Schizophrenia Research 212 (2019) 186–195

Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

Classification of schizophrenia and normal controls using 3D convolutional neural network and outcome visualization

Kanghan Oh [a], Woosung Kim [b, c], Guangfan Shen [b, c], Yanhong Piao [b, c], Nam-In Kang [d], Il-Seok Oh [a], Young Chul Chung [b, c, *]

[a] Division of Computer Science and Engineering, Chonbuk National University, 567 Baekje-daero, Deokjin-gu, Jeonju-si 54896, Republic of Korea
[b] Research Institute of Clinical Medicine of Chonbuk National University-Biomedical Research Institute of Chonbuk National University Hospital, Jeonju-si 54907, Republic of Korea
[c] Department of Psychiatry, Chonbuk National University Medical School, Jeonju-si 54907, Republic of Korea
[d] Department of Psychiatry, Maeumsarang Hospital, Wanju, Jeollabuk-do, Republic of Korea

ARTICLE INFO

Article history:
Received 11 September 2018
Received in revised form
16 May 2019
Accepted 21 July 2019
Available online 6 August 2019

Keywords:
Classification accuracy
Convolutional neural network
Support vector machine
Saliency map
Schizophrenia

ABSTRACT

Background: The recent deep learning-based studies on the classification of schizophrenia (SCZ) using MRI data rely on manual extraction of feature vector, which destroys the 3D structure of MRI data. In order to both identify SCZ and find relevant biomarkers, preserving the 3D structure in classification pipeline is critical.
Objectives: The present study investigated whether the proposed 3D convolutional neural network (CNN) model produces higher accuracy compared to the support vector machine (SVM) and other 3D-CNN models in distinguishing individuals with SCZ spectrum disorders (SSDs) from healthy controls. We sought to construct saliency map using class saliency visualization (CSV) method.
Methods: Task-based fMRI data were obtained from 103 patients with SSDs and 41 normal controls. To preserve spatial locality, we used 3D activation map as input for the 3D convolutional autoencoder (3D-CAE)-based CNN model. Data on 62 patients with SSDs were used for unsupervised pretraining with 3D-CAE. Data on the remaining 41 patients and 41 normal controls were processed for training and testing with CNN. The performance of our model was analyzed and compared with SVM and other 3D-CNN models. The learned CNN model was visualized using CSV method.
Results: Using task-based fMRI data, our model achieved 84.15%~84.43% classification accuracies, outperforming SVM and other 3D-CNN models. The inferior and middle temporal lobes were identified as key regions for classification.
Conclusions: Our findings suggest that the proposed 3D-CAE-based CNN can classify patients with SSDs and controls with higher accuracy compared to other models. Visualization of salient regions provides important clinical information.

1. Introduction

Schizophrenia (SCZ) is a complex psychiatric disorder characterized by aberrant sensory perceptions, concrete thinking, and a restricted range of emotion. As SCZ is a heterogenous disorder that is currently diagnosed based on operational criteria, reliable diagnosis is challenging. A diagnostic system aided by neuroimaging data would allow for a more objective approach and may increase diagnostic accuracy and our ability to predict treatment response and prognosis.

Traditional analytic methods frequently used to explore the neuroimaging biomarkers for the classification of neuropsychiatric disorders have been based on mass univariate statistics under the assumption that different regions act independently. However, this assumption is not appropriate in terms of our current understanding of brain functioning (Foxet et al., 2005). Since machine learning (ML) methods have been able to take the inter-correlation between regions into account, ML has been gaining considerable attention due to its advantages over conventional analytic methods. Support vector machine (SVM), a supervised technique that works by estimating the optimal hyperplane that best separates two classes, is among the most popular ML methods. Despite its popularity, SVM has been criticized for not performing well on raw data and requiring the expert use of design techniques to extract the less redundant and more informative features (LeCun et al., 2015; Plis et al., 2014). An alternative family of ML methods, known as deep learning (DL), has been introduced to medical image analysis and has yielded promising results in multiple applications, including computerized prognoses for Alzheimer's disease (Hosseini-Asl et al., 2018; Kam et al.,

* Corresponding author at: Chonbuk National University Medical School, Geonjiro 20, Jeonju, Republic of Korea.
E-mail address: chungyc@jbnu.ac.kr (Y.C. Chung).

https://doi.org/10.1016/j.schres.2019.07.034
0920-9964/

2018; Korolev et al., 2017; Liu et al., 2014; Rieke et al., 2018; Suk and Shen, 2013), tumor segmentation (Havaei et al., 2017), and histopathological diagnoses (Cireşan et al., 2013; Litjens et al., 2016). An essential aspect of DL that differentiates it from other ML methods is that the features are not manually extracted; instead, they are learned from the data, resulting in a more objective and less bias-prone process. Additionally, the ability to achieve higher orders of abstraction and complexity relative to other ML methods, such as SVM, renders DL better suited for detecting complex, scattered, and subtle patterns in data (Plis et al., 2014).

With respect to SCZ, a limited number of studies have been performed using DL methods. Using structural magnetic resonance imaging (sMRI) data, Plis et al. (2014) and Pinaya et al. (2016) applied a deep belief network (DBN) to the original pre-processed images and obtained accuracy rates of 91% and 73.6%, respectively. Using resting-state functional MRI (rsfMRI) data, Kim et al. (2016) classified SCZ patients and controls with an accuracy of 85.5% based on a new model consisting of a stacked autoencoder (SAE) with weight sparsity control. Other studies (Han et al., 2017; Patel et al., 2016; Zeng et al., 2018) reported accuracy rates ranging from 79 to 92% in schizophrenia using AE-based two or three-stage architecture. All of the aforementioned methods adopted DBN or SAE-based network that requires one-dimensional (1D) input feature vector. It is critical to note that they lose important spatial information between the voxels or regions (i.e., spatial locality) in the process of transforming 3D or 4D volume data into 1D data. Losing the spatial locality, we also lose a way to identify and visualize important biomarkers from the learned neural networks. The convolutional neural networks (CNNs), a special type of feed-forward neural network, can use original 3D data while preserving spatial locality. Owing to two intrinsic properties of CNNs, weight sharing and local connectivity, this approach can also reduce the number of trainable weights to be estimated, making it computationally possible to operate the network at the voxel level. Despite the evidence that CNNs can generate the most diagnostically accurate results (e.g., in Alzheimer's disease) (Hosseini-Asl et al., 2018, Korolev et al., 2017, Rieke et al., 2018), there has been no research on the use of CNNs in SCZ.

The primary aim of the present study was to investigate whether our proposed model using a 3D-convolutional autoencoder (CAE) based CNN produces higher classification accuracy compared to the SVM in distinguishing individuals with SCZ spectrum disorders (SSDs) from healthy controls. The performance of our model was also compared to other 3D CNN models. Furthermore, we sought to develop an easily understood saliency map of the outcomes of our CNN model using a class saliency visualization method, which is available only by applying the end-to-end DL technique.

2. Methods

2.1. Participants

All participants ($n = 103$) met DSM-IV-TR criteria for SSDs (SCZ: 74; schizoaffective disorder: 7; and schizophreniform disorder: 22) according to the Structured Clinical Interview for DSM-IV (SCID) (First et al., 1996; Han and Hong, 2000). Other inclusion criteria were between 18 and 59 years of age and ability to provide written informed consent. We excluded patients who reported alcohol or drug dependency within the past 6 months and those with intellectual disabilities, a current or historical neurological illness, a serious internal or external illness, or a positive urine pregnancy test. The severity of symptoms within 1 week before the fMRI experiment was evaluated with the Positive and Negative Syndrome Scale (Kay et al., 1987), which was administered by trained psychiatrists. Healthy individuals ($n = 41$) were recruited for the control group via advertisements; these participants were age- and sex-matched to the patient group and interviewed using the screening module of the SCID-IV, non-patient edition. They were required to have no previous or current psychiatric disorder, neurological disorder, or significant medical condition. The detailed characteristics of participants and statistical comparisons between patients with SSDs and controls are presented in Table 1. All subjects joined this study voluntarily and provided written informed consent. The study was approved by the Ethics Committee of Chonbuk National University Hospital (Approval number CUH 2012-08-001).

Table 1
Demographic and clinical characteristics of the participants.

Variable	Schizophrenia spectrum disorder ($N = 103$)	Normal control ($N = 41$)	P
Age (years)	32.46 ± 9.21	33.98 ± 8.53	0.470
Education (years)	13.59 ± 2.76	16.22 ± 2.57	0.727
Sex (male/female)	57/46	24/17	0.727
Age of onset (years)	27.20 ± 9.00	–	–
Duration of illness (months)	56.28 ± 50.96	–	–
PANSS-total	49.93 ± 18.98	–	–
PANSS-positive	13.09 ± 6.70	–	–
PANSS-negative	12.09 ± 5.48	–	–
PANSS-general	24.76 ± 9.10	–	–

Values are presented as mean ± standard deviation or number.
PANSS, Positive and Negative Syndrome Scale.

2.2. Experimental task

Participants performed a task that was designed to elicit ideas of reference (IORs) using negative and neutral images drawn from Google images and the International Affective Picture System (Lang et al., 1998). Candidate negative images judged to evoke feelings similar to IORs were first selected by the investigators. Normal volunteers evaluated candidate images in terms of appropriateness for IOR evocation, ease of understanding, and negative valence using a visual analogue scale (VAS) ranging from 0 to 10. Images rated ≥7 on all items were selected for inclusion in the task. The scripts for auditory stimuli depicting negative images (i.e., talking behind your back or whispering about you) were initially created by the investigators and then selected based on appropriateness for eliciting IORs, ease of understanding, and negative valence (scores rated ≥7 by volunteers). The scripts were recorded by voice actors. The procedure for selecting neutral images and matched scripts were the same. The task consisted of three conditions: the activation condition, which involved negative images and matched auditory stimuli; the control condition, which involved neutral images and matched auditory stimuli; and the resting condition, which involved a fixation cross and no auditory stimuli. One trial (90 s) consisted of the presentation of activating (three audio-visual stimuli, each lasting 9 s) and control (three audio-visual stimuli, each lasting 9 s) blocks and two resting conditions (each lasting 18 s). The duration of the entire experiment, which consisted of five trials, was 450 s. Stimuli were presented and responses recorded using Integrated Functional Imaging System software (MRI Devices Corporation, Waukesha, WI, USA) and E-Prime system (Psychology Software Tools, Inc., Pittsburgh, PA, USA).

Participants were instructed to press the right or left button if their reaction to the stimulus was unpleasant or neutral, respectively. They could respond at any point before presentation of the subsequent stimulus. Response types and reaction times were automatically transferred to the computer files. The behavioral response was intended to maintain the participant's attention. A few hours before fMRI scanning, all participants practiced the fMRI task. After completing the task in the scanner, participants were again presented with the stimuli and asked to use the VAS to rate the valence, self-referential quality, and malevolence of each stimulus.

2.3. Image acquisition and preprocessing

The fMRI data were acquired at the Chonbuk National University Hospital on a 3 T Siemens Verio scanner (Magnetom Sonata A.G.; Siemens Medical Solutions, Erlangen, Germany) with 45 mT/m gradients and a 12-channel standard quadrature head coil (Siemens, MAGNETON Sonata). The transaxial functional scans were acquired using a single-shot, T2*-weighted gradient-echo-planar imaging sequence depicting the blood-oxygenation level-dependent signal with the following parameters: repeat time (TR) = 3 s, echo time (TE) = 30 ms, field of view = 240 mm, acquisition matrix = 64 × 64, flip angle = 90°, gap = 0.8 mm, 152 volumes (36 contiguous axial 3-mm-thick slices), and even-first interleaved acquisition. Axial 1.0-mm-thick T1-weighted images (TR = 1.9 s; TE = 2.5 ms, flip angle = 9°; field of view = 240 mm; and image matrix = 256 × 246 mm) were also collected. The first three scans were discarded to allow for signal saturation. Spatial preprocessing was performed using SPM8 software (www.fil.ion.ucl.ac.uk/spm). Corrections for differences in slice acquisition time were performed using user-specified sequences. Head motion was corrected by realignment, and corrected images were co-registered to the T1-weighted image for each participant. T1- weighted images were normalized to the standard T1 template, and the resulting transformation matrices were applied to the co-registered functional images. Normalized images were smoothed with an 8-mm full-width-at-half maximum Gaussian filter. All fMRI data were individually examined for motion artifacts, and subjects with excessive interscan motion, defined by a translation >2 mm in one of the directions x, y, or z and a rotation >2° around one of the axes, were excluded.

2.4. Input data for deep learning and SVM models

Using the Statistical Parametric Mapping software (SPM8, http://www.fil.ion.ucl.ac.uk/spm8) and General Linear Model (GLM) method, 3D activation map was constructed based on the contrast images (activation vs. control) (Fig. 1). Representative images of activation map are given in the supplement. The whole-brain contrast images were used as input data for DL applications. For the SVM applications, we conducted experiments in three ways: a) directly fed the whole-brain betas to SVM without any preprocessing; b) used 40 features extracted by applying principal component analysis (PCA) to the whole-brain voxels; and c) used the β-values of five AAL regions chosen based on the significant differences of contrast between patients and controls (uncorrected $p < 0.01$, cluster size >50) (Pereira et al., 2009; Verma and Ghosh, 2017).

2.5. Deep learning model selection

The architectures of proposed DL model are summarized in Fig. 2. This model was selected by comparing classification performances of 3D-CAE based CNN with or without regularization techniques to those of 3D CNN with or without regularization techniques.

2.6. Unsupervised learning using 3D-CAE

Each of the proposed modules comprises the following components: convolution, batch normalization, activation, and pooling (or up-sampling). Specifically, 3D convolution was applied to input data employing a padding function to preserve the original scale. The use of a CAE generally consists of encoding and decoding steps: input data are transformed into a lower-dimensional feature space during the encoding phase; then, the encoded data are reconstructed to the original space in the output layer during the decoding phase. For a 3D tensor $\mathcal{X}$ (of size $h \times w \times d$), the encoder is defined as follows:

$$\mathcal{H}_k = \varphi(\sigma(\boldsymbol{BN}(\mathcal{W}_k * \mathcal{X} + b_k))) \quad (1)$$

where $\mathcal{W}_k$ and b_k are the 3D filter and feature-specific bias for the k-th feature map, respectively, * denotes 3D convolution filter, and $\sigma(\cdot)$ is the element-wise non-linear activation function. ***BN*** denotes the batch normalization (BN). Following the encoding phase, the decoder can be defined as follows:

$$\mathcal{X}\prime = \varphi\prime\left(\sigma\left(\boldsymbol{BN}\left(\sum_{k\in\mathbb{F}} \mathcal{W}\prime_k * \mathcal{H}_k + b\prime_k\right)\right)\right) \quad (2)$$

where $\mathcal{X}\prime$ denotes the 3D reconstructed map; $\mathcal{W}\prime_k$ and b'_k are the 3D filter and bias for the decoding process, respectively; $\mathbb{F}$ represents the group of latent 3D feature maps; φ and φ' are max-pooling and up-sampling, respectively. The cost function used to minimize the error between the input $\mathcal{X}$ and reconstructed input $\mathcal{X}\prime$ is the mean-squared error function:

$$E(\theta) = \frac{1}{N}\sum_{k=1}^{N} \|\mathcal{X} - \mathcal{X}'\|_2^2, \quad (3)$$

where N is the number of training samples and $\theta = \{\mathcal{W}, b, \mathcal{W}\prime, b\prime\}$ denotes the set of trainable parameters. Finally, the error given by Eq. (3) was minimized using the Adam optimizer (Mesnil et al., 2012), and the weight parameters for the layers were updated by backpropagation.

The data distribution was regularized by batch normalization (BN) and rectified linear unit (ReLU). The BN, a robust regularization technique, is known to be effective for overcoming overfitting and gradient vanishing problems. The d-dimensional input $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ was regularized as follows:

$$\widehat{x}_i = \left(\frac{x_i - \mu_b}{\sigma_b}\right)\gamma + \beta, \quad (4)$$

where x and $\widehat{x}$ are the input and output of BN, respectively; μ_b and σ_b are the mean and variance of the mini-batch, respectively; and γ and β are trained by backpropagation. Also, the ReLU has been widely used to overcome the gradient vanishing problem in DL.

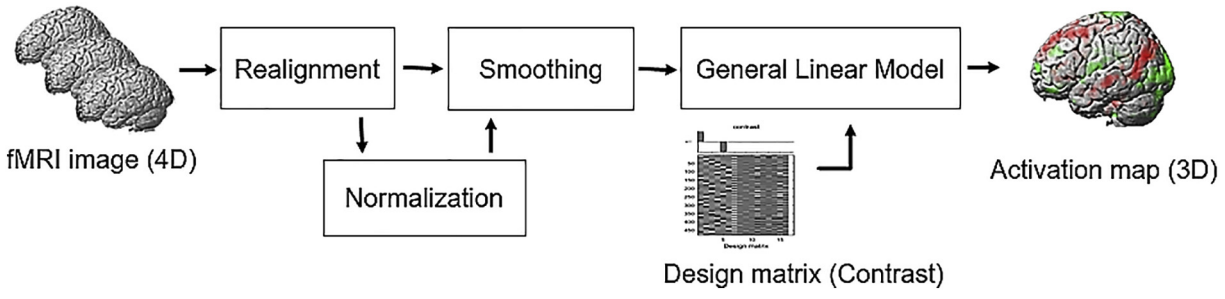

Fig. 1. Framework of generating 3D activation map.

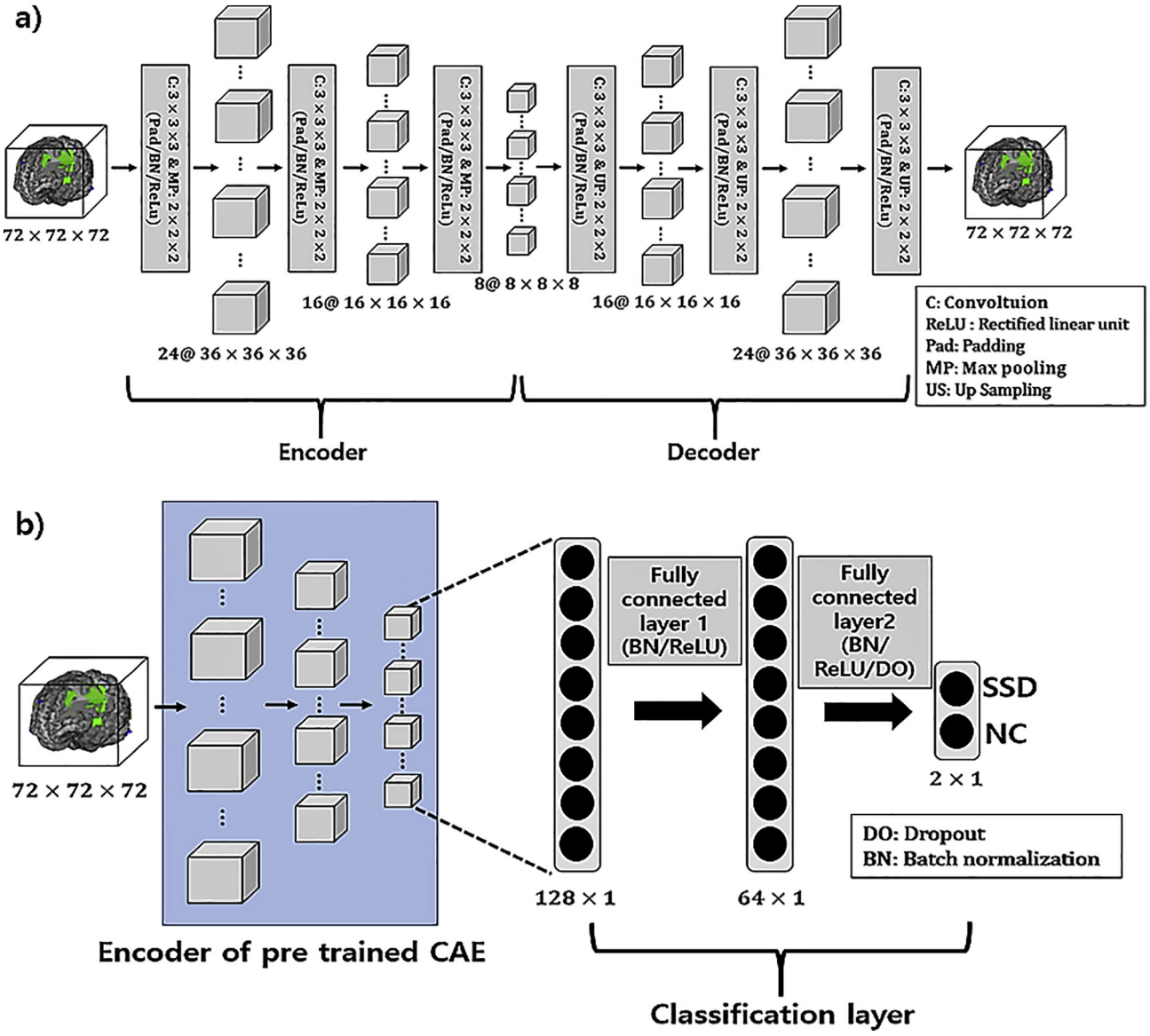

Fig. 2. Architecture of the proposed model: a) convolutional autoencoder (CAE) and b) 3D-CAE-based convolutional neural network (CNN).

2.7. Supervised learning using 3D CNNs

To overcome the problem of high-dimensional feature space and low-sample size data, we employed a fine-tuning technique (Ioffe and Szegedy, 2015; Raina et al., 2006) using pre-defined initial weights drawn from related domain knowledge rather than training from scratch. Because the initial data distribution is estimated from related sources, fine-tuning-based learning is effective for preventing overfitting problems due to limited data (Raina et al., 2006). In the proposed network, the first three convolutional layers (feature extraction layer) of the 3D CNN were the encoder of the learned 3D-CAE. The network involved the following two main phases: a) relevant features were extracted in the feature extraction layer, and b) the extracted features were classified into SSDs and control groups in the classification layer. Specifically, for the initial 3D fMRI data (size $72 \times 72 \times 72$), three feature extraction layers, including $3 \times 3 \times 3$ convolutions, BN, ReLU, and $2 \times 2 \times 2$ max-pooling, were implemented, and the number of feature maps was gradually decreased from 24 to 8 in the CAE process. The last feature maps were then transformed into 1D (4096×1) and fully connected to the classification layer neurons, which include two 128×1 and 64×1 hidden layers and a 2×1 output layer. As in the feature extraction layer, BN and ReLU were also employed in the fully connected (FC) layers. To encourage the sparsity of our network, drop out (DO) with a dropping rate of 0.8 was applied before the softmax. The key idea of DO is to randomly deactivate units during the training, and it has been known to extract sparse representations and have data augmentation effect (Srivastava et al., 2014). Additionally, we also applied Lasso regression (L1) regularization, which can decrease the coefficient of the less important features to zero, which leads to adaptive feature selection (Kim et al., 2016). Given the target labels, the complete set of trainable weights for the feature extraction and classification layers was updated during the training phase by backpropagation using the Adam optimizer (Mesnil et al., 2012). Its algorithms were mini-batch size of 30, $\beta\_1 = 0.9$, $\beta\_2 = 0.99$, 150 epochs, L1 regularization parameter of 1×10^{-3}, weight decay of 1×10^{-6}, and learning rate of 1×10^{-3}. The model without 3D-CAE has been trained from the scratch using random initialization of weight values.

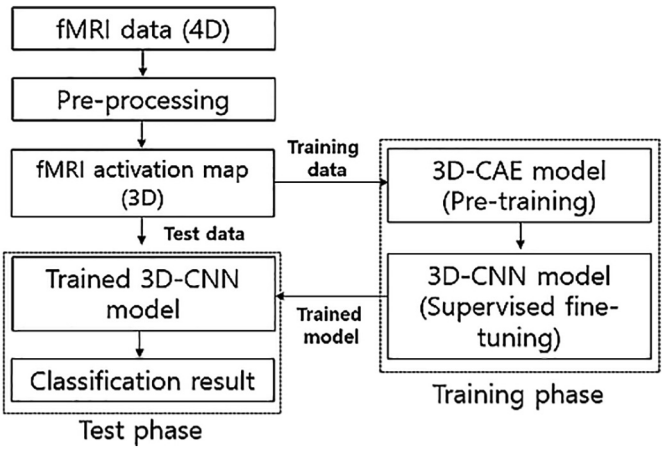

Fig. 3. Framework of the classification procedure.

2.8. Classifiers evaluation

The framework of the classification procedure using 3D-CAE based CNN is illustrated in Fig. 3. The 3D activation maps from 62 subjects with SSDs were used for pretraining with a 3D-CAE. The encoding layers of the 3D-CAE were used as 3D-CNN feature extraction layers for supervised fine-tuning. The classification evaluation was based on the data from the remaining 41 patients with SSDs and the 41 controls. To obtain a relatively unbiased estimate of generalizability, 10-fold cross-validation was performed. To avoid possible bias caused by the random dataset partitioning, the validation process was repeated 100 times independently. The proposed 3D-CAE-based CNN was trained and tested using the Keras (https://keras.io) open-source library for DL in Python. This scheme was applied the same in the model selection. Other 3D network models were selected from previous structural MRI studies using CNN in differentiating patients with Alzheimer's disease from controls (Hosseini-Asl et al., 2018; Korolev et al., 2017; Rieke et al., 2018). Though patient's diagnosis and MRI data were different from ours, these models were compared because no paper was found with regard to the classification of schizophrenia vs. controls using 3D CNN. We plotted the Receiver operating characteristic (ROC) curves of our model and other comparing models. Also, the proposed model and SVM models were evaluated using the nested 10-fold cross-validation 10 times. For the SVM, Bayesian optimization with 500 iterations on two hyperparameters (Kernel size and Box constraint) was applied using the fitcsvm function provided by the Matlab. In the proposed model, two hyperparameters reflecting DO ratio and weighting factor of L1 regularizer were optimized.

2.9. Visualization approaches

Two different visualizations are important in medical image analysis, i.e., visualizing the feature maps in each layer of the learned neural network and visualizing for explanation on why the neural network reached the decisions (Rieke et al., 2018). The first visualization is straightforward since the 3D feature maps are available and display of interesting slices of 3D feature maps are enough (Fig. 6). For the second visualization, we employed the class saliency visualization (CSV) method (Simonyan et al., 2014) to interpret the network outcomes. The CSV method calculates how much each input voxel $\mathcal{X}_{ijk}$ contributes to the final activation of the target class c. In detail, saliency can be computed for a given MRI data $\mathcal{X}$, a target class c, and a classification model f. In our case, the class c corresponds to SCZ. It provides us which voxels in the MRI contribute most to the model's classification into SCZ class. The class saliency map M, which has the same scale with the original input I, are calculated by the partial derivatives:

$$M_{ijk} = \left|\omega_{h(i,j,k)}\right| = \left|\frac{\partial f_c}{\partial I_{ijk}}\right| \quad (5)$$

Because each voxel in MRI is connected by at least one weight, the saliency map can be expressed as $\omega_{h(i,j,k)}$, where $h(i,j,k)$ denotes the index of the ω that spatially corresponds to MRI voxel I_{ijk}. For this reason, the 3D saliency map can be obtained by rearranging the vector ω, which is calculated by back-propagation. Because it directly produces saliency voxels by interpreting the pre-trained CNN, this model requires no additional annotation, such as a segmentation mask or bounding box, to localize important regions of neuroimaging data. This approach is useful in neuroimaging domains in which the differentiating characteristics of the various disease classes may not be well-established as it helps to identify meaningful neuroimaging findings.

Additionally, we illustrate how well each layer discriminates patients with SSDs from healthy controls. It is accomplished by transforming a high-dimensional feature space into 2D space by t-Distributed Stochastic Neighbor Embedding (t-SNE). The t-SNE (Maaten and Hinton, 2008) consists of the following phases: first, a probability distribution is constructed by computing the pairwise distance between the high-dimensional points; second, t-SNE computes the similarity matrix, which is the joint probability distribution, and the projected points are iteratively updated by minimizing the Kullback–Leibler divergence between a Gaussian distribution in the high-dimensional space and a t distribution in the low-dimensional space.

3. Results

3.1. Model selection

As the number of epochs increased, the performance of our model initially improved and later converged. In particular, an increase in performance was observed between 100 and 500 epochs (Table 2). The use of different numbers of neurons in the classification layers (FC 1 and FC 2, respectively) did not have a major influence on the final results. Overall, 128–64 neurons provided maximum stability (Table 3). Applying different L1 and DO regularization parameters produced slightly different performances. The best performance was observed when λ=0.001 and DO rate = 0.8. Based on these observations, we fixed our networks to

Table 2
Performance comparison for increasing numbers of epoch in the CAE.

Number of epochs	Accuracy	Sensitivity	Specificity	PPV	NPV
100	82.99%	86.21%	78.87%	80.65%	85.61%
500	83.51%	86.83%	80.41%	80.54%	86.73%
1000	84.43%	88.42%	80.06%	81.14%	88.10%
1500	84.01%	87.59%	80.67%	80.86%	87.45%

CAE, convolutional auto-encoder; NPV, negative predictive values; PPV, positive predictive values.

Table 3
Performance comparison for different numbers of neurons in the classification layer.

Number of neurons	FC1: 32 FC2: 16	FC1: 64 FC2: 32	FC1:128 FC2:64	FC1:256 FC2:128
#prams	202 K	390 K	770 K	1542 K
Accuracy	83.80%	83.94%	84.43%	84.17%

FC, fully connected; #prams, number of parameters.

Table 4
Performance comparison of various 3D CNN models.

Network model	Accuracy	Sensitivity	Specificity	PPV	NPV
3D-CNN (training from scratch)	60.56%	60.55%	60.57%	60.78%	60.34%
3D-CNN (training from scratch with BN)	68.92%	70.94%	66.98%	67.18%	70.77%
3D-CNN (training from scratch with DO)	66.40%	64.34%	68.44%	66.92%	65.90%
3D-CNN (training from scratch with BN + DO)	67.94%	76.15%	61.28%	61.48%	75.99%
3D-CAE + CNN	76.16%	77.40%	74.92%	75.43%	76.91%
3D-CAE + CNN with BN	81.22%	83.36%	79.17%	79.32%	83.23%
3D-CAE + CNN with DO	78.15%	81.55%	75.00%	75.16%	81.41%
3D-CAE + CNN with BN & DO	84.43%	88.42%	80.06%	81.14%	88.10%

BN, batch normalization; CAE, convolutional auto-encoder; CNN, convolutional neural network; DO, dropout; NPV, negative predictive values; PPV, positive predictive values.

Table 5
Classification accuracy between our model vs. other 3D CNN models.

Network model	Accuracy	Sensitivity	Specificity	PPV	NPV
Korolev et al., 2017	74.85%	77.61%	72.35%	71.79%	78.08%
Rieke et al., 2018	68.30%	71.09%	65.59%	66.06%	70.74%
Hosseini-Asl et al., 2018	78.04%	81.34%	74.92%	75.43%	80.91%
Proposed model	84.43%	88.42%	80.06%	81.14%	88.10%

NPV, negative predictive values; PPV, positive predictive values.

have two fully connected layers with 128 and 64 nodes, respectively. The learning stops at 150 epoch. Comparing various models, 3D-CAE based CNN with BN and DO achieved best accuracy, 84.43% (Table 4) which was selected as our model.

3.2. Classification

Other 3D CNN models reported performances ranging from 68.30% to 78.04%. Our model using 10-fold cross-validation achieved the best performance, 84.43% (Table 5). The ROC curves showed the best area under the ROC curve (AUC), 0.8989 for our model (Fig. 4). The result of permutation test (5000 times) for our model was significant (p value <0.001) suggesting the accuracy of deep-learning was high. In the SVM, when raw whole brain betas and PCA extracted features were used as input, classification accuracies were 57.32% and 70.73%. With beta-values extracted from the five areas, it was 67.07%. Our model using the nested cross-validation showed 84.15% accuracy (Table 6 and Fig. 5).

Table 6
Classification accuracy between our model vs. SVM models using the nested cross validation.

Network model	Accuracy	Sensitivity	Specificity	PPV	NPV
SVM-raw	57.32%	68.29%	46.34%	56.00%	59.30%
SVM-beta	67.07%	73.17%	60.98%	65.22%	69.44%
SVM-PCA	70.73%	78.05%	63.41%	68.09%	74.28%
Proposed model	84.15%	87.80%	80.49%	81.82%	86.84%

NPV, negative predictive values; PPV, positive predictive values.

3.3. Visualization

Visualizations of feature maps for the three convolutional layers of our model showed some intense regions and segmentation clues in the first two layers. Global feature maps that look like binary patterns were extracted in the third layer (Fig. 6). Note that this phenomenon is also observed in the CNNs processing natural images (Zeiler and Fergus, 2014). Fig. 7 shows the results of applying t-SNE to output of each layer (feature maps). The t-SNE used a 50-dimensional feature vector reduced by principal component analysis (PCA) and transformed the vector into a point in 2-dimensional space. For the FC layers, PCA was not applied since size of the original vectors are small, i.e., 128 for FC 1 and 64 for FC 2. As moving forward from convolutional layer 1 (Conv 1) to FC 2 layer, two classes are more clearly separated. The visualization results by CSV method revealed a relatively large single area corresponding to the left inferior and middle temporal lobes as key regions differentiating patients with SSDs from healthy controls (Fig. 8).

4. Discussion

The use of DL to investigate psychiatric and neurological disorders has the practical advantage of not requiring manual feature selection, and an increasing number of neuroimaging studies are using DL techniques to elucidate the neural correlates of these disorders. For the first time, we applied a CNN model to distinguish patients with SSDs from healthy controls using task-based fMRI data. Our proposed model performed better than traditional SVM and other 3D CNN models.

4.1. Methodological aspects

The strength of our study was its novel use of a CNN model to distinguish patients with SSDs from healthy controls. CNN is a particular type of feed-forward neural network inspired by how the

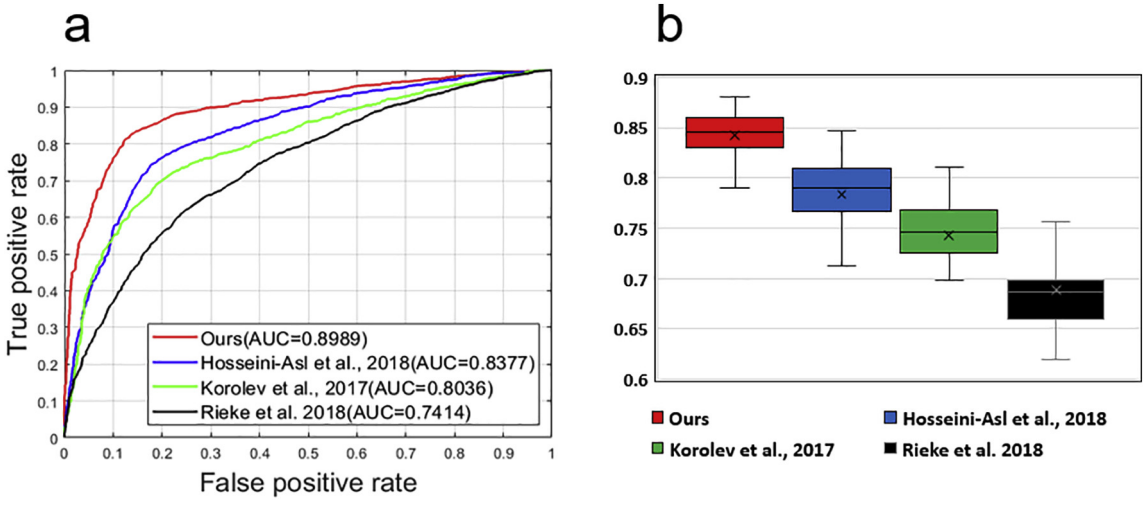

Fig. 4. Quantitative performance comparison of our model and other 3D CNN models: a) ROC curves and AUC scores, and b) boxplot graph.

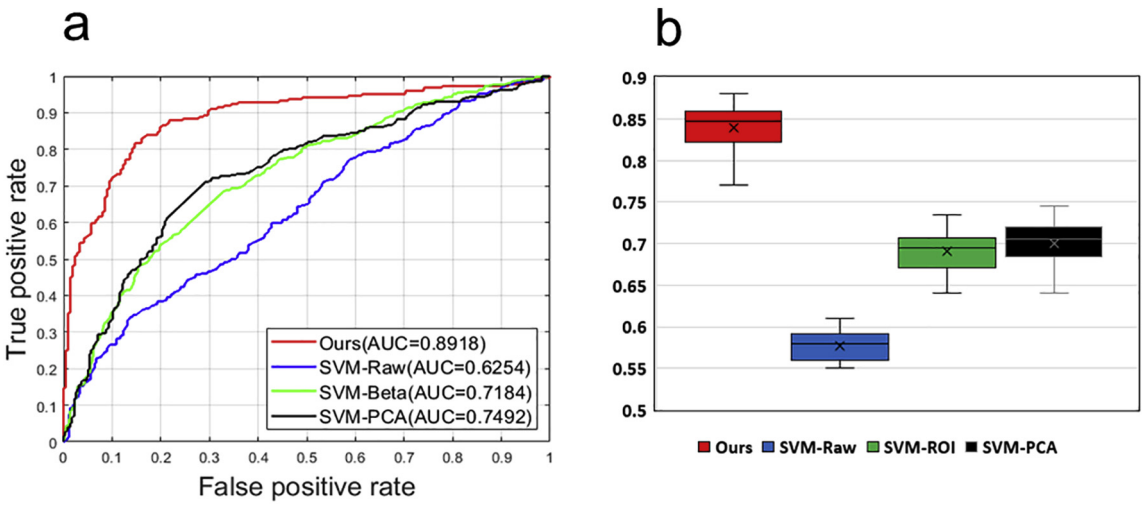

Fig. 5. Quantitative performance comparison of our model and SVM models: a) ROC curves and AUC scores, and b) boxplot graph.

human visual cortex processes information. Over the past decade, CNN has been breaking records in computer vision in several competitions (Krizhevsky et al., 2012) and has rapidly become a methodology of choice for analyzing medical images. Application areas are image classification, object detection, segmentation, registration, and other tasks; for a concise overview, see the article by Litjens et al. (2017). Few studies in the field of neuropsychiatry attempted to differentiate patients with Alzheimer's disease from subjects with mild cognitive impairment or normal controls using sMRI data in a way of preserving 3D spatial locality. Our proposed network model, incorporating 3D-CAE and 3D-CNN, was designed to encourage end-to-end learning without using hand-crafted features. As a result, spatial locality of 3D fMRI was preserved during the training and testing phases, resulting in more robust feature representation and good performance. We believe that this is a more useful way to take full advantage of CNNs.

Another characteristic of our model was the application of robust regularization techniques, such as CAE, BN, DO, L1 regularizer, and ReLU, which are all known to be effective for analyzing fMRI data with large variations in a limited dataset. The network's reinforcement of sparsity used CAE, DO (with a high dropping ratio), and L1 boosted performance. In particular, due to the use of BN, our model does not require the layer-wise approach used in the existing stacked CAE (CAE_s) model (Jonathan et al., 2011), which involves a complex and cumbersome training process, independent training of all CAE convolution layers. The advantage of using regularization techniques seems evident because other models not using those techniques performed worse. As models (from first to fourth rows in Table 4) were trained from scratch, no pre-training process was considered, leading to an overfitting problem and poor performance. For efficient hyperparameter optimization, stopping at 150th epoch, 128 and 64 neurons for FC 1 and FC 2, respectively, and L1-regulazation parameters set at λ=0.001 ensured superior classification performance. Due to exponential combinations and excessive time for learning model (Bergstra and Bengio, 2012), we took an approach usually adopted by DL community, starting from default values and changing important hyperparameters, and finding out relevant values.

4.2. Clinical implications

The classification accuracy of our model was much higher than those (57.32–70.73%) of conventional SVM. It was somewhat unexpected to see low performances with SVM models. Possible explanation may be that significance level for contrast images was

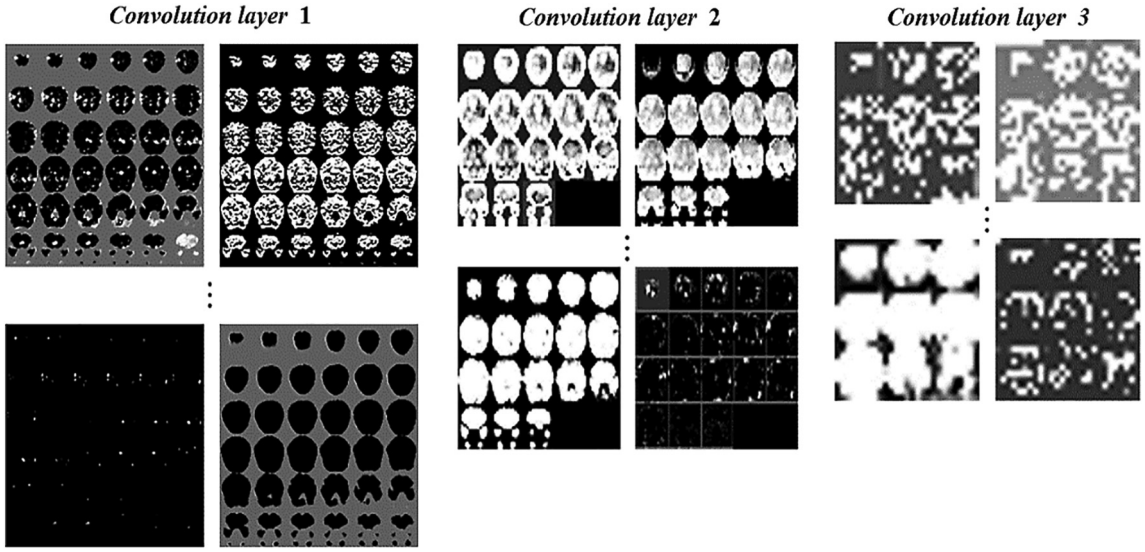

Fig. 6. Visualization results of convolutional layer feature maps. From left to right: first, second and third convolution layers.

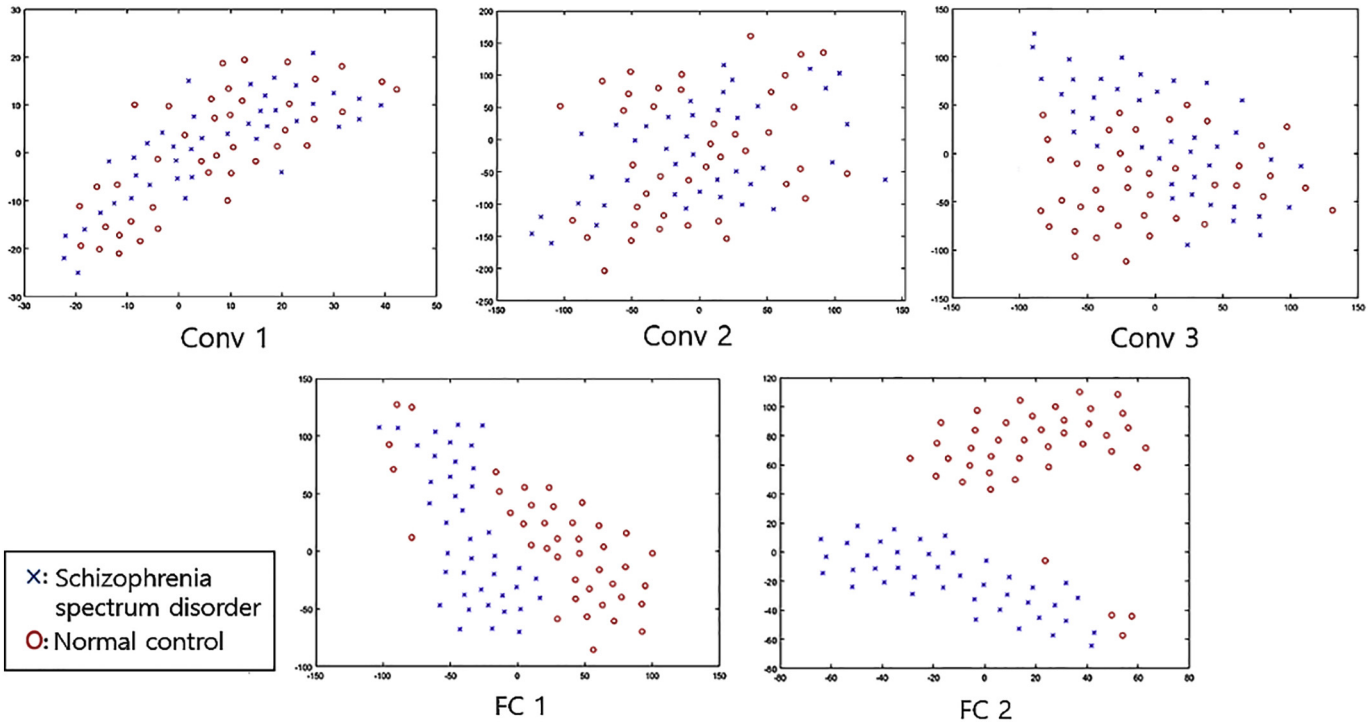

Fig. 7. Visualization of hidden layers (three convolution layers, Conv 1, 2 and 3 and two fully connected layers, FC 1, and FC 2) using t-SNE.

very low (uncorrected $p < 0.01$) in the two sample *t*-test. In anyway, the modest result (84.43%) with 3D-CAE based CNN gives us a new hope that automated individual-level diagnosis of SCZ may be possible in the near future. When compared to the performances of other 3D CNN models, ours was also relatively higher. However, direct comparison should be cautious because patient's diagnosis and architecture of the CNNs were different. Other studies using DBN or deep neural network (DNN) with sMRI data of SCZ reported similar results to ours (Pinaya et al., 2016; Plis et al., 2014). However, it should be noted that they used 1D input feature vector. As the structural brain abnormalities become more evident in chronic SCZ, the potential of using task-based fMRI and rsfMRI data for the early diagnosis of SCZ may be much greater. From a clinical perspective, an accuracy rate of 84.43% obtained by the proposed model is not sufficient for use in real-world situations. In clinical practice, we often need to differentiate individuals with SCZ from not only healthy controls but also from those with other psychiatric disorders (i.e., multiclass classification). Given that the accuracy of our model would decrease in such complex situations, further improvement of our model may be required.

In terms of the visualization results of the convolution and fully connected layers, as the hidden layer depth increased, our ability to use the features of the model to discriminate between the two groups gradually improved. These features may have contributed to the enhanced classification performance. To verify the features that contributed the most to classification accuracy, we developed a saliency map using the CSV method. According to the map, the left inferior and middle temporal lobes were critical areas for

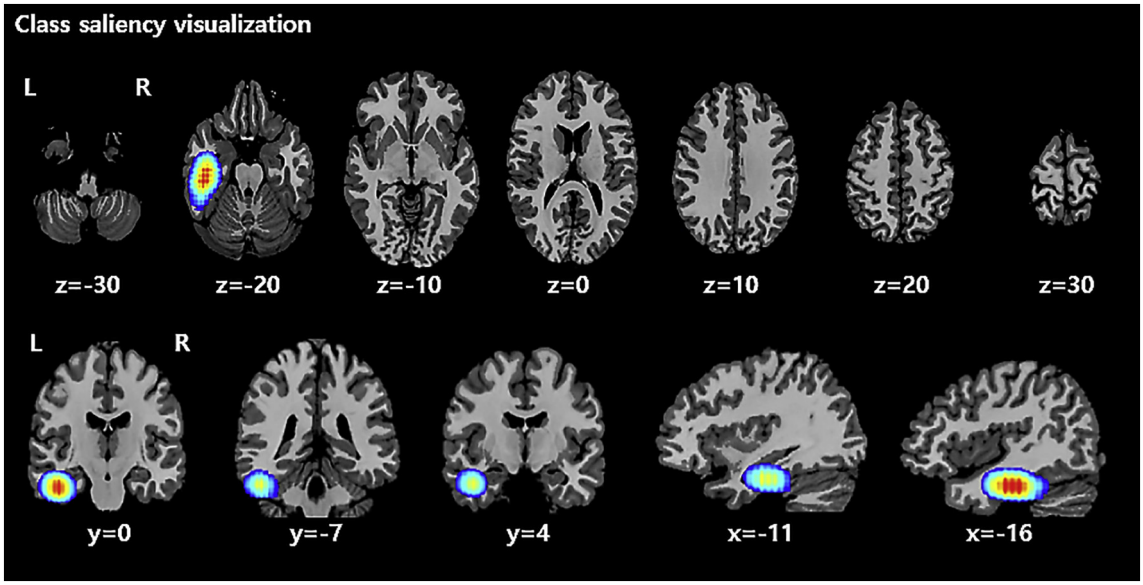

Fig. 8. Discriminative brain region in classification of schizophrenia spectrum disorders vs. normal controls using class saliency visualization method.

differentiating patients with SSDs from normal controls. Group comparisons of task-based fMRI data revealed significantly higher activation in several brain areas (left fusiform gyrus, left ligual gyrus, left cuneus, left cerebellum 6, etc.) of patients with SSDs compared to normal controls (for details, see Supplementary Table 1). Given that the left fusiform gyrus is included in the left inferior temporal lobe, it seems that the results of the DL model did not significantly differ from the findings of the univariate analysis. There is strong evidence indicating abnormalities in the temporal lobe, especially the superior temporal gyrus and the amygdala–hippocampal complex, in SCZ (Shenton et al., 2001). The inferior and middle temporal gyri, by contrast, have received far less attention. However, several lines of evidence point to the possibility of less gray matter volume in the middle or inferior temporal gyrus in patients with first-episode (Kuroki et al., 2006) and chronic (Onitsuka et al., 2004) SCZ, and to less activation in the left posterior inferior temporal gyrus and bilateral superior/middle temporal gyri in individuals with SCZ performing the silent lip-reading task (Surguladze et al., 2001). Additionally, accumulating evidence suggests white matter abnormalities in the temporal lobe in SCZ (Eastwood and Harrison, 2003; Mitelman et al., 2003; Whitford et al., 2007). Previous functional neuroimaging studies have suggested that the inferior temporal gyrus and middle temporal gyrus are involved in several cognitive processes (Cabeza and Nyberg, 2000), including visual perception (inferior temporal gyrus) (Herath et al., 2001; Ishai et al., 1999), language and semantic memory processing (middle temporal gyrus) (Chao et al., 1999; Tranel et al., 1997), and multimodal sensory integration (Mesulam, 1998). It is of interest that, compared to static images of visual speech, moving visual images (lip-reading) induced activation in the superior/middle temporal cortex, bilateral lingual gyrus, bilateral parietal lobule, and bilateral inferior frontal gyrus (Calvert et al., 1997; Calvert and Campbell, 2003). Specifically, the inferior temporal gyrus was activated when individuals observed facial gestures (Bernstein et al., 2011), talking faces (Ludman et al., 2000; Campbell et al., 2001), or symbolic gestures (Xu et al., 2009) and while individuals processed the multi-modal semantic signals associated with the meaning of speech (Mummery et al., 1999; Giraud and Truy, 2002). Considering the function of the inferior and middle temporal lobes, it may be inferred that the IOR-evoking task in the present study may be associated with processing the semantic signals of the visual speech images; thus, our DL model identified the inferior and middle temporal lobes as critical areas for classification.

This study has several limitations. First, as the number of subjects used for the training and test phases was small, it is not clear how well these findings will generalize to different samples. Given the difficulty of obtaining large fMRI datasets of individuals with SCZ at one site, the development of consortia that can facilitate the collection of large datasets is highly required. Indeed, the Asia Consortium on MRI Studies in Psychosis (ACMP) was organized in 2017. We hope it will provide the resources necessary for larger-scale research. Second, because of limited sample size, we could not conduct a nested cross-validation which is reported to provide an almost unbiased estimate of the true error (Varma and Simon, 2006). Third, in the process of CSV, we had to approximate a coarse scale saliency map on the template brain image because an exact localization of salient regions was limited. This limitation should be overcome in subsequent research. Despite these shortcomings, this is the first study to apply a CNN model to distinguish individuals with SCZ from normal controls and to obtain a salient map. In conclusion, our findings suggest that 3D-CAE-based CNN can accurately (84.43%) differentiate patients with SSDs from normal controls. Furthermore, it enabled identification of the inferior and middle temporal lobes as salient areas.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.schres.2019.07.034.

Contributors

Young-Chul Chung designed the study. Kang Han Oh designed the model. All authors acquired the data. Young-Chul Chung, Kang Han Oh and Woo Sung Kim analyzed. Young-Chul Chung, Il-Seok Oh, Kang Han Oh and Woo Sung Kim wrote the article, which all authors reviewed and approved for publication.

Acknowledgements

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) and the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C2383), research funds of Chonbuk National University in 2018 and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (grant number: 2018R1A6A3A01013251).

References

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305.
Bernstein, L.E., Jiang, J., Pantazis, D., Lu, Z.L., Joshi, A., 2011. Visual phonetic processing localized using speech and nonspeech face gestures in video and point-light displays. Hum. Brain Mapp. 32 (10), 1660–1676.
Cabeza, R., Nyberg, L., 2000. Imaging cognition, II: an empirical review of 275 PET and fMRI studies. J. Cogn. Neurosci. 12 (1), 1–47.
Calvert, G.A., Campbell, R., 2003. Reading speech from still and moving faces: the neural substrates of visible speech. J. Cogn. Neurosci. 15 (1), 57–70.
Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Ivesen, S.D., David, A.S., 1997. Activation of auditory cortex during silent lipreading. Science 276 (5312), 593–596.
Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., Brammer, M.J., David, A.S., 2001. Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). Brain Res. Cogn. Brain Res. 12 (2), 233–243.
Chao, L.L., Haxby, J.V., Martin, A., 1999. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. Nat. Neurosci. 2 (10), 913–919.
Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 411–418.
Eastwood, S.L., Harrison, P.J., 2003. Interstitial white matter neurons express less reelin and are abnormally distributed in schizophrenia: towards an integration of molecular and morphologic aspects of the neurodevelopmental hypothesis. Mol. Psychiatry 8, 821–831.
First, M.B., Spitzer, R.L., Gibbon, M., Williams, J.B.W., 1996. Structured Clinical Interview for DSM-IV Axis I Disorders (Patient Ed. (SCID-P)) (Version 2). Biometrics Research, New York State Psychiatric Institute, New York, NY.
Foxet, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E., 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc. Natl. Acad. Sci. U. S. A. 102, 9673–9678.
Giraud, A.L., Truy, E., 2002. The contribution of visual areas to speech comprehension: a PET study in cochlear implants patients and normal-hearing subjects. Neuropsychologia 40, 1562–1569.
Han, O.S., Hong, J.P., 2000. Structured Clinical Interview for DSM-IV Axis I Disorders, Research Version. Korea Hana Medical Publishing, Seoul.
Han, S., Huang, W., Zhang, Y., Zhao, J., Chen, H., 2017. Recognition of early-onset schizophrenia using deep-learning method. Appl. Informatics 4 (1), 1–11.
Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H., 2017. Brain tumor segmentation with deep neural networks. Med. Image Anal. 35, 18–31.
Herath, P., Kinomura, S., Roland, P.E., 2001. Visual recognition: evidence for two distinctive mechanisms from a PET study. Hum. Brain Mapp. 12, 110–119.
Hosseini-Asl, E., Ghazal, M., Mahmoud, A., Aslantas, A., Shalaby, A.M., Casanova, M.F., Barnes, G.N., Gimel'farb, G., Keynton, R., El-Baz, A., 2018. Alzheimer's disease diagnostics by a 3D deeply supervised adaptable convolutional network. Front. Biosci. (Landmark Ed.) 23, 584–596.
Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. ICML 448–456 (2015).
Ishai, A., Ungerleider, L.G., Martin, A., Haxby, J.V., 1999. Distributed representation of objects in the human ventral visual pathway. Proc. Natl. Acad. Sci. U. S. A. 96 (16), 9379–9384.
Jonathan, M., Ueli, M., Dan, C., Jurgen, S., 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. ICANN 21, 52–59.
Kam, T.E., Zhang, H., Shen, D., 2018. A novel deep learning framework on brain functional networks for early MCI diagnosis. MICCAI 2018, 293–301.

Kay, S.R., Fiszbein, A., Opler, L.A., 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr. Bull. 13, 261–276.
Kim, J., Calhoun, V.D., Shim, E., Lee, J.H., 2016. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. NeuroImage 124, 127–146.
Korolev, S., Safiullin, A., Belyaev, M., Dodonova, Y., 2017. Residual and Plain Convolutional Neural Networks for 3D Brain MRI Classification. ISBI. https://doi.org/10.1109/ISBI.2017.7950647.
Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems, 25, pp. 1097–1105.
Kuroki, N., Shenton, M.E., Salisbury, D.F., Hirayasu, Y., Onitsuka, T., Ersner-Hershfield, H., Yurgelun-Todd, D., Kikinis, R., Jolesz, F.A., Mccarley, R.W., 2006. Middle and inferior temporal gyrus gray matter volume abnormalities in first-episode schizophrenia: an MRI study. Am. J. Psychiatry 163 (12), 2103–2110.
Lang, P.J., Bradley, M.M., Cuthbert, B.N., 1998. International Affective Picture System (IAPS): Photographic Slides. Center for Research in Psychophysiology University of Florida, Gainsville.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Litjens, G., Sanchez, C.I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-van de kaa, C., Bullt, P., Ginneken, B., Laak, J., 2016. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci. Rep. 6, 26286.
Litjens, G., Kooi, T., Bejnordi, E.B., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Laak, J.A.W.M., Ginneken, B., Sanchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.
Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., Feng, D., 2014. Early diagnosis of Alzheimer's disease with deep learning. In: IEEE 11th International.
Ludman, C.N., Summerfield, A.Q., Hall, D., Elliott, M., Foster, J., Hykin, J.L., Bowtell, R., Morris, P.G., 2000. Lip-reading ability and patterns of cortical activation studied using fMRI. Br. J. Audiol. 34 (4), 225–230.
Maaten, L., Hinton, G., 2008. Visualizing Data Using t-SNE. JMLR, pp. 2579–2605.
Mesnil, G., Dauphin, Y., Glorot, X., Rifai, S., Bengio, Y., Goodfellow, I.J., Lavoie, E., Muller, X., Desjardins, G., Warde-Farley, D., 2012. Unsupervised and Transfer Learning Challenge: A Deep Learning Approach. ICML, pp. 97–110.
Mesulam, M.M., 1998. From sensation to cognition. Brain 121, 1013–1052.
Mitelman, S.A., Shihabuddin, L., Brickman, A.M., Hazlett, E.A., Buchsbaum, M.S., 2003. MRI assessment of gray and white matter distribution in Brodmann's areas of the cortex in patients with schizophrenia with good and poor outcomes. Am. J. Psychiatry 160 (12), 2154–2168.
Mummery, C.J., Patterson, K., Wise, R.J., Vandenberghe, R., Price, C.J., Hodges, J.R., 1999. Disrupted temporal lobe connections in semantic dementia. Brain 122 (Pt 1), 61–73.
Onitsuka, T., Shenton, M.E., Salisbury, D.F., Dickey, C.C., Kasai, k., Toner, S.K., Frumin, M., Kikinis, R., Jolesz, F.A., McCarley, R.W., 2004. Middle and inferior temporal gyrus gray matter volume abnormalities in chronic schizophrenia: an MRI study. Am. J. Psychiatry 161 (9), 1603–1611.
Patel, P., Aggarwal, P., Gupta, A., 2016. Classification of Schizophrenia Versus Normal Subjects Using Deep Learning. ICVGIP (December 18–22).
Pereira, F., Mitchell, T., Botvinck, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45, 199–209.
Pinaya, W.H.L., Gadelha, A., Doyle, O.M., Noto, C., Zugman, A., Cordeiro, Q., Jackowski, A.P., Bressan, R.A., Sato, J.R., 2016. Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. Nat. Sci. Rep. 6 (38897).
Plis, S.M., Hjelm, D.R., Salakhutdinov, R., Allen, E.A., Bockholt, H.J., Long, J.D., Johnson, H.J., Paulsen, J.S., Turner, J., Calhoun, V.D., 2014. Deep learning for neuroimaging: a validation study. Front. Neurosci. 8, 1–11.
Raina, R., Ng, A.Y., Koller, D., 2006. Constructing Informative Priors Using Transfer Learning. ICML, pp. 713–720.
Rieke, J., Fabian, E., Weygandt, M., Haynes, J.D., Ritter, K., 2018. Visualizing convolutional networks for MRI-based diagnosis of Alzheimer's disease. MICCAI 24–31.
Shenton, M.E., Dickey, C.C., Frumin, M., McCarley, R.W., 2001. A review of MRI findings in schizophrenia. Schizophr. Res. 49, 1–52.
Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps (arXiv:1312.6034v2 [cs.CV] 19 Apr 2014).
Srivastava, N., Hinton, Geoffrey, Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.
Suk, H.I., Shen, D., 2013. Deep learning-based feature representation for AD/MCI classification. MICCAI 2013, 583–590.
Surguladze, S.A., Calvert, G.A., Brammer, M.J., Campbell, R., Bullmore, E.T., Giampietro, V., David, A.S., 2001. Audio–visual speech perception in schizophrenia: an fMRI study. Psychiatry Res. 106 (1), 1–14.
Tranel, D., Damasio, H., Damasio, A.R., 1997. A neural basis for the retrieval of conceptual knowledge. Neuropsychologia 35 (10), 1329–1339.
Varma, S., Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinf. 7, 91.
Verma, N.K., Ghosh, A.K., 2017. Computational Intelligence: Theories, Applications and Future Directions-Volume II. Springer, Singapore.
Whitford, T.J., Grieve, S.M., Farrow, T.S.D., Gomes, L., Brennan, J., Harris, A.W.F., Gordon, E., Williams, L.M., 2007. Volumetric white matter abnormalities in first-episode schizophrenia: a longitudinal, tensor-based morphometry study. Am. J. Psychiatry 164, 1082–1089.
Xu, J., Gannon, P.J., Emmorey, K., Smith, J.F., Braun, A.R., 2009. Symbolic gestures and spoken language are processed by a common neural system. Proc. Natl. Acad. Sci. U. S. A. 106 (49), 20664–20669.
Zeiler, M.D. and Fergus, R., 2014. Visualizing and Understanding Convolutional Networks. ECCV 2014.
Zeng, L.L., Wang, H., Hu, P., Yang, B., Pu, W., Shen, H., Chen, X., Liu, Z., Yin, H., Tan, Q., Wang, K., Hu, D., 2018. Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. EBioMedicine 30, 74–85.