reproduction performances (MAEs for the COBRE dataset) of the sixteen 3D-CAE models with respect to the number of channels and number of blocks. Regarding the number of blocks, it can be seen that the larger the number of blocks, the larger the reproduction error. This result is intuitively understandable, in that models with smaller blocks are easier to reconstruct because extracted latent features do not abstract the original image as much (**Figure 4**). Regarding the number of channels, although the differences were small, there was a tendency for the larger number of channels to be associated with smaller reproduction errors (see **Supplementary Table 2** for more details). This result is consistent with the fact that the models with more channels have more expressive capability.

Clinical Evaluation: Relation to Clinical Information

The efficacy of the proposed method was evaluated using linear regressions for predicting demographic and clinical information related to a psychiatric disorder, i.e., schizophrenia. Demographic and clinical information, including age, the dose of antipsychotic medication (CPZE), and scores of positive and negative symptoms (PANSS), were used as an objective variable, and all extracted features of 3D-CAE were used as explanatory variables. Feature using the ROI-based method was

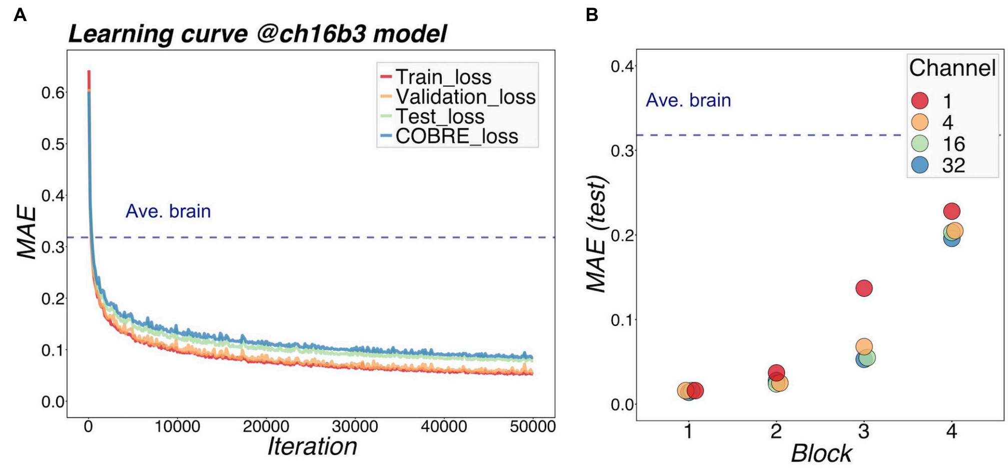


FIGURE 3 | Learning performance of models. **(A)** shows the learning loss curve for a 16-channel and 3-block model. The red line shows the training loss, indicating that the learning has progressed, and the loss has fallen sufficiently. The validation loss and test loss were also decreased, so the model was not overfitting. The blue line indicates the loss at the other site (COBRE), and the loss degraded as well. It can be seen that the MAE of our proposed models was well below the level of Ave.brain at which the model was assumed to output the average brain. This suggested that our 3D-CAE models have successfully reproduced the brain images with individual characteristics. Similar learning curves were found for other models. In **(B)**, the reproduction performance of each of the 16 models were compared. The relationships between MAE, number of channels, and number of blocks are shown. The horizontal axis indicates the number of blocks, which is color-coded by the number of channels. As the number of blocks increased, the MAE tended to be larger, and as the number of channels increased, the MAE tended to be slightly smaller. 3D-CAE, three-dimensional convolutional autoencoder; COBRE, Center for Biomedical Research Excellence; MAE, mean absolute error.

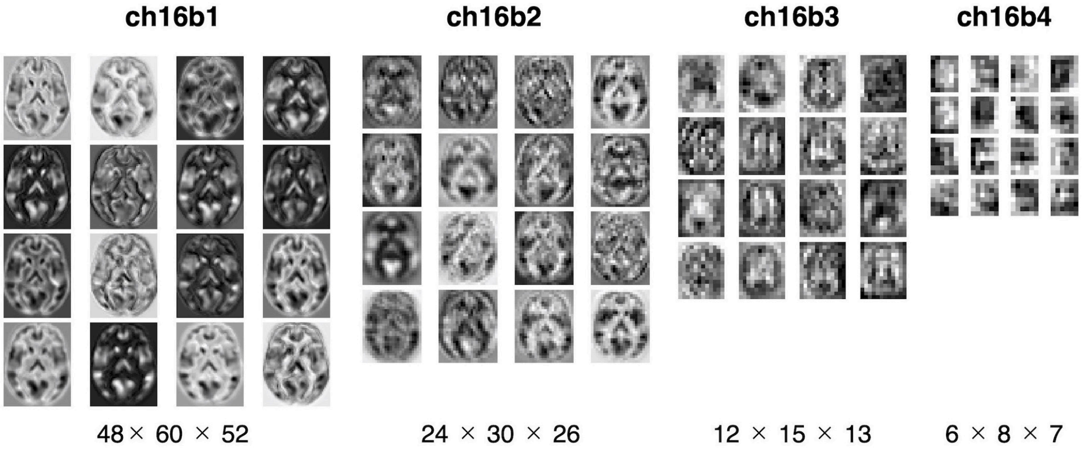


FIGURE 4 | Visualization of extracted feature. The extracted features were mapped for four models with 16 channels. From left to right: the model with one, two, three, and four blocks. The middle slices of the horizontal slice from 3D feature are shown. In the one-block model, the morphology of the brain can be seen, but with four blocks, the images are more abstract.

also used for comparison with the conventional method. A linear regression analysis was used as the simplest method to confirm if extracted features from 3D-CAEs with different hyperparameters (numbers of blocks and channels) preserved useful information. Each of the 16 3D-CAE models were analyzed 10 times, and the difference in predictive performance of the models was examined statistically.

Figure 5 illustrates a representative example of the regression analysis results. Differences in the performance of regression models (RMSE) with respect to the number of channels with 3 blocks (**Figures 5A–D**) and respect to the number of blocks with 16 channels (**Figures 5E–H**) were demonstrated as representative examples. The results of the comparison with the ROI method are shown in **Table 2**. The detailed results are described in **Supplementary Tables 3–5**, respectively.

Regarding the prediction of age, there were tendencies for the RMSEs to be smaller with increases in the number of channels (**Figure 5A**) and with decreasing number of blocks (**Figure 5B**). Indeed, statistical analysis revealed that there were significant differences between the models (channel: $p < 0.001$; block: $p < 0.001$). However, even the model with 32 channels and 1 block, which is considered one of the most predictive models, is equivalent to the ROI method ($p = 0.346$; **Table 2**), suggesting that for the

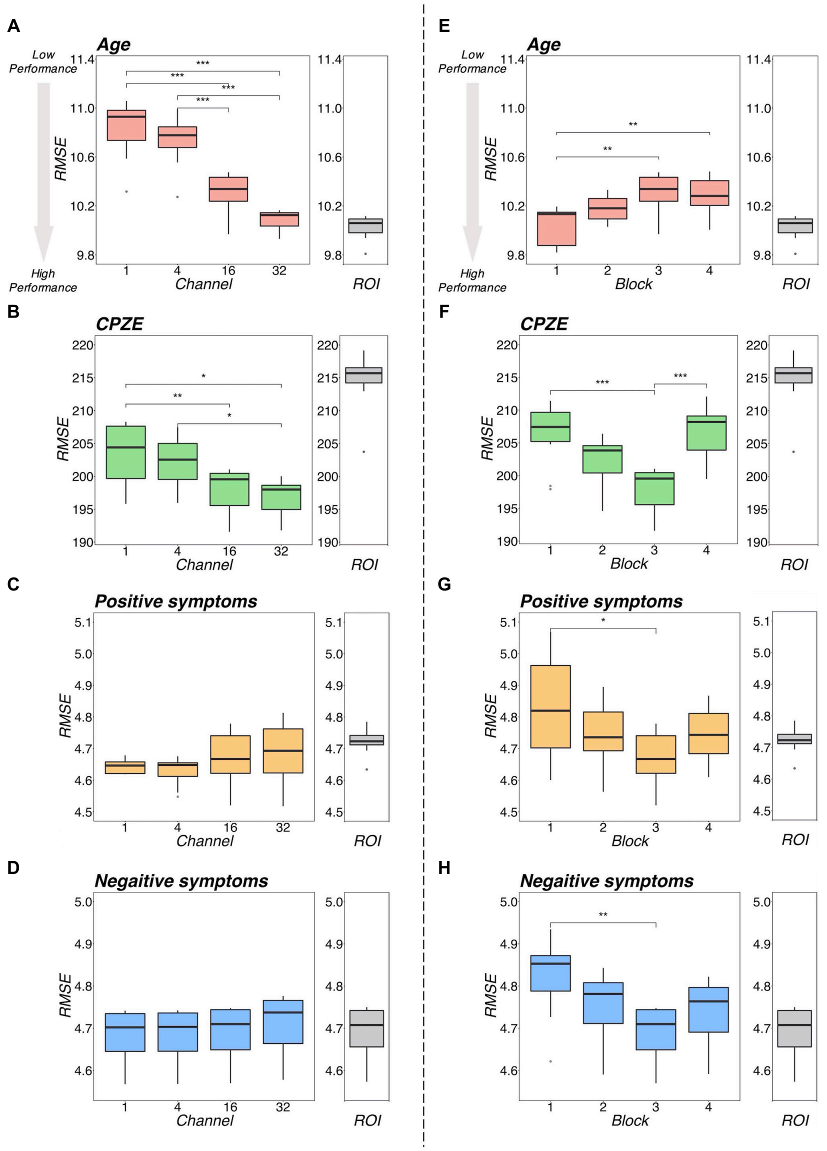


FIGURE 5 | Regression performance plot. The left side **(A–D)** shows the model differences by number of channels for the four models with 3 blocks as an example. The right side **(E–H)** shows the model differences by number of blocks for the four models, with 16 channels as representative examples. Regarding age, as shown in **(A,E)**, the RMSEs were smaller with increasing number of channels and decreasing number of blocks. Regarding CPZE, as shown in **(B,F)**, the RMSEs were smaller with increasing number of channels. On the other hand, the RMSEs may be smaller in block 3. Regarding positive symptoms and negative symptoms, as shown in **(C,D)**, there was no apparent trend in the number of channels. As shown in **(G,H)**, the RMSE may be smaller in block 3. The results of each regression with the ROI method is also included for reference. It suggests that a model with 3 blocks may be appropriate for extracting schizophrenia-related information. ****p* < 0.001, ***p* < 0.01, **p* < 0.05 (two-way analysis of variance followed by Tukey's multiple comparison test). CPZE, chlorpromazine equivalent; RMSE, root mean square error; ROI, region of interest.

prediction of age, 3D-CAE-based features were comparable to a conventional method.

In addition, the superiority of the 1 block condition was observed in the prediction of VIQ, PIQ and duration of illness (**Supplementary Tables 2–4**). However, 3D-CAEs with 1 block were not superior to the ROI method in predicting those information (VIQ: $p < 0.001$; PIQ: $p < 0.001$; duration of illness: $p = 0.100$; **Table 2**).

Regarding the prediction for CPZE, there was a tendency for the RMSEs to be smaller with increases in the number of channels (**Figure 5C**); on the other hand, the RMSEs were smallest with the condition of 3 blocks (**Figure 5D**). Statistical analysis revealed that there were significant differences between the models (channel: $p < 0.001$; block: $p < 0.001$). *Post hoc* analysis revealed that there were significant differences between 1 block and 3 blocks, and 3 blocks and 4 blocks. Moreover, the lowest level of RMSE of 3D-CAE was significantly lower than the RMSE from the ROI-based feature ($p < 0.001$; **Table 2**), indicating that for the prediction of CPZE, 3D-CAE based features outperformed a conventional method.

TABLE 2 | The results of the *t*-test.

16 channels and 3 blocks model

		3D-CAE (ch16b3)	**ROI**	***P*-value**
SZ-related clinical information	*CPZE*	197.85 (3.76)	214.75 (4.33)	< 0.001***
	Positive symptoms	4.67 (0.09)	4.72 (0.04)	$0.088^{\dagger}$
	Negative symptoms	4.67 (0.07)	4.69 (0.07)	0.968
	Duration of illness	11.87 (0.10)	11.23 (0.16)	< 0.001***
	Age of onset	7.00 (0.11)	7.47 (0.15)	< 0.001***
	Diagnosis	0.668 (0.03)	0.634 (0.02)	0.005**
Other information	*Age*	10.29 (0.18)	10.03 (0.10)	0.001**
	VIQ	14.92 (0.17)	14.72 (0.05)	0.003**
	PIQ	14.65 (0.11)	13.83 (0.09)	< 0.001***

32 channels and 1 block model

		3D-CAE (ch32b1)	**ROI**	***P*-value**
SZ-related clinical information	*CPZE*	206.57 (4.61)	214.75 (4.33)	0.001**
	Positive symptoms	4.84 (0.16)	4.72 (0.04)	0.037*
	Negative symptoms	4.89 (0.10)	4.69 (0.07)	< 0.001***
	Duration of illness	11.36 (0.17)	11.23 (0.16)	0.100
	Age of onset	7.05 (0.13)	7.47 (0.15)	< 0.001***
	Diagnosis	0.632 (0.03)	0.634 (0.02)	0.868
Other information	*Age*	9.97 (0.16)	10.03 (0.10)	0.346
	VIQ	15.17 (0.17)	14.72 (0.05)	< 0.001***
	PIQ	14.56 (0.13)	13.83 (0.09)	< 0.001***

*The differences between 3D-CAE and ROI are presented as mean (standard deviation) and p-value of RMSE. Regarding the diagnosis, it is presented as accuracy. The significantly better performances are marked in red. The 3D-CAE model with 16 channels and 3 blocks was superior to the ROI method in predicting CPZE, age of onset, and diagnosis. The model also appeared comparable or better than the ROI method in positive symptoms. The 3D-CAE model with 32 channels and 1 block was also superior to the ROI method in predicting the CPZE and age of onset. Meanwhile, that the model was comparable to the ROI method for age prediction is different from the model with 16 channels and 3 blocks. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, $^{\dagger}p < 0.1$ (t-test).*

3D-CAE, three-dimensional convolutional autoencoder; ROI, region of interest; SZ, Schizophrenia; CPZE, chlorpromazine equivalent.

Regarding the prediction of positive symptoms, there was no clear tendency with respect to the number of channels (**Figure 5E**). On the other hand, with respect to the number of blocks, the RMSEs seemed to be the smallest with the condition of 3 blocks (**Figure 5F**). Statistical analysis indicated that there were significant differences between the models (channel: $p < 0.001$; block: $p < 0.001$). *Post hoc* analysis revealed that there were significant differences between 1 block and 3 blocks. Similar trends could be observed in the prediction of negative symptoms (**Figures 5G,H**), where there were significant differences between the models (channel: $p < 0.001$; block: $p < 0.001$). In comparison to the conventional method, the 3D-CAE model with 3 blocks showed a trend toward a smaller prediction error for positive symptoms than the ROI method ($p = 0.088$; **Table 2**), the mean RMSE (SD) was 4.67 (0.09) and 4.72 (0.04), respectively, suggesting that the 3D-CAE might be comparable or better than the ROI method. Regarding the prediction of negative symptoms, there was no significant difference between 3D-CAE and the conventional method ($p = 0.968$; **Table 2**).

In addition, the superiority of the 3 blocks condition was observed in the prediction of age of onset and diagnosis (**Supplementary Tables 2–4**). Furthermore, 3D-CAEs with 3 blocks performed better than the ROI method in predicting those clinical information (age of onset: $p < 0.001$; diagnosis: $p = 0.005$; **Table 2**).

To summarize the regression analysis results, in terms of clinical information related to schizophrenia, specifically for predicting CPZE, positive symptom score, age of onset, and diagnosis, 3D-CAE with 3 blocks had better prediction than other numbers of blocks models, regardless of the number of channels. In addition, 3D-CAE with 3 blocks performed better than the ROI method in predicting clinical information. On the other hand, in terms of information not directly related to schizophrenia, such as age and intelligence, 3D-CAE with 1 block had better prediction than 3D-CAE with other numbers of blocks, regardless of the number of channels. However, 3D-CAE with 1 block did not perform better than the ROI method in predicting information not directly related to schizophrenia.

The saliency map was calculated to examine the correspondence between the features and the brain (**Figure 6**). The map showed that the regions contributing to CPZE prediction using 3D-CAE were the cerebellum, right middle temporal gyrus, the insula, posterior cingulate cortex, and precuneus. The regions that contributed to predicting the positive symptoms were found to the cerebellum, right inferior temporal gyrus, the insula, anterior and middle cingulate cortex. The other visualization results are described in **Supplementary Figure 1**.

DISCUSSION

We have shown that (1) the proposed 3D-CAEs successfully reproduced 3D MRI data with sufficiently low errors, and (2) the diagnostic label-free features extracted using 3D-CAE retained the relation of various clinical information. In addition, we explored the appropriate hyperparameter range of 3D-CAE, and our results suggest that a model with 3 blocks-based features might preserve information related to the medication dose and the severity of positive symptoms in patients with schizophrenia.

The reproduction errors of 3D-CAE were lower than the average brain level, indicating that the proposed 3D-CAEs successfully reproduced 3D brain MRI data with individual characteristics. In addition, the 3D-CAE trained with the Kyoto dataset was applicable to the COBRE dataset with different

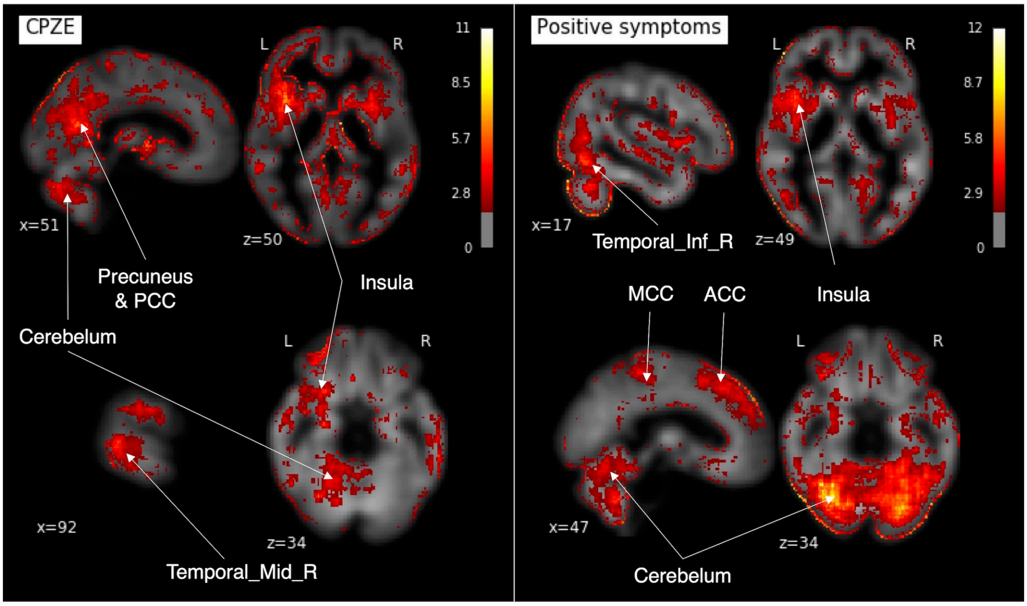


FIGURE 6 | The saliency maps. In our developed 3D-CAE model, the saliency maps of the signals contributing to the prediction of each clinical information were obtained by calculating the gradient of the neural network. 3D-CAE, three-dimensional convolutional autoencoder; PCC, posterior cingulate cortex; MCC, middle cingulate cortex; ACC, anterior cingulate cortex; Temporal_Mid_R, right middle temporal gyrus; Temporal_Inf_R, right inferior temporal gyrus.

scanners and scanning parameters. Although the current study was tested using only two datasets, the results suggested that the proposed method may have applicability to data from multiple sites and scanners, itself a challenging issue in neuroimaging studies (Jovicich et al., 2009; Schnack et al., 2010; Fortin et al., 2018; Dewey et al., 2019; Yamashita et al., 2019).

Regression analyses demonstrated that the 3D-CAE-based feature was comparable or more effective than the ROI-based feature in predicting the medication dose and the severity of positive symptoms in patients with schizophrenia, even though 3D-CAE-based features were extracted without using a diagnostic label of schizophrenia. Because this approach enabled us to extract neuroimaging features of individuals without information on the clinical diagnosis, it can be useful for heterogeneous population data. Furthermore, using this approach, we were able to predict clinical variables. These imply that our approach in this study could be an alternative method to conventional methods based on categorical diagnostic information. This study showed that the prediction of CPZE, positive symptoms, and age of onset might be more improved in 3D-CAE than ROI. These are clinically meaningful because the model would help clinicians decide the treatment plan by predicting based on an objective indicator. In contrast, the current medication dose is mainly adjusted based on the patient's self-reported condition.

Regarding the number of channels, 16— to 32-channel models demonstrated better performance. This is easy to understand because the more channels the model has, the more expressive it is (Zhu et al., 2019). However, since increasing number of channels inevitably results in increasing computational power needs, estimation of the appropriate number of channels is still important. Our results suggest that the number of channels may be sufficient at 16 or 32 for reproducing structural brain MRI scans. Regarding the number of blocks, our results indicated that information from a local receptive field (small number of block) was sufficient for predicting age. However, predicting schizophrenia-related clinical data required information from more global receptive fields (larger block numbers, such as 3-block). As the number of blocks increase, the effective receptive fields expand, and the global feature of the brain can be extracted (Szegedy et al., 2015; Le and Borji, 2017; Luo et al., 2017). In our model, the 3 blocks model contained eight convolutional layers, and effective receptive fields of the feature unit were about 68 × 68 × 68 voxels, corresponding to about 30% of the brain. This fact is consistent with the previous neuroimaging studies showing that the medication dose and symptoms severity are associated with the volume of multiple brain regions, including the temporal lobe, frontal lobe, and various subcortical regions (García-Martí et al., 2008; Palaniyappan et al., 2013; Van Erp et al., 2016; van Erp et al., 2018; Bullmore, 2019; Fan et al., 2019). The 3D-CAE-based feature's superiority may be related to the detection of local signal interactions inherent in the convolutional methods; this contrasts with the ROI-based method, in which signals within each ROI are averaged and the interactions of local signals are discarded.

In our model, the saliency map showed that the cerebellum, temporal lobe, cingulate gyrus, and insular cortex had greater contributions in predicting the severity of symptoms and dose of antipsychotic medication. The present study results were consistent with the results of previous studies showing that positive symptoms and CPZE correlated with cortical thickness thinning in the temporal lobe (van Erp et al., 2018), and that cerebellar atrophy was associated with positive symptoms (Cierpka et al., 2017). The insular and cingulate cortices, which were shown to be significant contributors to clinical variables in the present study, have been repeatedly reported to be reduced in the regional brain volume in schizophrenia (Glahn et al., 2008; Takayanagi et al., 2013; Gupta et al., 2015; Uwatoko et al., 2015). However, the relationship between these areas and positive symptoms and CPZE requires further investigation. As a side note, because the relatively high values of the edge of the brain may be influenced by the traits of Smoothgrad (Smilkov et al., 2017) that emphasize the edge, it was difficult to consider them from a neuroimaging study perspective.

There are some limitations to our study. First, this study only explored a limited range of hyperparameters. In CAE, there are several hyperparameters than those explored, such as activate function, optimizer, and learning rate. However, because we focused on the total dimension of the extracted features and the effective receptive field's size, the numbers of blocks and channels were explored as the target variables. In addition, the exploration range of hyperparameters was limited due to practical reasons including the computational power and costs of the experiments.

Second, the differences in preprocessing of neuroimaging data may affect the robustness of the study results. In this study, we employed the standard preprocessing methods (e.g., image resolution, standardization, smoothing), which have been used in neuroimaging studies, such as voxel-based morphometry (Ashburner and Friston, 2000). Nevertheless, further studies may evaluate the effects of the preprocessing methods on results.

Third, the datasets used in this study only included patients diagnosed with schizophrenia as well as healthy subjects. Considering the heterogeneity of psychiatric disorders, it will be necessary to examine the applicability of diagnostic label-free feature extraction using 3D-CAE to other psychiatric disorders in the future.

Fourth, regressions were used to predict clinical and demographic scores, but the 3D-CAE-based feature outperformed the feature of the ROI does not necessarily prove that the predictive value generated is clinically useful. In the present study, the main goal was feature extraction, and only simple regression was used for prediction. The additional experiments with the development of a fine-tuned model and evaluation using longitudinal data of disease process are needed in the future. These may improve clinical decisions for assessing patients' prognosis and estimating an appropriate medication dose.

In this paper, we presented 3D-CAE-based feature extraction for brain structural imaging of psychiatric disorders. We found that 3D-CAE can extract features that retained their relation to clinical information from 3D MRI data without diagnostic labels. Our data suggest that 3D-CAE models with effective hyperparameter settings may extract information related to the medication dose and the severity of symptoms in patients with schizophrenia. The feature extraction without using diagnostic labels based on the current diagnostic criteria is scientifically significant and may lead to the development of alternative data-driven diagnostic criteria.

DATA AVAILABILITY STATEMENT

All data generated or analyzed during this study are included in this published article. The primary data can be obtained from public databases, including the Decoded Neurofeedback (DecNef) Project Brain Data Repository (https://bicr-resource.atr.jp/srpbs1600/) and the Centers for Biomedical Research Excellence (COBRE; http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html).

ETHICS STATEMENT

All study participants signed an informed consent form. The study was performed in accordance with the current Ethical Guidelines for Medical and Health Research Involving Human Subjects in Japan and was approved by the Committee on Medical Ethics of Kyoto University and National Center of Neurology and Psychiatry.

AUTHOR CONTRIBUTIONS

HY, YH, and YY conceived, designed the research, and drafted the manuscript. HY and YH conducted the deep learning experiments and analyzed the data. GS, JM, TM, and HT collected MRI data. GS, JM, TM, HT, MH, and AH provided critical revisions. All authors contributed to and have approved the final manuscript.

FUNDING

This work was partially supported by JSPS KAKENHI Grant Nos. JP17H00740, JP17H04248, JP17H06039, JP18KT0021, JP18K07597, JP19H04998, JP19K17077, JP20H00001, and JP20H00625, Grant-in-Aid for Scientific Research on Innovative Areas from the MEXT Grant No. JP20H05064, Strategic International Brain Science Research Promotion Program by the Japan Agency for Medical Research and Development Grant No. 19dm0307008h0002, JST CREST Grant No. JPMJCR16E2 and a Grant from SENSHIN Medical Research Foundation.

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at https://www.biorxiv.org/content/10.1101/2020.08.24.213447v1. We thank all the authors of the included studies. We also would like to thank Editage (www.editage.com) for English language editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.652987/full#supplementary-material

REFERENCES

Aghdam, M. A., Sharifi, A., and Pedram, M. M. (2019). Diagnosis of autism spectrum disorders in young children based on resting-state functional magnetic resonance imaging data using convolutional neural networks. *J. Digit. Imaging* 32, 899–918. doi: 10.1007/s10278-019-00196-1

American Psychiatric Association. (1994). *Diagnostic And Statistical Manual Of Mental Disorders?: DMS-IV*. Washington, DC: American Psychiatric Publishing.

American Psychiatric Association. (2013). *Diagnostic And Statistical Manual Of Mental Disorders (DSM-5)*. Washington, DC: American Psychiatric Publishing.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage* 38, 95–113. doi: 10.1016/j.neuroimage.2007.07.007

Ashburner, J., and Friston, K. J. (2000). Voxel-based morphometry – the methods. *Neuroimage* 11, 805–821. doi: 10.1006/nimg.2000.0582

Bullmore, E. (2019). Cortical thickness and connectivity in schizophrenia. *Am. J. Psychiatry* 176, 505–506. doi: 10.1176/appi.ajp.2019.19050509

Cierpka, M., Wolf, N. D., Kubera, K. M., Schmitgen, M. M., Vasic, N., Frasch, K., et al. (2017). Cerebellar contributions to persistent auditory verbal hallucinations in patients with schizophrenia. *Cerebellum* 16, 964–972. doi: 10.1007/s12311-017-0874-5

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31, 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Dewey, B. E., Zhao, C., Reinhold, J. C., Carass, A., Fitzgerald, K. C., Sotirchos, E. S., et al. (2019). DeepHarmony: a deep learning approach to contrast harmonization across scanner changes. *Magn. Reson. Imaging* 64, 160–170. doi: 10.1016/j.mri.2019.05.041

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019). A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29. doi: 10.1038/s41591-018-0316-z

Fan, F., Xiang, H., Tan, S., Yang, F., Fan, H., Guo, H., et al. (2019). Subcortical structures and cognitive dysfunction in first episode schizophrenia. *Psychiatry Res. Neuroimaging* 286, 69–75. doi: 10.1016/j.pscychresns.2019.01.003

Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., and Fair, D. A. (2019). The Heterogeneity problem: approaches to identify psychiatric subtypes. *Trends Cogn. Sci.* 23, 584–601. doi: 10.1016/j.tics.2019.03.009

First, M. B., Spitzer, R. L., Gibbon, M., and Williams, J. B. W. (1997). *Structured Clinical Interview for DSM-IV Axis I Disorders SCID-I*. Washington, DC: American Psychiatric Publishing.

Fornito, A., Zalesky, A., Pantelis, C., and Bullmore, E. T. (2012). Schizophrenia, neuroimaging and connectomics. *Neuroimage* 62, 2296–2314. doi: 10.1016/j.neuroimage.2011.12.090

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* 167, 104–120. doi: 10.1016/j.neuroimage.2017.11.024

Fusar-Poli, P., Howes, O., Bechdolf, A., and Borgwardt, S. (2012). Mapping vulnerability to bipolar disorder: a systematic review and meta-analysis of neuroimaging studies. *J. Psychiatry Neurosci.* 37, 170–184. doi: 10.1503/jpn.110061

Gao, J., Chen, M., Li, Y., Gao, Y., Li, Y., Cai, S., et al. (2021). Multisite autism spectrum disorder classification using convolutional neural network classifier and individual morphological brain networks. *Front. Neurosci.* 14:629630. doi: 10.3389/fnins.2020.629630

García-Martí, G., Aguilar, E. J., Lull, J. J., Martí-Bonmatí, L., Escartí, M. J., Manjón, J. V., et al. (2008). Schizophrenia with auditory hallucinations: a voxel-based morphometry study. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 32, 72–80. doi: 10.1016/j.pnpbp.2007.07.014

Glahn, D. C., Laird, A. R., Ellison-Wright, I., Thelen, S. M., Robinson, J. L., Lancaster, J. L., et al. (2008). Meta-analysis of gray matter anomalies in schizophrenia: application of anatomic likelihood estimation and network analysis. *Biol. Psychiatry* 64, 774–781. doi: 10.1016/j.biopsych.2008.03.031

Guo, X., Liu, X., Zhu, E., and Yin, J. (2017). "Deep clustering with convolutional autoencoders," in *Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, eds D. Liu, S. Xie, Y. Li, D. Zhao, and E. S. El-Alfy (Cham: Springer), 373–382. doi: 10.1007/978-3-319-70096-0_39

Gupta, C. N., Calhoun, V. D., Rachakonda, S., Chen, J., Patel, V., Liu, J., et al. (2015). Patterns of gray matter abnormalities in schizophrenia based on an international mega-analysis. *Schizophr. Bull.* 41, 1133–1142. doi: 10.1093/schbul/sbu177

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *Neuroimage Clin.* 17, 16–23. doi: 10.1016/j.nicl.2017.08.017

Hinton, G. E. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Jovicich, J., Czanner, S., Han, X., Salat, D., van der Kouwe, A., Quinn, B., et al. (2009). MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *Neuroimage* 46, 177–192. doi: 10.1016/j.neuroimage.2009.02.010

Kay, S. R., Fiszbein, A., and Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr. Bull.* 13, 261–276. doi: 10.1093/schbul/13.2.261

Kingma, D. P., and Ba, J. L. (2015). "Adam: a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015* (Ithaca, NY).

Le, H., and Borji, A. (2017). What are the Receptive, Effective Receptive, and Projective Fields of Neurons in Convolutional Neural Networks?. Available onlne at: http://arxiv.org/abs/1705.07049 (accessed May, 2020).

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, S., Ripke, S., Neale, B. M., Faraone, S., Purcell, S., Rh, P., et al. (2014). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs cross-disorder group of the psychiatric genomics consortium. *Nat. Genet.* 45, 984–994. doi: 10.1038/ng.2711

Linden, D. E. J. (2012). The challenges and promise of neuroimaging in psychiatry. *Neuron* 73, 8–22. doi: 10.1016/j.neuron.2011.12.014

Luo, W., Li, Y., Urtasun, R., and Zemel, R. (2017). Understanding the effective receptive field in deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 4905–4913. Available at: http://arxiv.org/abs/1701.04128 (accessed May, 2020).

Martinez-Murcia, F. J., Ortiz, A., Gorriz, J.-M., Ramirez, J., and Castillo-Barnes, D. (2020). Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders. *IEEE J. Biomed. Health Inform.* 24, 17–26. doi: 10.1109/JBHI.2019.2914970

Nelson, B. G., Bassett, D. S., Camchong, J., Bullmore, E. T., and Lim, K. O. (2017). Comparison of large-scale human brain functional and anatomical networks in schizophrenia. *Neuroimage Clin.* 15, 439–448. doi: 10.1016/j.nicl.2017.05.007

Nishio, M., Nagashima, C., Hirabayashi, S., Ohnishi, A., Sasaki, K., Sagawa, T., et al. (2017). Convolutional auto-encoders for image denoising of ultra-low-dose CT. *Heliyon* 3:e00393. doi: 10.1016/j.heliyon.2017.e00393

Oh, K., Kim, W., Shen, G., Piao, Y., Kang, N. I., Oh, I. S., et al. (2019). Classification of schizophrenia and normal controls using 3D convolutional neural network and outcome visualization. *Schizophr. Res.* 212, 186–195. doi: 10.1016/j.schres.2019.07.034

Olesen, J., Gustavsson, A., Svensson, M., Wittchen, H.-U., and Jönsson, B. (2012). The economic cost of brain disorders in Europe. *Eur. J. Neurol.* 19, 155–162. doi: 10.1111/j.1468-1331.2011.03590.x

Owen, M. J. (2014). New approaches to psychiatric diagnostic classification. *Neuron* 84, 564–571. doi: 10.1016/j.neuron.2014.10.028

Owen, M. J., Sawa, A., and Mortensen, P. B. (2016). Schizophrenia. *Lancet* 388, 86–97. doi: 10.1016/S0140-6736(15)01121-6

Palaniyappan, L., Marques, T. R., Taylor, H., Handley, R., Mondelli, V., Bonaccorso, S., et al. (2013). Cortical folding defects as markers of poor treatment response in first-episode psychosis. *JAMA Psychiatry* 70, 1031–1040. doi: 10.1001/jamapsychiatry.2013.203

Pinaya, W. H. L., Mechelli, A., and Sato, J. R. (2019). Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: a large-scale multi-sample study. *Hum. Brain Mapp.* 40, 944–954. doi: 10.1002/hbm.24423

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., et al. (2014). Deep learning for neuroimaging: a validation study. *Front. Neurosci.* 8:229. doi: 10.3389/fnins.2014.00229

Poldrack, R. A. (2007). Region of interest analysis for fMRI. *Soc. Cogn. Affect. Neurosci.* 2, 67–70. doi: 10.1093/scan/nsm006

Quaak, M., van de Mortel, L., Mani Thomas, R., and van Wingen, G. (2021). Deep learning applications for the classification of psychiatric disorders using neuroimaging data: systematic review and meta-analysis. *Neuroimage Clin.* 30:102584. doi: 10.1016/j.nicl.2021.102584

Qureshi, M. N. I., Oh, J., and Lee, B. (2019). 3D-CNN based discrimination of schizophrenia using resting-state fMRI. *Artif. Intell. Med.* 98, 10–17. doi: 10.1016/j.artmed.2019.06.003

Ratnanather, J. T., Poynton, C. B., Pisano, D. V., Crocker, B., Postell, E., Cebron, S., et al. (2013). Morphometry of superior temporal gyrus and planum temporale in schizophrenia and psychotic bipolar disorder. *Schizophr. Res.* 150, 476–483. doi: 10.1016/j.schres.2013.08.014

Sarraf, S., DeSouza, D. D., Anderson, J., and Tofighi, G. (2017). DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. *bioRxiv*[Preprint] 070441. doi: 10.1101/070441

Schnack, H. G., van Haren, N. E. M., Brouwer, R. M., van Baal, G. C. M., Picchioni, M., Weisbrod, M., et al. (2010). Mapping reliability in multicenter MRI: voxel-based morphometry and cortical thickness. *Hum. Brain Mapp.* 31, 1967–1982. doi: 10.1002/hbm.20991

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). *SmoothGrad: Removing Noise by Adding Noise*. Available online at: http://arxiv.org/abs/1706.03825 (accessed June, 2019).

Sugihara, G., Oishi, N., Son, S., Kubota, M., Takahashi, H., and Murai, T. (2017). Distinct patterns of cerebral cortical thinning in schizophrenia: a neuroimaging data-driven approach. *Schizophr. Bull.* 43, 900–906. doi: 10.1093/schbul/sbw176

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Washington, DC: IEEE Computer Society), 1–9. doi: 10.1109/CVPR.2015.7298594

Takayanagi, M., Wentz, J., Takayanagi, Y., Schretlen, D. J., Ceyhan, E., Wang, L., et al. (2013). Reduced anterior cingulate gray matter volume and thickness in subjects with deficit schizophrenia. *Schizophr. Res.* 150, 484–490. doi: 10.1016/j.schres.2013.07.036

Tokui, S., Oono, K., Hido, S., and Clayton, J. (2015). "Chainer: a next-generation open source framework for deep learning," in *Proceedings of Workshop On Machine Learning Systems (LearningSys) in the 29th Annual Conference On Neural Information Processing Systems (NIPS)*, (San Diego, USA: The Neural Information Processing Systems Foundation), 1–6.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978

Uwatoko, T., Yoshizumi, M., Miyata, J., Ubukata, S., Fujiwara, H., Kawada, R., et al. (2015). Insular gray matter volume and objective quality of life in schizophrenia. *PLoS One* 10:e0142018. doi: 10.1371/journal.pone.0142018

Van Erp, T. G. M., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., et al. (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Mol. Psychiatry* 21, 547–553. doi: 10.1038/mp.2015.63

van Erp, T. G. M., Walton, E., Hibar, D. P., Schmaal, L., Jiang, W., Glahn, D. C., et al. (2018). Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the enhancing neuro imaging genetics through meta analysis (ENIGMA) consortium. *Biol. Psychiatry* 84, 644–654. doi: 10.1016/j.biopsych.2018.04.023

van Os, J., and Kapur, S. (2009). Schizophrenia. *Lancet* 374, 635–645. doi: 10.1016/S0140-6736(09)60995-8

Vieira, S., Pinaya, W. H. L., and Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci. Biobehav. Rev.* 74, 58–75. doi: 10.1016/j.neubiorev.2017.01.002

Wang, S. H., Phillips, P., Sui, Y., Liu, B., Yang, M., and Cheng, H. (2018). Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling. *J. Med. Syst.* 42:85. doi: 10.1007/s10916-018-0932-7

Wang, Z., Sun, Y., Shen, Q., and Cao, L. (2019). Dilated 3D convolutional neural networks for brain MRI data classification. *IEEE Access* 7, 134388–134398. doi: 10.1109/ACCESS.2019.2941912

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., et al. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. *Lancet* 382, 1575–1586. doi: 10.1016/S0140-6736(13)61611-6

World Health Organization. (1992). *International Statistical Classification Of Diseases And Related Health Problems*. Geneva: World Health Organization.

Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., et al. (2019). Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLoS Biol* 17:e3000042. doi: 10.1371/journal.pbio.3000042

Zhu, H., An, Z., Yang, C., Hu, X., Xu, K., and Xu, Y. (2019). *Rethinking the Number of Channels for the Convolutional Neural Network*. Available online at: http://arxiv.org/abs/1909.01861 (accessed May, 2020).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.