NeuroImage 223 (2020) 117328

Contents lists available at ScienceDirect

NeuroImage

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

fMRI volume classification using a 3D convolutional neural network robust to shifted and scaled neuronal activations

Hanh Vu, Hyun-Chul Kim, Minyoung Jung, Jong-Hwan Lee*

Department of Brain and Cognitive Engineering, Korea University, Anam-ro 145, Seongbuk-gu, Seoul 02841, Republic of Korea

ARTICLE INFO

Keywords:
Classification
Convolutional neural networks
Deep neural networks
Functional MRI
Human Connectome Project
Machine learning
Real-time fMRI
Sensorimotor tasks

ABSTRACT

Deep-learning methods based on deep neural networks (DNNs) have recently been successfully utilized in the analysis of neuroimaging data. A convolutional neural network (CNN) is a type of DNN that employs a convolution kernel that covers a local area of the input sample and moves across the sample to provide a feature map for the subsequent layers. In our study, we hypothesized that a 3D-CNN model with down-sampling operations such as pooling and/or stride would have the ability to extract robust feature maps from the shifted and scaled neuronal activations in a single functional MRI (fMRI) volume for the classification of task information associated with that volume. Thus, the 3D-CNN model would be able to ameliorate the potential misalignment of neuronal activations and over-/under-activation in local brain regions caused by imperfections in spatial alignment algorithms, confounded by variability in blood-oxygenation-level-dependent (BOLD) responses across sessions and/or subjects. To this end, the fMRI volumes acquired from four sensorimotor tasks (left-hand clenching, right-hand clenching, auditory attention, and visual stimulation) were used as input for our 3D-CNN model to classify task information using a single fMRI volume. The classification performance of the 3D-CNN was systematically evaluated using fMRI volumes obtained from various minimal preprocessing scenarios applied to raw fMRI volumes that excluded spatial normalization to a template and those obtained from full preprocessing that included spatial normalization. Alternative classifier models such as the 1D fully connected DNN (1D-fcDNN) and support vector machine (SVM) were also used for comparison. The classification performance was also assessed for several *k*-fold cross-validation (CV) schemes, including leave-one-subject-out CV (LOOCV). Overall, the classification results of the 3D-CNN model were superior to that of the 1D-fcDNN and SVM models. When using the fully-processed fMRI volumes with LOOCV, the mean error rates (± the standard error of the mean) for the 3D-CNN, 1D-fcDNN, and SVM models were 2.1% (± 0.9), 3.1% (± 1.2), and 4.1% (± 1.5), respectively ($p = 0.041$ from a one-way ANOVA). The error rates for 3-fold CV were higher (2.4% ± 1.0, 4.2% ± 1.3, and 10.1% ± 2.0; $p < 0.0003$ from a one-way ANOVA). The mean error rates also increased considerably using the raw fMRI 3D volume data without preprocessing (26.2% for the 3D-CNN, 75.0% for the 1D-fcDNN, and 75.0% for the SVM). Furthermore, the ability of the pre-trained 3D-CNN model to handle shifted and scaled neuronal activations was demonstrated in an online scenario for five-class classification (i.e., four sensorimotor tasks and the resting state) using the real-time fMRI of three participants. The resulting classification accuracy was 78.5% (± 1.4), 26.7% (± 5.9), and 21.5% (± 3.1) for the 3D-CNN, 1D-fcDNN, and SVM models, respectively. The superior performance of the 3D-CNN compared to the 1D-fcDNN was verified by analyzing the resulting feature maps and convolution filters that handled the shifted and scaled neuronal activations and by utilizing an independent public dataset from the Human Connectome Project.

1. Introduction

For more than a decade, deep learning based on a variety of artificial neural networks, including fully connected feedforward neural networks, convolutional neural networks (CNNs), and recurrent neural networks, has become widespread in a number of research fields, including vision, speech, and natural language processing (Goodfellow et al., 2016; LeCun et al., 2015). For example, CNNs have been crucial to object recognition within a 2D image since the pioneering model AlexNet (Krizhevsky et al., 2012) and have been the cornerstone of computer-vision research, such as facial recognition (Parkhi et al., 2015; Schroff et al., 2015) and video captioning (Karpathy et al., 2014; Simonyan and Zisserman, 2014). This is because CNNs can automatically extract topographically organized image features from an input image in a biologically plausible manner that approximates the human visual cortex (Güçlü and van Gerven, 2015; Horikawa and Kamitani, 2017; Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014).

* Corresponding author.
E-mail address: jonghwan_lee@korea.ac.kr (J.-H. Lee).

https://doi.org/10.1016/j.neuroimage.2020.117328
Received 31 October 2019; Received in revised form 16 July 2020; Accepted 20 August 2020
Available online 5 September 2020

Deep-learning techniques have also proven to be efficacious in neuroimaging data analysis, particularly for structural magnetic resonance imaging (sMRI) (Hazlett et al., 2017; Kleesiek et al., 2016; Plis et al., 2014; Shen et al., 2017). For instance, Plis et al. (2014) applied a deep belief network (DBN), a fully connected greedy layer-wise pre-trained restricted Boltzmann machines, to sMRI data as a feature extractor. The classification accuracy for patients with schizophrenia (SZ) or Huntington disease from healthy controls was superior when using the features of the DBN with an increased number of hidden layers (Plis et al., 2014).

Deep learning has also been beneficial for functional MRI (fMRI) in the investigation of human brain function (Güçlü and van Gerven, 2015; Horikawa and Kamitani, 2017; Jain and Huth, 2018; Kell et al., 2018; Khaligh-Razavi and Kriegeskorte, 2014; Wen et al., 2018; Yamins et al., 2014) and the classification of fMRI data (Jang et al., 2017; Kim et al., 2016; Plis et al., 2014; Suk et al., 2014; Zhao et al., 2017b). There have been several studies that have investigated the classification of 3D whole-brain fMRI volumes such as using fully connected deep neural networks (fcDNNs) and support vector machines (SVMs) by vectorizing a single fMRI volume into a 1D vector (Fig. 1(a)) (Jang et al., 2017; Ramasangu and Sinha, 2014; Song et al., 2011; Zhang et al., 2018). For example, Jang et al. (2017) classified fMRI volumes acquired from sensorimotor tasks using a 1D-fcDNN which was pretrained using a DBN. As a result, the weights of the 1D-fcDNN exhibited spatial patterns that were remarkably sensorimotor-specific, particularly for the higher hidden layer connected to the output layer (Jang et al., 2017).

It is important to note, however, that anatomical regions and functional networks may not be perfectly aligned across sessions and/or subjects (Fig. 1(b)) even after spatial normalization due to the variability in the spatial layout of neuronal activations for each of the functional networks across the whole brain, confounded further by imperfections in spatial normalization algorithms (Aguirre et al., 1998; Calhoun et al., 2017; Dohmatob et al., 2018). Thus, it can be argued that the classification of fMRI volumes using vectorized 1D patterns is inherently limited due to both the shift in neuronal activations across sessions and/or subjects from spatial misalignment and the scale in neuronal activations that are affected by various spatial extents from the over- or underestimation of neuronal activations (Eklund et al., 2016; Lee et al., 2009a). A more suitable approach for fMRI volume classification would be a classifier model that can handle 3D volume information, such as the 3D-CNN model (Dou et al., 2016; Kleesiek et al., 2016; Maturana and Scherer, 2015), thus avoiding these potential issues.

In this context, CNN models have only recently been employed as machine-learning models for fMRI data (Huang et al., 2018; Nie et al., 2016; Zhao et al., 2017a; Zhao et al., 2017b). For example, Huang et al. (2018) employed a 1D-CNN-based autoencoder for fMRI time-series reconstruction and demonstrated the ability of the 1D-CNN to extract hierarchical features from a task-based fMRI time series. In addition, another 3D-CNN model, trained using multimodal neuroimaging data including fMRI data, has been utilized to extract features and to predict the lifetime of patients with brain tumors using an SVM classifier applied to the output of the fully connected layer (Nie et al., 2016). In a more recent study, a 3D-CNN was adopted to learn the latent representation for decoding task states using a larger cohort (Wang et al., 2018), with a 4D fMRI time series from Human Connectome Project (HCP) data used as input to the 3D-CNN model rather than a single fMRI volume. The multiple fMRI 3D volumes across the time points in each block were averaged and subsequently used as input for the first convolutional layer to classify the corresponding task information of each block. In another study, Zhao et al. (2017b) used a 3D-CNN model to automatically assign a label to 3D functional network input, which was decomposed from a fully preprocessed fMRI volume series using sparse dictionary learning, to one of the resting-states networks.

Despite this recent research on the use of 3D-CNNs, no previous study has employed a 3D-CNN for the classification of individual raw whole-brain blood-oxygenation-level-dependent (BOLD) fMRI volumes or has presented an in-depth interpretation of the 3D convolution kernel and extracted 3D feature representation maps in addition to reporting classification performance. The average fMRI volume obtained from one task block and/or 3D volume of functional networks decomposed from a 3D fMRI volume series would have less noise and less spatial variability than a single fMRI volume. Thus, we believe that classifying a single fMRI volume by employing a 3D-CNN model is more challenging than the problems investigated in previous studies. Our motivation for using a 3D-CNN as a classifier for fMRI volumes is based on the characteristics of 2D-CNNs, which have been shown to be able to extract scale- and shift-invariant features from an object recognition task using 2D images (Kanazawa et al., 2014; Norouzi et al., 2009; Pinto et al., 2008; Sharif Razavian et al., 2014). Inspired by this, we hypothesized that a 3D-CNN with convolution operations followed by down-sampling such as stride and/or pooling would produce robust feature maps of shifted and scaled 3D neuronal activation patterns in local brain regions while preserving the overall spatial layout of the neuronal activations across the whole brain. This would lead to superior classification performance by resolving potential spatial misalignment issues compared to conventional classifier models using 1D vectorized multivoxel patterns across the whole brain. If the 3D-CNN demonstrates superior classification performance, it would support investigations into how brain networks shift during task performance in real-time using a simultaneously acquired fMRI dataset. Thus, we demonstrated the feasibility of the 3D-CNN model in this regard (i.e., please refer to "***2.10. Online classification of the sensorimotor tasks via real-time fMRI***" for more details).

2. Materials and methods

2.1. Overview

Fig. 2 illustrates the overall process used to systematically compare the classification performance using classifier models including the 3D-CNN, 1D-fcDNN, and SVM models for various fMRI volume input scenarios, including fully preprocessed data with standard preprocessing steps (Fig. 2(a); red arrow), different combinatorial steps for minimal preprocessing without spatial normalization (Fig. 2(b); green arrow), and online classification (Fig. 2(c); blue arrow). Details regarding the adopted fMRI volumes acquired from the four sensorimotor tasks can be found in the next subsection. The 3D-CNN model is explained in Section 2.4 and the 1D-fcDNN and SVM models are described in Section 2.5. To investigate the efficacy of our 3D-CNN model in comparison to the 1D-fcDNN and SVM models for a public dataset, the minimally preprocessed 3T data from the S1200 release of the HCP (Van Essen et al., 2012) were used as described in Section 2.11.

2.2. Participants and fMRI data acquisition

The overall study protocol was approved by the Institutional Review Board (IRB) at Korea University. The participants submitted written consent forms and were compensated as outlined in the IRB documents. Twelve young right-handed male volunteers (age = 25.0 ± 2.0 years) who did not report any neurologic or neuropsychiatric problems participated in the research by performing four sensorimotor tasks: left-hand clenching (LH), right-hand clenching (RH), auditory attention (AD), and visual stimulation (VS) (Jang et al., 2017). The fMRI run for each task consisted of three task blocks (20 s per block), each followed by a resting state (RS) period of 20 s consisting of a white cross for fixation on a black background, with a 30-s RS period at the beginning of the run (Fig. 3(a)). The participants clenched their left or right hand approximately twice per second for the LH and RH runs, respectively, listened to 1.1-kHz sound waveforms during the task block and 1-kHz sound waveforms during the RS period for the AD run, and watched an alternating black-and-white checkerboard with a frequency of 8 Hz during the task period for the VS run. BOLD fMRI data for each run were acquired using a 3T Tim Trio MRI scanner (Siemens, Erlangen,

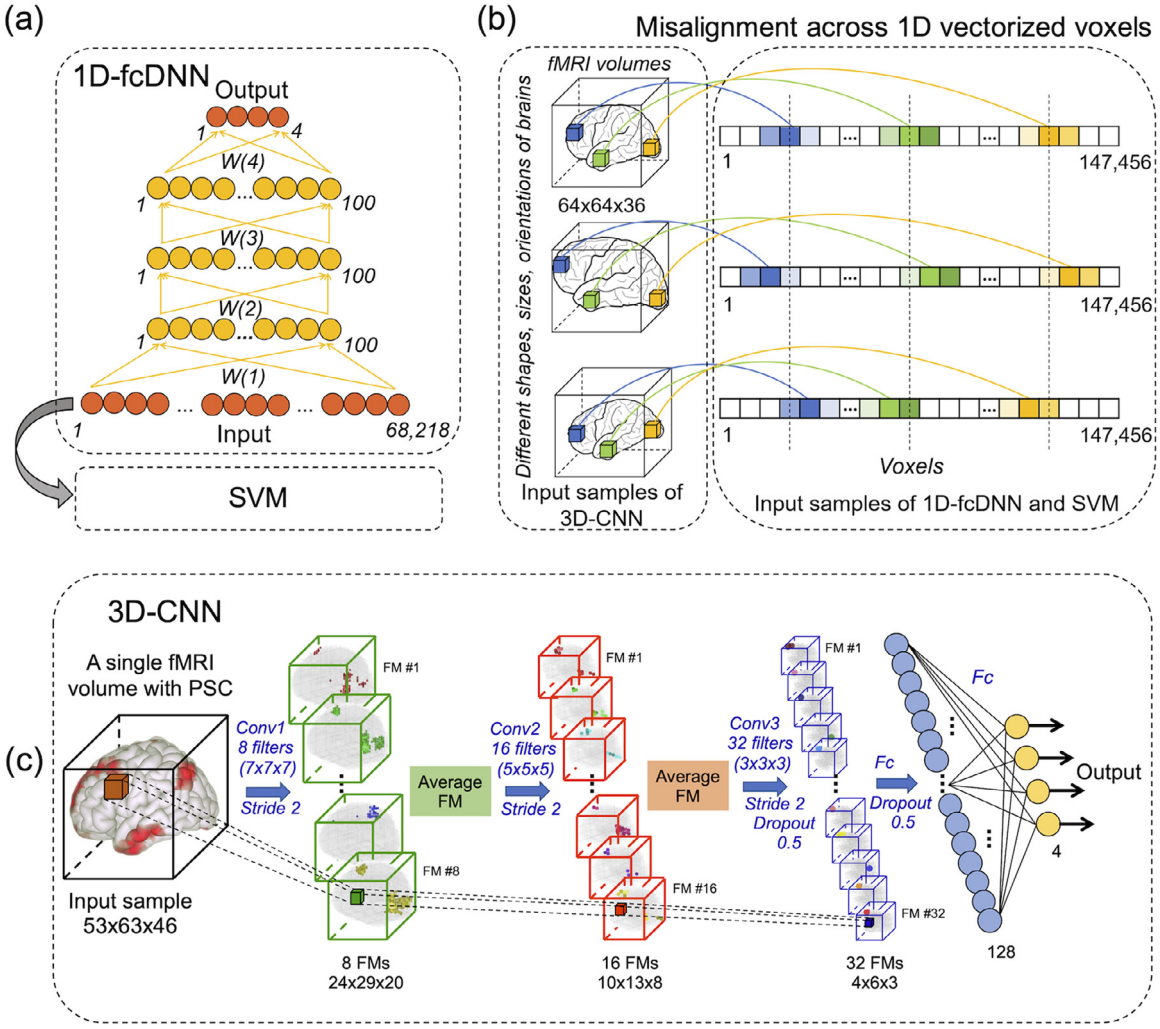

Fig. 1. (a) 1D-fcDNN, SVM, and linear classifier models for fMRI volume classification. Each volume was vectorized across in-brain voxels and 1D-vectorized voxels were used as input. (b) An illustration of potential misalignment due to imperfections in the spatial normalization algorithms confounded by individual variability in BOLD responses across runs, sessions, and/or subjects. In the raw and minimally preprocessed fMRI 3D volume series without spatial normalization, misalignment is substantial because of the varying sizes, shapes, and orientations within the field-of-view of the brain across subjects. (c) A schematic of the 3D-CNN for fMRI volume classification including the dimensions of the input volumes and feature maps and the size of the convolutional filters in each Conv layer. 1D-fcDNN, 1D fully connected deep neural network; W, weight; SVM, support vector machine; 3D-CNN, 3D convolutional neural network; PSC, percentage signal change; Conv1, the first convolutional layer; Conv2, the second convolutional layer; Conv3, the third convolutional layer; FM, feature map of a convolutional layer; Fc, fully connected layer.

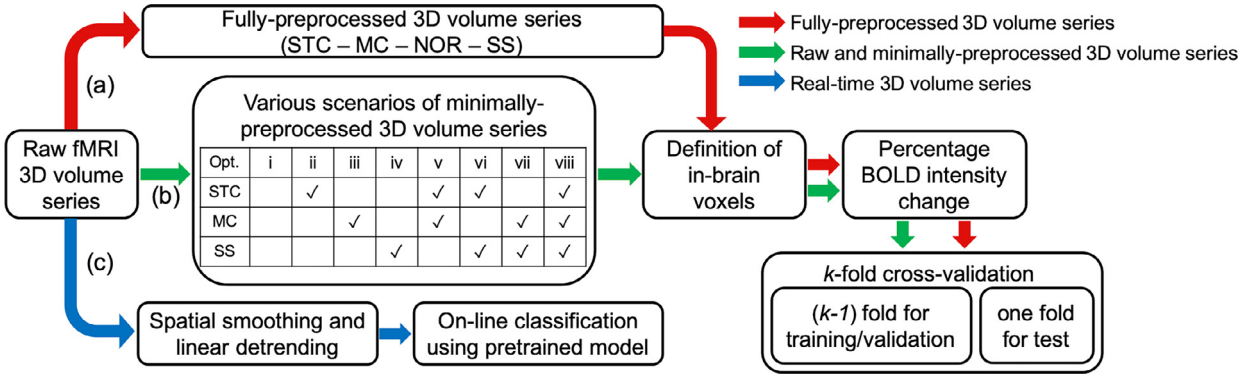

Fig. 2. Flow diagram of the fMRI volume classification process, including full preprocessing, minimal preprocessing, and real-time fMRI scenarios. STC, slice timing correction; MC, motion correction; NOR, spatial normalization; SS, spatial smoothing; BOLD, blood-oxygenation-level-dependent.

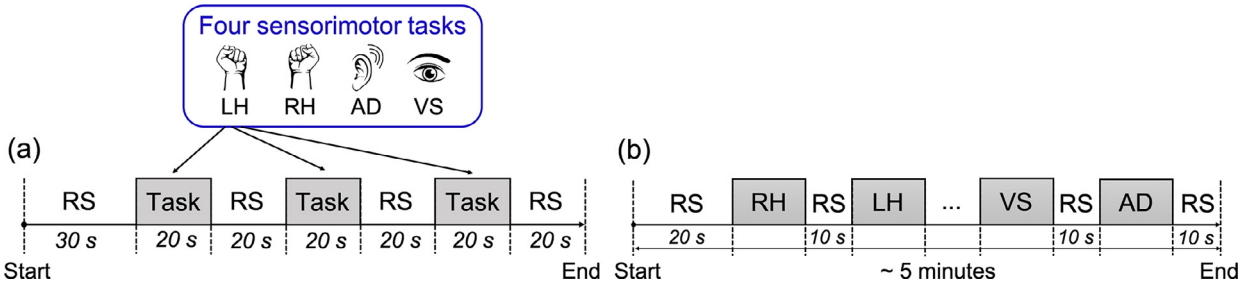

Fig. 3. Experimental paradigms for fMRI data acquisition: (a) off-line classification based on fMRI runs (150 s) for each of the four sensorimotor tasks, consisting of three task blocks interleaved with a resting state (RS) period; and (b) online classification using an fMRI run (~ 5 min) consisting of counter-balanced task blocks (two blocks for each of the four tasks; 8–11 repetition times [TRs] per block; TR = 2 s) interleaved with a RS period (10 s). LH, left-hand clenching; RH, right-hand clenching; AD, auditory attention; VS, visual stimulation.

Germany) with a 12-channel head coil and a standard gradient-echo echo-planar-imaging (EPI) pulse sequence (TR/TE = 2000/30 ms; field-of view = 240 × 240 mm^2; slice thickness = 4 mm; 36 axial slices without a gap; voxel size = 3.75 × 3.75 × 4 mm^3).

2.3. Preprocessing the fMRI data

The first five volumes (10 s) in each fMRI run were discarded to allow the T1 effect to equilibrate. The remaining 70 volumes (140 s) were subject to standard preprocessing using SPM8, with slice timing correction, motion correction, spatial normalization to the Montreal Neurological Institute (MNI) space with a 3-mm isotropic voxel size, and spatial smoothing with an 8-mm full-width at half-maximum (FWHM) Gaussian kernel (Fig. 2(a)), which were applied in that order. Following this, 30 preprocessed task-related fMRI volumes (60 s across the three task blocks delayed 6 s from task onset to allow the hemodynamic response to peak) were available for each task and for each subject (i.e., 120 task-related fMRI volumes across the four tasks for each subject; 1440 fMRI volumes across all 12 subjects). In addition to this fully-preprocessed data, various preprocessing strategies using subsets of the preprocessing steps (excluding spatial normalization) were applied to the data in order to assess the ability of the classifier model to handle potential misalignment and variability in neuronal activations between fMRI volumes across sessions and/or subjects (Fig. 2(b); refer to Section 2.8 for details). The intensities of the 30 task-related fMRI volumes for each task run and for each subject were normalized to percentage signal change (PSC, %) using the average BOLD intensity across the RS volumes (80 s) as a baseline. Classification performance using fMRI volumes subject only to spatial smoothing followed by linear detrending was also analyzed for the three classifier models in online classification via real-time fMRI (Fig. 2(c); refer to Section 2.10 for details).

2.4. 3D-CNN model for fMRI volume classification

Our 3D-CNN model for 3D fMRI volume classification (Fig. 1(c)) was developed by modifying LeNET5, a 2D-CNN model with three convolutional (Conv) layers and two fully connected (Fc) layers for the classification of MNIST digit images (LeCun et al., 1998). Our 3D-CNN consisted of three 3D Conv layers and two Fc layers: 8 Conv kernels/filters (i.e., channels) in the first Conv layer (Conv1) with a kernel size of 7 × 7 × 7 and a stride of 2 (i.e., the dimensions of each of the 8 feature maps are 24 × 29 × 20 from a 53 × 63 × 46 fully processed fMRI volume); 16 Conv filters in Conv2 with a kernel size of 5 × 5 × 5 and a stride of two (i.e., 16 10 × 13 × 8 feature maps); and 32 Conv filters in Conv3 with a kernel size of 3 × 3 × 3 and a stride of 2 (i.e., 32 4 × 6 × 3 feature maps). The 8 feature maps of Conv1 and the 16 feature maps of Conv2 for a single fMRI volume in the input layer were averaged across the filters/channels and the average feature maps were used as the input for the subsequent Conv layers. All 2,304 nodes from the 32 feature maps in Conv3 were then connected to one Fc hidden layer with 128 nodes, and these 128 hidden nodes were connected to the four output nodes for four-class classification. A rectified linear unit (ReLU) and sigmoid activation functions were used for all Conv/Fc and output layers, respectively.

The cost function was the cross-entropy between the target one-hot vector (i.e., a value of 1 for the output node assigned as a target class for the input fMRI volume and 0 for the remaining output nodes) and the obtained output across the four output nodes. Stochastic gradient descent with a mini-batch size of 50 (without momentum) was used to minimize the cost function with an initial learning rate of 10^{-3}, which was annealed to 10^{-6} throughout the training (i.e., 300 epochs). Dropouts with a probability of 0.5 were adopted in the Conv3 layer and Fc layer to alleviate potential overfitting (Srivastava et al., 2014). The 3D-CNN implemented in the MATLAB environment (github.com/hagaygarty/mdCNN) was modified to fit our 3D-CNN model for fMRI volume classification. Our code and sample data are publicly available (https://github.com/bsplku/3dcnn4fmri). Both the code and data comply with the requirements of our funding bodies, institution, and institutional ethics committee.

2.5. Alternative classifier models

2.5.1. 1D fully connected DNN (1D-fcDNN)

A 1D-fcDNN model with three hidden layers (100 hidden nodes per layer) and an output layer (with four nodes) was compared with the 3D-CNN (Fig. 1). This 1D-fcDNN model was pretrained using a DBN to initialize the weights across the three hidden layers and was fine-tuned using the four target class labels by adding an output layer for classification (Jang et al., 2017). Each of the 3D fMRI volumes was vectorized into a 1D vector across in-brain voxels and used as input. The activation functions were the ReLU for the hidden node and the linear function (followed by the softmax layer) for the output node. The parameters used to train the DBN were as follows: a maximum epoch of 300; learning rates of 10^{-5}, 10^{-2}, and 10^{-2} for the first, second, and third layers of the DBN, respectively; a momentum factor of 0.5; and a mini-batch size of 10 (Jang et al., 2017). The 1D-fcDNN was then fine-tuned from the pretrained weights of the DBN across 300 epochs using the following parameters: a cost function representing the mean-squared error (MSE) between the target output and obtained output values; an initial learning rate of 10^{-3} with annealing after 100 epochs to a minimum of 10^{-6} at 300 epochs; a momentum factor of 0.1; a mini-batch size of 10; a small step of 10^{-2} to change the L1-norm parameter (μ); and a small positive interval of 10^{-3} to define the non-zero ratio for weight sparsity optimization (ε) (Jang et al., 2017). Based on the best performing 1D-fcDNN from our adopted four-task sensorimotor dataset (i.e., error rate = 3.4%) (Jang et al., 2017), the target weight sparsity level was set to stringent, with a non-zero ratio of 0.001 only between the input layer and the first hidden layer. Since cross-entropy-based cost functions have been widely used and may perform better than MSE, the 1D-fcDNN was also trained using a cross-entropy cost function with the same parameters as those

used to train the 1D-fcDNN model with the MSE cost function. The code is publicly available at https://github.com/bsplku/dnnwsp.

2.5.2. Support vector machine (SVM) classifier

A linear SVM and a non-linear SVM with a radial basis function (RBF) kernel were also applied to the 1D-vectorized fMRI volume within in-brain voxels for the classification of the corresponding task information. The hyperparameters, including the soft margin parameter C and the RBF kernel size γ_{SVM}, were optimized based on a grid search (i.e., $C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$ and $\gamma_{SVM} = 2^{-15}, 2^{-13}, \ldots, 2^{3}$) using the training data (*k-2* folds) and the validation data (one fold) in the inner loop of the nested *k*-fold CV scheme (Cristianini and Shawe-Taylor, 2000; Lee et al., 2009a; Varoquaux et al., 2017). Using the optimized hyperparameters, the SVM models were re-trained using the training and validation data (*k-1* folds), and test performance was assessed using the test data (one remaining fold) in the outer loop of the nested *k*-fold CV. LIBSVM models implemented in MATLAB were employed to perform classification (www.csie.ntu.edu.tw/~cjlin/libsvm).

2.6. Training and testing of the classifier models via k-fold cross-validation

The 1,440 fMRI volumes were divided into training and test data by splitting the 12 subjects into subjects for the training and testing of the classifier models in the *k*-fold CV framework. Various *k*-fold scenarios (i.e., $k = 3, 4, 6,$ or 12) were adopted to evaluate performance depending on the proportion of omitted test data (Varoquaux et al., 2017). Using *k*-fold CV, training data (*k-1* folds) and test data (the one remaining fold) were presented in each loop, and all possible scenarios for splitting the data were analyzed. For example, in the 6-fold CV, the data from 10 subjects were used for training and the data from the two remaining subjects were used for testing the trained model. This framework was then repeated 66 times for the one remaining fold across all combinations of two subjects in our dataset (i.e., C(12, 2)). Similarly, a total of 220 folds for the 4-fold CV (i.e., C(12, 3)) and 495 folds for the 3-fold CV (i.e., C(12, 4)) were tested. The mean error rate and standard error (SE) of the mean across all folds are reported for each of the classifier models. Using the error rates obtained from randomly permuted *k*-fold sets for each subject, the McNemar test (McNemar, 1947), a non-parametric statistical test that has recently been applied to compare the performance of pairs of machine-learning classifiers (Dietterich, 1998; Raschka, 2018), was used to determine the statistical significance of any difference between classifier pairs. Using the average error rate across randomly permuted *k*-fold sets for each subject, the McNemar test and one-way analysis of variance (ANOVA) were used to assess the statistical significance across all subjects for pairs of classifiers and for all classifiers, respectively. To test for significance, a null distribution for the chi-square values from the McNemar test was obtained using a total of 10,000 sets of randomly permuted indices for 120 volumes across the four tasks for each subject; the resulting null distribution was used to correct the *p*-value (Manly, 2018). To correct the statistical significance from the ANOVA, 10,000 random permutations were conducted using randomized indices of subjects (*, corrected $p < 0.05$; **, corrected $p < 0.01$; ***, corrected $p < 0.001$). The "randperm" function implemented in MATLAB (R2018a) was used for randomization. The computational time for training and testing was recorded for each of the three classifiers (3D-CNN, 1D-fcDNN, and SVM) with a single hardware system (Intel i7-6850K 3.6GHz, Nvidia GeForce GTX1080, 128 GB RAM) using either the central processing unit (CPU) or the graphics processing unit (GPU) via the "gpuArray" command in MATLAB.

2.7. Interpretation of the trained models

2.7.1. 3D-CNN

The trained 3D convolution filters and the filtered 3D feature maps were visualized for each of the three Conv layers. The feature maps extracted from the 30 fMRI volumes with PSC intensities for each task and for each subject in the training set were averaged for each of the convolution filters (i.e., $n = 8$, 16, and 32 for Conv1, Conv2, and Conv3, respectively). The average 3D feature map for each filter was z-scored with zero mean and unit variance across in-brain voxels, and the z-scored feature maps were subject to one-sample *t*-tests across all training subjects for group-level inference. Group-level 3D feature maps for the top one percentile of *t*-scores were obtained for each filter in the Conv1, Conv2, and Conv3 layers and for each task. Consequently, the 8, 16, and 32 3D group-level feature maps for Conv1, Conv2, and Conv3, respectively, were visualized using distinct colors in both multiple 2D-axial slice images and 3D-rendered volumes.

We postulated that the feature maps of Conv3 would be highly task-specific compared to those in Conv1 and Conv2. Thus, the coarse feature maps of Conv3 (i.e., $4 \times 6 \times 3$) were back-projected to MNI space (i.e., the input volume space, $53 \times 63 \times 46$; bottom of Fig. 5(c)) via effective receptive field (ERF) mapping (Luo et al., 2016). Please refer to the section "***Back projection of a feature map at Conv3 to the input layer using effective receptive field mapping***" in the Supplementary Materials for more details. Consequently, the back-projected ERF maps at the input layer for the feature maps at the Conv3 were visualized with an intensity threshold higher than the top one percentile from the average ERF map for each of the four tasks.

The 3D filters for each of the three Conv layers (i.e., 8 filters with dimensions of $7 \times 7 \times 7$ for Conv1; 16 filters with dimensions of $5 \times 5 \times 5$ for Conv2; and 32 filters with dimensions of $3 \times 3 \times 3$ for Conv3) were also visualized using an intensity threshold of higher than the top five percentile and lower than the bottom five percentile for the positive and negative weights, respectively, across all filters for each Conv layer. Because the kernels/filters at the last Conv layer (i.e., Conv3) extract the most highly task-relevant features, the convolution filters and the brain regions found to be dominant in these features were further investigated in association with each of the four tasks (please see the subsection, "***Interpretation of the filters at the last Conv layer in association with each of the four target tasks***" in the Supplementary Material and Fig. S1 for more details).

In addition to the visualization of the feature maps and filters, two more straightforward approaches – the visualization of class saliency maps (Simonyan et al., 2013) and class activation maps (CAM) (Zhou et al., 2016) – that have been proposed to interpret the deep features of CNN models in recognizing 2D images were applied to our trained 3D-CNN model. Please refer to the section "***Visualization of the 3D-CNN model using class saliency and class activation maps***" in the Supplementary Materials for more details. Using these two approaches, the 3D volume for the class saliency or class activation maps for each of the 30 fMRI volumes and for each of the four tasks (i.e., LH, RH, AD, and VS) per subject were obtained. The 30 3D volumes for the class saliency or class activation maps for each task and each subject were then averaged and subjected to one-sample *t*-tests across all 12 subjects for group inference. The resulting group-level *t*-scored maps for the class saliency or class activation maps for each of the four tasks were overlaid on top of the T1-weighted anatomical template (MNI152 with 3-mm isotropic voxel resolution).

2.7.2.1D-fcDNN

The weights of the 1D-fcDNN model, particularly between the input layer and each of the 100 nodes in the first hidden layer, appeared to parcellate the whole brain into multiple brain regions (Jang et al., 2017). Therefore, the weight feature maps that were strongly associated with each of the four tasks were visualized. More specifically, the brain regions in the top one percentile in the weight feature maps were first defined as active regions. The average PSC BOLD intensity maps across all 30 volumes for each task and for each subject were obtained, and the average PSC BOLD intensity maps across all training subjects were obtained for each task. The average PSC BOLD intensity map of the task with non-zero values only in the top one percentile was defined as active regions in the input fMRI volumes for each of the four tasks.

The task label for each weight feature map was then assigned based on the highest overlap ratio between (a) the active regions in the corresponding weight feature map and (b) the active regions in the input fMRI volumes for each of the four tasks.

2.8. Classifier input based on various fully and minimally preprocessed fMRI volume scenarios

The fMRI volumes with PSC intensities were used as input for the classifiers, in which the average BOLD signal for the four RS periods (80 s; 20 s/period) in each run was used as the baseline BOLD intensity to normalize the BOLD intensity during the task period as PSC intensity for fully and minimally preprocessed data (Fig. 2(a) and (b)). The mask used to include the brain area with non-zero PSC intensities (i.e., an in-brain mask) for each subject was defined from the voxels whose BOLD intensities were greater than 30% of the maximum BOLD intensity from the average fMRI volume across all volumes for the subject; the intersection of the in-brain masks across all subjects was defined as an overall in-brain mask. For the 3D-CNN model, the input fMRI volumes (i.e., $53 \times 63 \times 46$ for the fully preprocessed data; $64 \times 64 \times 36$ for the minimally preprocessed data) had non-zero PSC intensities within only the in-brain voxels of this overall in-brain mask. For the 1D-fcDNN and SVM models, the fMRI volumes with PSC intensity were vectorized across the in-brain voxels (i.e., $68{,}218 \times 1$ for the fully processed data; $17{,}222 \times 1$ for the minimally preprocessed data). Raw fMRI volumes without any preprocessing were also applied as input for the classifier to assess the ability of the model to handle variation in neuronal activations across volumes even from a single task for a single subject. Once the in-brain mask for each subject was defined, the raw fMRI volumes ($64 \times 64 \times 36$) with non-zero values only within the in-brain mask were used as input for the 3D-CNN model. Similarly, the vectorized voxel intensities across the whole brain ($147{,}456 \times 1$) were used as input for the 1D-fcDNN and SVM models.

It was possible that alternative linear and non-linear machine learning classifier models that are simpler than the 1D-fcDNN and SVM models may perform well for our sensorimotor dataset. Thus, the linear discriminant analysis (LDA) and logistic regression (LR) as alternative linear classifiers and random forest (RF) as an alternative non-linear classifier, were also employed to classify a single fMRI volume for comparison. The 1D vectorized in-brain voxels were used as an input sample and the classification was performed using LOOCV. The three classifier models implemented in the MATLAB (R2018a) environment were used with an optimized hyperparameter set selected from a grid search of hyperparameters including default parameters in the nested five-fold cross-validation: (a) for the LDA model using the "fitcdiscr.m", the linear coefficient threshold (δ) was searched among log-scaled positive values in the range of $[10^{-6}, 10^{3}]$ and the amount of regularization (γ) was searched among real values in the range of [0, 1]; (b) for the RF model using the "cartree.m" (https://www.mathworks.com/matlabcentral/fileexchange/31036-random-forest), the number of trees in the ensemble was searched from {10, 20, 30, 40, 50, 100, 150, 200, 250, …, 2000}, the maximum depth of each tree was searched from {5, 10, 15, 20, 25, …, 100, 'None'} with 'None' means the nodes can be expanded as much as possible, the minimum number of samples required to split an internal leaf node was searched from {2, 5, 10, 15, 20}, and the minimum number of samples required to be at a leaf node was searched from {1, 2, 5, 10, 15}; and (c) for the LR model using the "LogisticRegression.m" (https://www.mathworks.com/matlabcentral/fileexchange/42770-logistic-regression-with-regularization-used-to-classify-hand-written-digits), the regularization parameter (λ) was searched from $\{10^{-3}, 10^{-2}, \ldots, 10^{4}\}$.

The baseline machine learning classifiers (i.e., SVM, LDA, RF, and LR) may perform better using the input patterns obtained from optimally smoothed fMRI volumes because these classifiers do not have a mechanism to inherently smooth the activations of voxels, unlike the 3D-CNN model (via convolution kernels). Thus, several isotropic Gaussian smoothing kernel sizes (4 mm, 6 mm, 10 mm, and 15 mm) were additionally applied to the full preprocessing steps after spatial normalization. Using the smoothed fMRI volume data, the classification performance of the baseline machine learning algorithms was obtained using the 1D vectorized input patterns from the smoothed whole-brain voxel-level fMRI volume data.

The utility of each component of the 3D-CNN model (i.e., the Conv layer, stride, pooling layer, and Fc layer) was systematically investigated for various 3D-CNN model architectures, including a model with the Conv layers with or without stride, with or without average or max pooling layers, and with or without a Fc layer. In addition, network architecture without a Conv layer was also tested in order to investigate whether convolutional operations are essential. This series of investigations was conducted using both fully preprocessed and raw/minimally preprocessed fMRI volume data.

The 1D-fcDNN model was also investigated for the scenario in which it has a similar number of parameters as the 3D-CNN model (i.e., 301,220). To this end, the input fMRI volume (i.e., $53 \times 63 \times 46$) from the fully preprocessed fMRI volume data was down-sampled at a scale of 0.35 (i.e., $19 \times 23 \times 17$), which resulted in 3204 in-brain voxels, down from the original 68,218. The 1D activation pattern (3204×1) from the whole brain was then used as input for the 1D-fcDNN with three hidden layers and 100 hidden nodes per layer, meaning that the number of parameters for this 1D-fcDNN model was 321,240. The classification was conducted under the LOOCV framework, and the resulting performance was compared to that of the 3D-CNN model.

2.9. Classification using parcellated ROIs across the whole brain

The CNN model may have the ability to naturally smooth the activations of voxels via the pooling and/or stride operations and thus may be better suited for noisy voxel-level data across the whole brain than the 1D-fcDNN, SVM, and the three alternative classifier models. Thus, the mean activation in the regions-of-interest (ROIs) defined from parcellated atlases was used as input instead of the whole-brain voxels, with the mean activation in the ROIs potentially reducing the noise in voxel-level data. Four atlases were employed to define the ROIs: automated anatomical labeling (AAL) with 116 regions (Tzourio-Mazoyer et al., 2002), Shen's 268 atlas (Shen et al., 2013), the HCP 360 atlas (Glasser et al., 2016), and Gordon's 333 atlas (Gordon et al., 2014). For the raw and minimally preprocessed fMRI volumes, the atlases in the MNI space were registered to the subject's EPI space using the SPM8 toolbox and the corresponding ROIs were defined in the subject's EPI space. The mean PSC intensity for each of the ROIs was then calculated and the 1D mean PSC across the ROIs was used as input for the 1D-fcDNN, SVM, and three alternative classifier models (the input size was 116×1 for the AAL 116 atlas, 268×1 for Shen's 268 atlas, 360×1 for the HCP 360 atlas, and 333×1 for Gordon's 333 atlas for both the fully preprocessed and minimally preprocessed fMRI 3D volume data). For the 3D-CNN model, the PSC intensity of each voxel in an ROI was replaced by the mean PSC intensity of the ROI, and the whole-brain fMRI volume (i.e., $53 \times 63 \times 46$ for the fully preprocessed volume data; $64 \times 64 \times 36$ for the minimally preprocessed volume data) was used as input.

It is possible that our adopted four sensorimotor tasks were so distinct that the use of complicated models such as our proposed 3D-CNN might be redundant and that the use of only highly task-relevant regions would be sufficient to correctly classify the four tasks using simpler models such as the SVM, LDA, RF, and LR. In this context, the average BOLD intensities from only highly task-related ROIs defined from the 116 AAL regions were used as the input of the classifiers. Of the 116 AAL regions, the union of the left/right precentral gyrus, left/right supplementary motor area, and the left/right postcentral gyrus was defined as an ROI for the two hand motor tasks (i.e., labels 2, 20, and 58 for the LH task; labels 1, 19, and 57 for the RH task); the union of the superior temporal

gyrus and middle temporal gyrus was defined as an ROI for the AD task (i.e., labels 81, 82, 85, and 86); and the union of the superior occipital gyrus, middle occipital gyrus, and inferior occipital gyrus was defined as an ROI for the VS task (i.e., labels 49, 50, 51, 52, 53, and 54). The average PSC intensity for each of these four ROIs was calculated and the 1D vectorized average PSC values across the four ROIs (4×1) was used as input for the SVM, LDA, RF, and LR classifiers.

2.10. Online classification of the sensorimotor tasks via real-time fMRI

We also evaluated the efficacy of the 3D-CNN model in online classification in comparison to the 1D-fcDNN and SVM models. The fMRI volumes were acquired in real-time (Kim et al., 2015; Lee et al., 2012) using the same EPI parameters as for the data from the original 12 subjects, while three additional healthy right-handed male volunteers (34, 32, and 27 years old) performed the four counter-balanced tasks with interleaved RS periods (Fig. 3(b)). Spatial smoothing and linear detrending were applied as minimal preprocessing options to overcome potential artifacts arising from the use of the MRI scanner (the online classification data were acquired approximately seven years after the acquisition of the sensorimotor data from the original 12 subjects using the same MRI scanner). The baseline BOLD intensity used to calculate the PSC fMRI volume was defined based on the fMRI volumes in the 10 s (5 TRs) at the beginning of the run once the first five volumes (10 s) had been discarded to account for the T1 effect. In this online classification, the in-brain mask was defined individually (without intersecting the in-brain masks across all subjects). In addition, to prevent the potential loss of voxels from the training data from the original 12 subjects, all voxels in the $64 \times 64 \times 36$ volume were vectorized ($147{,}456 \times 1$) and the corresponding vectorized fMRI data were used to train the 1D-fcDNN and SVM models. The $64 \times 64 \times 36$ volumes were used to train the 3D-CNN model. In this on-line classification scenario, the RS was added as a target class to continuously classify fMRI volumes acquired from the real-time fMRI along with the four target classes for the sensorimotor tasks (Fig. 3(b)). The fMRI volumes corresponding to the four RS periods between the task blocks (Fig. 3(a)) were used as training samples for the RS class using a balanced number of samples across the five classes during the training of the classifier models. Each of the fMRI volumes in the real-time fMRI (rtfMRI) run obtained from each of the three participants was then classified as one of the four tasks or the RS in real-time.

From a neurobiological point of view, it would be interesting to interpret how brain networks shift during task performance. To this end, the whole brain voxels ($64 \times 64 \times 36$) were parcellated into each of the 116 regions of the AAL atlas in offline analysis. The classification of task information was then conducted using the average BOLD intensity for each of the 116 AAL regions. The 3D volume with this average BOLD intensity was used as input for the 3D-CNN model and the 1D-vectorized pattern of the 116 regions (116×1) was used to train the 1D-fcDNN and SVM models. For the interpretation of brain networks, the activation maps (AMs) of our real-time fMRI data were calculated across the time points and for each of the five output nodes from our trained 3D-CNN model. Then, the AMs ($64 \times 64 \times 36$) were also parcellated into each of the 116 AAL regions, and each of the 116 AAL regions was assigned to one of the eight networks including (i) the default mode (DM) network (Bell and Shine, 2015; Bressler and Menon, 2010; Buckner, 2012; Power et al., 2011; Thomas Yeo et al., 2011; Yuan et al., 2016), (ii) the dorsal attention (DA) network (Bell and Shine, 2015; Fox et al., 2006; Power et al., 2011; Thomas Yeo et al., 2011; Vossel et al., 2014; Yuan et al., 2016), (iii) the ventral attention (VA) network (Bell and Shine, 2015; Fox et al., 2006; Power et al., 2011; Thomas Yeo et al., 2011; Vossel et al., 2014), (iv) the salience (SA) network (Menon, 2015; Shirer et al., 2012; Steimke et al., 2017; Yuan et al., 2016), (v) the fronto-parietal (FP) network (Power et al., 2011; Thomas Yeo et al., 2011; Zanto and Gazzaley, 2013), (vi) the sensorimotor (SM) network (Agosta et al., 2011; Biswal et al., 1995; Chenji et al., 2016; Power et al., 2011; Shirer et al., 2012; Yuan et al., 2016), (vii) the auditory (AUD) network (Power et al., 2011; Shirer et al., 2012; Yuan et al., 2016), and (viii) the visual (VIS) network (Power et al., 2011; Thomas Yeo et al., 2011; Yuan et al., 2016). More specifically, the mask for each of the eight brain networks was extracted using Yeo's seven networks (i.e., DM, DA, VA, FP, and VIS) (Thomas Yeo et al., 2011) and the 90 functional ROIs (i.e., SA, SM, and AUD; http://findlab.stanford.edu) (Shirer et al., 2012). Each of the AAL 116 regions was then assigned to the most highly overlapping of the eight networks. Finally, the change in the average percentage BOLD intensity for each of the eight functional networks was calculated for each of the two hemispheres and visualized in an image plot.

2.11. Classification using a Human Connectome Project (HCP) dataset

In the HCP dataset (minimally preprocessed 3T data from the S1200 release; (Van Essen et al., 2012), the first 100 young healthy participants who were tested on 23 conditions across seven tasks (i.e., two conditions each for emotion, language, social, relational, and gambling tasks; five conditions for a motor task; eight conditions for a working memory task) were used. Please refer to earlier reports for more details on the conditions of each task and the imaging acquisition parameters (Barch et al., 2013; Glasser et al., 2013; Van Essen et al., 2013; Van Essen et al., 2012). For the HCP data classification, three target class definitions were used: (1) all seven tasks, (2) all 23 conditions, and (3) seven selected conditions from seven tasks (i.e., *right-hand clenching* for the motor task, *fear* for the emotion task, *story* for the language task, *mental* for the social task, *relational* for the relational task, *loss* for the gambling task, and *2-back places* for the working memory task; (Wang et al., 2018). Five-fold CV was applied to evaluate the classification performance by separating the 100 subjects into five folds. The subjects in four folds were used to train the classifier and the subjects in the one remaining fold were used to test the trained classifier. This was repeated using each of the five folds as testing data.

As with our sensorimotor dataset, classification performance was assessed using both fully preprocessed (i.e., motion correction and spatial normalization followed by spatial smoothing) 3D volume data and minimally preprocessed (i.e., motion correction only, or motion correction followed by spatial smoothing) 3D volume data. Because the fMRI 3D volume data from the HCP were already motion-corrected and spatially normalized to the MNI with a 2-mm isotropic voxel size, resampling to a 3-mm isotropic voxel size followed by spatial smoothing using an 8-mm FWHM Gaussian kernel was additionally applied using SPM8. For the minimally preprocessed fMRI 3D volume data, the unprocessed fMRI 3D volume series underwent motion correction followed by EPI distortion correction as described in the HCP preprocessing pipelines (Glasser et al., 2013). The motion-corrected fMRI 3D volume series was then smoothed using an 8-mm FWHM Gaussian kernel. Gradient distortion correction could not be applied due to an inability to access the Skyra gradient field nonlinearity coefficients for the HCP Connectome Skyra (Siemens Skyra 3T scanner). The voxel size was not resampled to minimize the number of computational steps required that may distort the fMRI measurements. The minimal preprocessing implementation available in the HCP pipeline scripts (github.com/Washington-University/HCPpipelines) was used. The BOLD intensities of the fMRI volumes for each task period were normalized to PSC using the mean fMRI volume across all fMRI volume series in the corresponding run for each subject as a baseline. As a result, the fully preprocessed fMRI volume (i.e., $53 \times 63 \times 46$) and minimally preprocessed fMRI volume (i.e., $92 \times 104 \times 72$) with the PSC in BOLD intensities were used as input for the 3D-CNN model. The vectorized 1D patterns of these volumes within the in-brain mask (i.e., the resampled mask of "brainmask_fs.2.nii.gz" for the fully processed volumes; a mask defined using the FSL software as described in the HCP pipeline for minimally preprocessed volumes) were used as input for the 1D-fcDNN and SVM models (i.e., $52{,}129 \times 1$ for fully preprocessed volumes; $96{,}149 \times 1$ for minimally preprocessed volumes). Using the four atlases, the mean BOLD

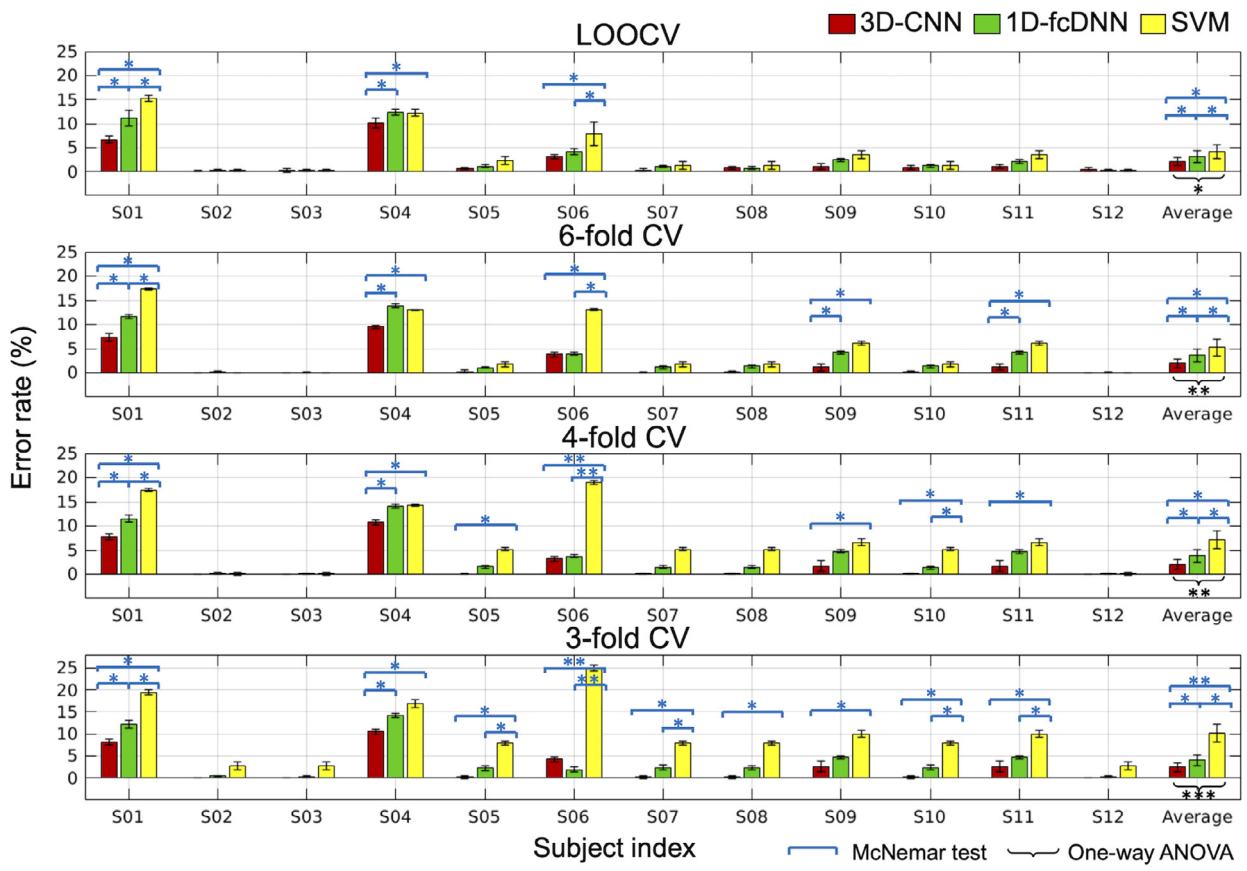

Fig. 4. Error rates (mean ± standard error) generated by the classifiers for each of the 12 subjects (as well as the summary statistics from all subjects in the final column) and from each of the scenarios for the *k*-fold CV scheme. The statistical significance was evaluated using (i) McNemar tests for within-subject evaluation using the results from randomly permuted *k*-fold sets and (ii) one-way ANOVA for between-subject evaluation using the average error rates across randomly permuted *k*-fold sets. 3D-CNN, 3D convolutional neural network; 1D-fcDNN, 1D fully connected deep neural network; SVM, support vector machine; LOOCV, leave-one-subject-out cross-validation; CV, cross-validation; S, subject; ANOVA, analysis-of-variance.

intensity within each of the ROIs was calculated for the fully preprocessed volumes and for the minimally preprocessed volumes as defined for our sensorimotor data (please refer to the section, "***2.9. Classification using parcellated ROIs across the whole brain***" for more details).

3. Results

3.1. Classification performance using the fully preprocessed data

The computation time using only the CPU to train the classifier model was approximately 8 h for the 3D-CNN model, 6.5 h for the 1D-fcDNN model (including 2.5 h for DBN pretraining), and 1 h for the linear SVM model for one fold using LOOCV (cf. approximately 6.5 h, 5.0 h, and 0.8 h for the 3-fold CV). The computation time using the GPU was substantially shorter at 2.0, 1.5, and 0.25 h for the 3D-CNN, 1D-fcDNN, and linear SVM models, respectively, using LOOCV and 1.5, 1.0, and 0.1 h using 3-fold CV. The computation time required to test a single fMRI volume using the trained classifier model was negligible (i.e., < 1 s) for all three models.

Fig. 4 presents the average error rates (± SE) for each subject and across all subjects for each of the classifier models and for each of the *k*-fold CV schemes. Overall, the 3D-CNN had significantly lower error rates (corrected $p < 0.05$ from one-way ANOVA; corrected $p < 0.05$ from McNemar tests) compared to the 1D-fcDNN and SVM models for all *k*-fold CV schemes. As the number of training samples became smaller (i.e., the lower the number of folds, *k*), the statistical significance of the difference in error rates between pairs of classifiers increased for each subject. Interestingly, individual variability in classification performance was also observed, with three subjects (S02, S03, and S12) exhibiting virtually perfect classification for all classifiers and *k*-fold CV schemes without any statistically significant difference between the three classifiers. On the other hand, another three subjects (S01, S04, and S06) had relatively higher error rates. For instance, S01 had significantly different error rates across all pairs among the three classifiers for all *k*-fold CV schemes (corrected $p < 0.05$ from the McNemar test). From the summary statistics across all subjects, the average error rates (%; ± SE) from the 3D-CNN, 1D-fcDNN, and SVM models were 2.1 (± 0.9), 3.1 (± 1.2), and 4.1 (± 1.5), respectively, using LOOCV (corrected $p < 0.035$ from the McNemar test; corrected $p = 0.041$ from one-way ANOVA). Using the 3-fold CV, the error rates were 2.4 (± 1.0), 4.2 (± 1.3), and 10.1 (± 2.0), respectively (corrected $p < 0.0082$ from the McNemar test; corrected $p = 2.5 \times 10^{-4}$ from one-way ANOVA). For the 1D-fcDNN model, the error rates were slightly lower from the cross-entropy loss function than the MSE for both the LOOCV and 3-fold CV schemes; however, there was no significant difference (Table S1).

3.2. Feature maps for the 3D-CNN in comparison to the 1D-fcDNN

Fig. 5 presents the average feature maps for the input volumes obtained from the trained 3D-CNN using LOOCV for (a) Conv1 ($n = 8$ from 8 3D convolution filters), (b) Conv2 ($n = 16$), and (c) Conv3 ($n = 32$). Each of the 3D feature maps was color-coded. Overall, the resulting feature maps of Conv1 were slightly shifted but covered task-specific brain areas such as the motor and cerebellum areas for the LH and RH tasks, the auditory areas for the AD task, and the visual area for the VS task. The feature maps of Conv2 were coarser than those of Conv1, and the feature maps of Conv3 even more so due to the stride in the convolution operation. It is notable, however, that the back-projected feature maps

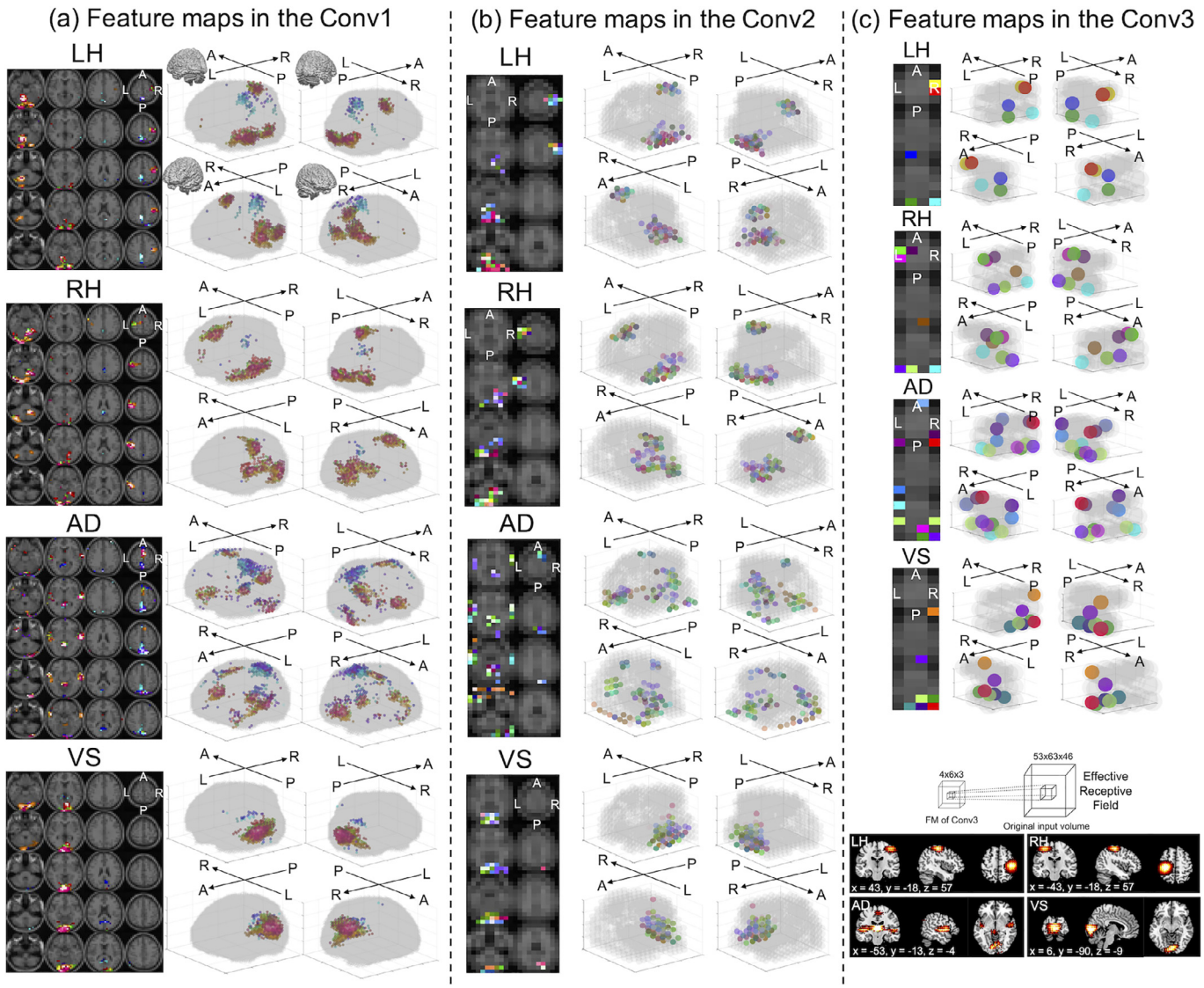

Fig. 5. Visualization of the feature maps obtained from the trained 3D-CNN model using LOOCV. Each feature map from a convolutional filter is presented as a different color, with the color gradient indicating the strength of the feature maps. Feature maps for (a) Conv1, (b) Conv2, and (c) Conv3 (feature maps back-projected onto the original input volume space using ERF mapping are presented at the bottom). Refer to the subsection "***3D-CNN***" in "***2.7. Interpretation of the trained models***" for details. LOOCV, leave-one-subject-out cross-validation; 3D-CNN, 3D convolutional neural network; FM, feature map; Conv1, the first convolutional layer; Conv2, the second convolutional layer; Conv3, the third convolutional layer; LH, left-hand clenching; RH, right-hand clenching; AD, auditory attention; VS, visual stimulation; L, left; R, right; A, anterior; P, posterior; ERF, effective receptive field.

of Conv3 into the input volume space covered remarkably task-specific brain areas for each of the four tasks (bottom of Fig. 5(c)). Fig. 6(a) displays the slight shift in the spatial feature patterns in Conv1 from the 3D-CNN in comparison to the mostly overlapping task-related weight feature maps in the first layer of the 1D-fcDNN. Fig. 6(b) presents the average feature maps obtained from two subjects for each task extracted from the scaled spatial extent of their neuronal activations.

3.3. Interpretations of the trained 3D-CNN

Fig. 7 shows all of the convolution filters in each of the three Conv layers and the resulting feature maps when the input was a single fMRI volume from the LH task or RH task. The top five percentile (red) and bottom five percentile (blue) values in the weights across all convolution filters were color-coded and the values in the top one percentile in the feature maps were visualized. The spatial layout of the convolution filters in each Conv layer varied, while the extracted feature maps were remarkably task-specific across the filters, resulting in slightly shifted feature maps for each of the four tasks (Figs. 5 and 6). Moreover, a single convolution kernel can extract features for multiple input patterns obtained from different target classes as we illustrated for the LH and RH tasks, thus representing a potential advantage of the proposed 3D-CNN model.

Fig. 8 shows the class saliency maps (a) and class activations maps (b) for each of the four tasks across all of the input samples from all of the subjects for the training of the 3D-CNN model in the LOOCV framework. Overall, both the resulting class saliency maps and class activation maps were remarkably task-specific (i.e., mainly the contralateral motor areas for the LH and RH tasks, the left auditory areas for the AD task, and the primary visual area for the VS task) in comparison to the feature maps and their active regions shown in Figs. 5 and 7, respectively.

3.4. Classification performance using raw and minimally preprocessed fMRI volumes

Table 1 summarizes the average error rates for the 3D-CNN, 1D-fcDNN, and SVM models using the raw and minimally preprocessed fMRI volumes for both LOOCV and 3-fold CV. Overall, the 3D-CNN

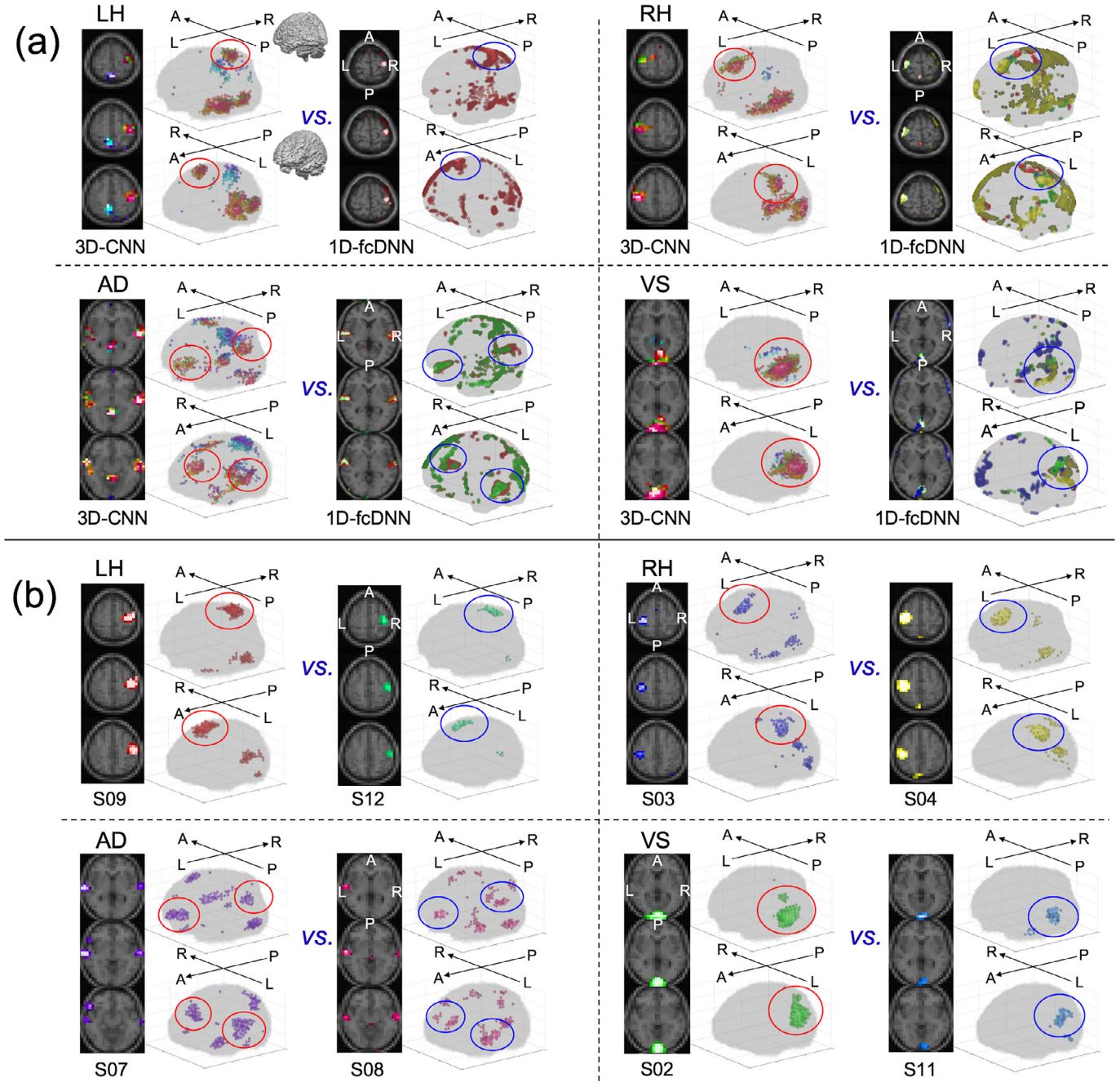

Fig. 6. (a) Comparison of the feature representations obtained from the 3D-CNN and 1D-fcDNN. (b) The feature maps in Conv1 obtained from a pair of two subjects who demonstrated scaled neuronal activations. 3D-CNN, 3D convolutional neural network; 1D-fcDNN, 1D fully connected deep neural network; S, subject; LH, left-hand clenching; RH, right-hand clenching; AD, auditory attention; VS, visual stimulation; L, left; R, right; A, anterior; P, posterior; Conv1, the first convolutional layer.

model outperformed the 1D-fcDNN and SVM models for all preprocessing scenarios, with the 1D-fcDNN and SVM models exhibiting a considerably lower performance when compared to the results from the fully preprocessed data. In addition, spatial smoothing (SS) more substantially enhanced classification performance compared to slice timing correction (STC) and motion correction (MC). For all three classifier models, the lowest error rates were obtained from the processing sequence of STC, MC, and SS (mean error rates of 2.6%, 3.9%, and 4.0% for the 3D-CNN, 1D-fcDNN, and SVM models, respectively, using LOOCV). The error rates were substantially higher for the raw fMRI 3D volume data (26.2%, 75.0%, and 75.0% for the 3D-CNN, 1D-fcDNN, and SVM models, respectively), with only the 3D-CNN model achieving a classification accuracy (73.8%) that was greater than the level of chance (25%). The error rates were substantially higher when 3-fold CV was employed (mean error rates of 4.0%, 12.2%, and 18.1% for the 3D-CNN, 1D-fcDNN, and SVM models, respectively), particularly for the 1D-fcDNN and SVM models, possibly indicating overfitting to a reduced number of training samples.

Fig. 9(a) presents the classification results for the 3D-CNN, 1D-fcDNN, SVM, and the three alternative classifier models using fully/minimally preprocessed and raw fMRI data in the LOOCV framework. Overall, it is evident that the 3D-CNN model had the lowest error rates compared to the other five classifiers for all preprocessing scenarios. The three alternative classifiers (LDA, RF, and LR) also performed

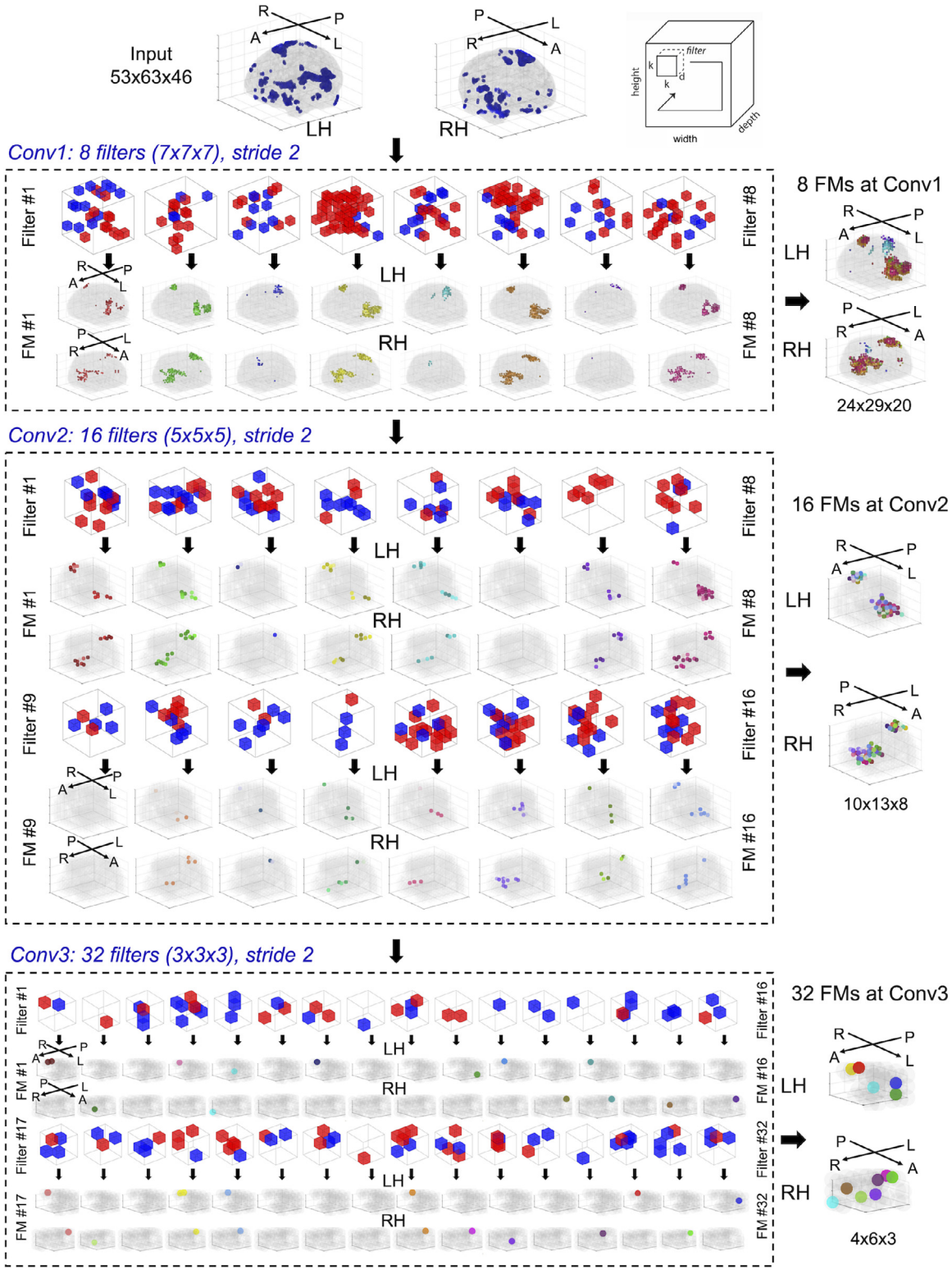

Fig. 7. The learned filters and the corresponding feature maps obtained from each of the three Conv layers of our 3D-CNN model. The positive weights (red) and the negative weights (blue) were collected from the top 5 percentile of the weight values across all filters. The feature maps of each filter were constructed when a single input volume from the LH or RH task was fed into the 3D-CNN model by visualizing only the top one percentile of values. 3D-CNN, 3D convolutional neural network; Conv, convolutional; FM, feature map; L, left; R, right; A, anterior; P, posterior; LH, left-hand clenching; RH, right-hand clenching.

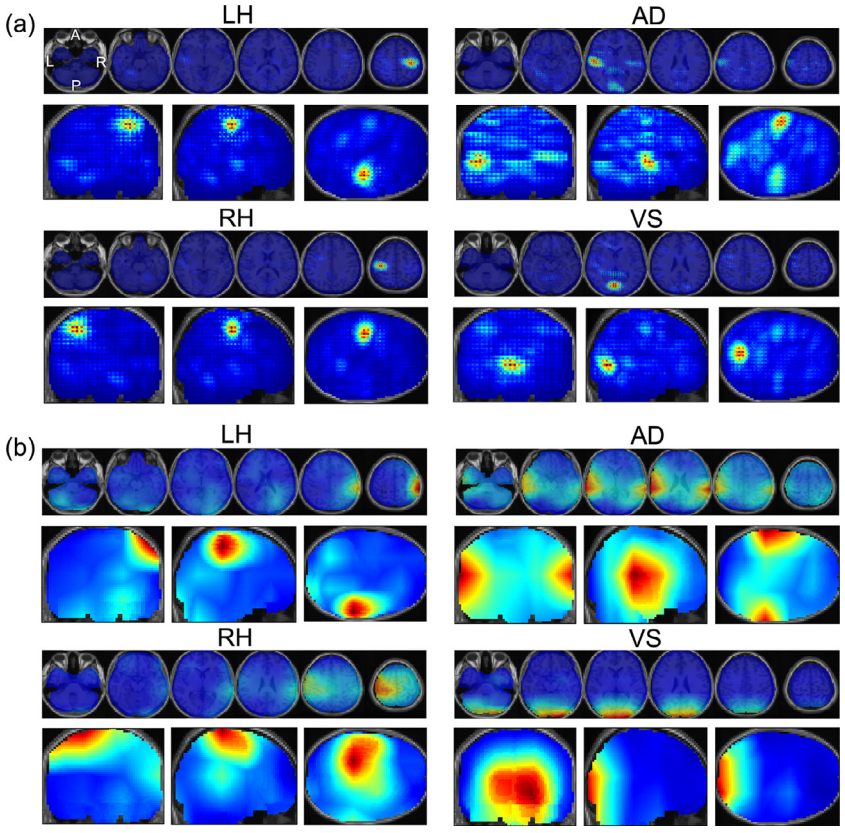

Fig. 8. Visualization of (a) the class saliency maps and (b) class activation maps obtained from the trained 3D-CNN model in the LOOCV for each of the four sensorimotor tasks (i.e., LH, RH, AD, and VS). Please refer to the sections, "***3D-CNN***" in "***2.7. Interpretation of the trained models***" and "***Visualization of 3D-CNN model using class saliency map and class activation map***" in the Supplementary Materials for more details. LH, left-hand clenching; RH, right-hand clenching; AD, auditory attention; VS, visual stimulation; L, left; R, right; A, anterior; P, posterior.

Table 1
Error rates (mean ± standard error) using raw and minimally preprocessed fMRI data as input for each of the three classifiers (3D-CNN, 1D-fcDNN, and SVM) using LOOCV or 3-fold CV. The error rates from the fully preprocessed fMRI volumes (Fig. 4) are presented in the last row for comparison. Statistical significance of the error rates obtained from the three classifier models was established using one-way ANOVA, with the *F*-score and the corresponding *p*-value reported.

Preprocessing scenario	LOOCV (%)				3-fold CV (%)			
	3D-CNN	1D-fcDNN	SVM	F-scorep-value	3D-CNN	1D-fcDNN	SVM	F-scorep-value
None	26.2 ± 6.3	75.0 ± 0.0	75.0 ± 0.0	$F = 21.7$ $p = 9.5 \times 10^{-7}$	39.5 ± 3.3	75.0 ± 0.0	75.0 ± 0.0	$F = 720.7$ $p < 10^{-10}$
STC	3.9 ± 2.1	8.8 ± 4.7	10.4 ± 5.7	$F = 6.1$ $p = 0.038$	5.6 ± 1.1	16.3 ± 2.4	20.6 ± 2.7	$F = 10.5$ $p = 6.9 \times 10^{-5}$
MC	3.5 ± 1.9	8.7 ± 4.0	11.0 ± 5.9	$F = 6.5$ $p = 0.025$	5.4 ± 1.0	15.8 ± 2.0	21.3 ± 2.5	$F = 22.3$ $p = 1.0 \times 10^{-8}$
SS	2.8 ± 1.5	4.6 ± 2.5	5.1 ± 3.2	$F = 5.2$ $p = 0.043$	4.4 ± 0.9	12.2 ± 1.8	18.7 ± 2.1	$F = 11.6$ $p = 5.3 \times 10^{-5}$
STC+MC	3.4 ± 2.1	9.1 ± 5.4	9.9 ± 4.6	$F = 9.4$ $p = 0.011$	5.5 ± 1.5	16.6 ± 2.9	21.1 ± 2.3	$F = 21.3$ $p = 3.1 \times 10^{-7}$
STC+SS	3.4 ± 2.2	4.4 ± 2.6	5.4 ± 3.5	$F = 6.9$ $p = 0.03$	4.7 ± 0.9	12.5 ± 2.1	18.8 ± 2.2	$F = 10.1$ $p = 5.8 \times 10^{-5}$
MC+SS	3.4 ± 2.0	4.5 ± 2.5	4.2 ± 2.8	$F = 5.5$ $p = 0.041$	4.6 ± 1.1	12.3 ± 1.6	19.2 ± 2.5	$F = 11.9$ $p = 6.2 \times 10^{-5}$
STC+MC+SS	2.6 ± 1.6	3.9 ± 2.1	4.0 ± 2.5	$F = 6.05$ $p = 0.04$	4.0 ± 0.8	12.2 ± 1.2	18.1 ± 2.0	$F = 12.7$ $p = 4.7 \times 10^{-5}$
STC+MC+ NOR+SS	2.1 ± 0.9	3.1 ± 1.2	4.1 ± 1.5	$F = 6.0$ $p = 0.041$	2.4 ± 1.0	4.2 ± 1.3	10.1 ± 2.0	$F = 8.6$ $p = 2.5 \times 10^{-4}$

fMRI, functional magnetic resonance imaging; BOLD, blood-oxygenation-level-dependent; 3D-CNN, 3D convolutional neural network; 1D-fcDNN, 1D fully connected deep neural network; SVM, support vector machine; STC, slice timing correction; MC, motion correction; SS, spatial smoothing; NOR, spatial normalization; LOOCV, leave-one-subject-out cross-validation; CV, cross-validation; ANOVA, analysis-of-variance.

well using the fully preprocessed data (e.g., the error rate of 4.0% from the LDA is comparable to the error rate of 4.1% from the SVM). However, the difference in the error rates between the 3D-CNN model and other alternative classifiers widened when only MC and/or STC were applied.

Table 2 summarizes the classification performance for the systematically altered 3D-CNN architectures. Overall, the 3D-CNN model with a max pooling layer following a Conv layer exhibited the lowest error rates. The error rates (a) from the 3D-CNN model with stride and without a pooling layer and (b) from the 3D-CNN model with an average pooling layer but without stride appeared to be comparable and were slightly higher than the error rates from the 3D-CNN model with a max pooling layer particularly when there was no Fc layer. The error rates from the 3D-CNN model without a down-sampling operation (i.e., without

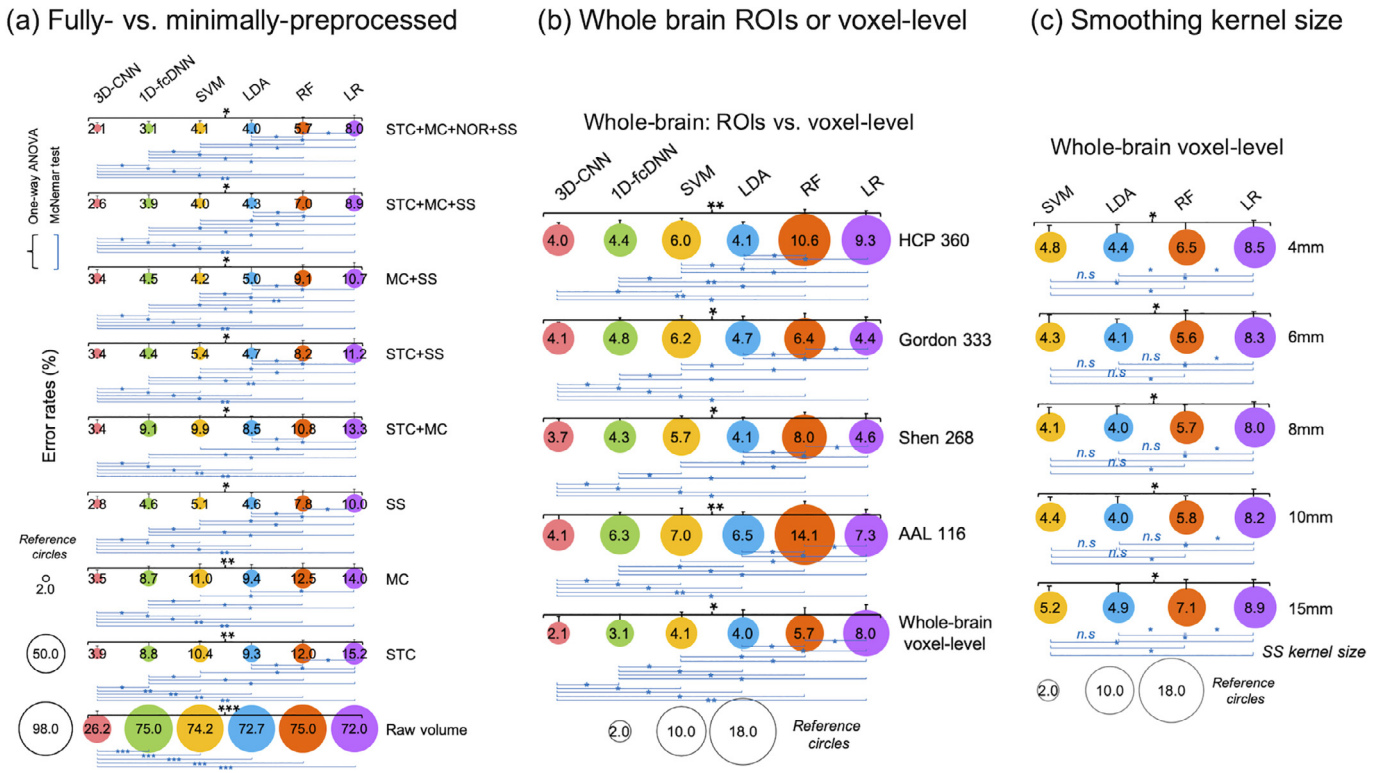

Fig. 9. (a) Average error rates obtained from each of the classifiers using fully processed and minimally preprocessed fMRI volume data in the LOOCV scheme. (b) Average error rates from each of the classifiers using the ROI-based input patterns in comparison to the whole-brain voxel-level data in the LOOCV scheme using the fully preprocessed fMRI volume data. (c) Average error rates obtained from each of the four baseline machine learning classifiers (i.e., SVM, LDA, RF, and LR) using whole-brain voxel-level input patterns from fully preprocessed fMRI volume data that had been spatially smoothed with isotropic Gaussian smoothing kernels of various sizes in the LOOCV scheme. The statistical significance was assessed with one-way ANOVA and McNemar tests (*, corrected $p < 0.05$; **, corrected $p < 0.01$; ***, corrected $p < 0.001$). 3D-CNN, 3D convolutional neural network; 1D-fcDNN, 1D fully connected deep neural network; SVM, support vector machine; LDA, linear discrimination analysis; RF, random forest; LR, logistic regression; STC, slice timing correction; MC, motion correction; SS, spatial smoothing; NOR, spatial normalization; AAL, Automated Anatomical Labeling 116 regions; Shen 268, Shen's 268 regions atlas; HCP 360, HCP 360 atlas; Gordon 333, Gordon's 333 regions atlas; ANOVA, analysis-of-variance.

stride and without a pooling layer) were moderately higher compared to those from the 3D-CNN model with a down-sampling operation. The error rates from the model without a Conv layer and with only a pooling layer were substantially higher. Overall error rates moderately increased when a Fc layer was not used compared to when a Fc layer was used; however, the number of parameters was substantially smaller in the 3D-CNN model without the Fc layer. The computational time using the GPU to train the model was substantially lower for the 3D-CNN model with stride (i.e., approximately 2 h) compared to the 3D-CNN model without stride and without a max pooling layer (approximately 3.5 h) using the fully-preprocessed data in the LOOCV framework.

From a 1D-fcDNN model that employed approximately the same number of parameters as the 3D-CNN model using the down-sampled fMRI volume data for 1D input patterns, the average error rate for the modified 1D-fcDNN model was 3.3 (± 1.1), which was slightly higher compared to the performance of the 1D-fcDNN using fMRI volumes before down-sampling (i.e., 3.0 ± 1.0) and was moderately lower than atlas-based classification (e.g., 4.3 ± 1.1 using Shen's 268 atlas; Fig. S2). On the other hand, the 3D-CNN model (with a stride of 2 in each Conv layer) had an average error rate of 2.1 (± 0.9) using whole-brain volume data in the LOOCV framework.

3.5. Classification performance using parcellated ROIs from atlases

Fig. 9(b) shows that the 3D-CNN model outperformed the 1D-fcDNN, SVM, and the three alternative classifier models using input patterns defined using the parcellated ROIs from each of the four atlases as well as the whole-brain voxel-level data. The confusion matrix for the classification results using the fully preprocessed data (Fig. S2a) indicates that classification performance for the AD task was particularly compromised in all three classifier models (i.e., 3D-CNN, 1D-fcDNN, and SVM) using the parcellated ROIs as input in comparison to the whole-brain voxel-level fMRI volume data. Figure S2b presents an example fMRI volume for the AD task, which was misclassified as the VS task using ROI-based classification for the AAL 116, Shen 268, and HCP 360 atlases but was correctly classified as the AD task for the Gordon 333 atlas. It is notable that the overall PSC intensities were blurred in the ROI-based patterns compared to the voxel-level PSC intensities, and the average PSC for the visual area from the three misclassified atlases was stronger than that from the Gordon 333 atlas. The number of parameters for the 1D-fcDNN model was drastically reduced using the ROI-based input patterns in comparison to the whole-brain voxel-level input patterns (Table S2). As a result, the computational time required to train the 1D-fcDNN model for whole-brain voxel-level input patterns (i.e., 1.5 h) using the GPU was substantially reduced using the ROI-based input patterns defined from the atlases (i.e., 0.20, 0.25, 0.30, and 0.30 h for the AAL 116, Shen 268, Gordon 333, and HCP 360 atlases, respectively).

Fig. S3 shows that the 3D-CNN model outperformed the 1D-fcDNN and SVM models for the classification of the four sensorimotor tasks using the minimally preprocessed fMRI volume data in the ROI-based and LOOCV framework. Overall, the error rates for the Shen 268, Gordon 333, and HCP 360 atlases were comparable, whereas the error rate for the AAL 116 atlas was slightly higher than the other three. Using the raw fMRI volumes, the 3D-CNN models showed significantly reduced error rates using the whole-brain voxel-level input patterns (i.e., 26.2%) compared to the whole-brain ROI-based input patterns (i.e., approximately

Table 2
Error rates (mean ± standard error) of the 3D-CNN models with varying architectural designs using both fully preprocessed volume data and raw/minimally preprocessed volume data.

3D-CNN model		With Conv layers				Without Conv layers(and without stride)	
		With stride and without pooling	Without stride Avg pooling	Max pooling	No pooling	Avg pooling	Max pooling
With a Fc layer	*Raw volume*	*26.2 ± 6.3*	27.5 ± 6.1	**25.1 ± 6.3**	28.8 ± 7.3	35.5 ± 7.3	31.6 ± 7.2
	STC	*3.9 ± 2.1*	4.0 ± 2.2	**3.8 ± 1.8**	4.9 ± 2.3	8.9 ± 3.1	8.8 ± 3.2
	MC	3.5 ± 1.9	*3.5 ± 1.8*	**3.3 ± 1.1**	4.8 ± 2.4	8.7 ± 3.2	8.7 ± 3.9
	SS	*2.8 ± 1.5*	2.9 ± 1.3	**2.8 ± 1.4**	3.9 ± 1.6	7.2 ± 2.6	7.0 ± 2.5
	STC+MC	**3.4 ± 2.1**	3.8 ± 1.6	*3.5 ± 2.0*	4.7 ± 2.2	8.8 ± 3.0	8.5 ± 3.1
	STC+SS	*3.4 ± 2.2*	3.6 ± 2.0	**3.3 ± 2.0**	4.0 ± 1.6	7.5 ± 2.3	7.4 ± 2.7
	MC+SS	*3.4 ± 2.0*	3.5 ± 2.2	**3.4 ± 1.9**	4.1 ± 2.0	7.5 ± 2.5	7.3 ± 2.1
	STC+MC+SS	*2.6 ± 1.6*	2.8 ± 1.7	**2.6 ± 1.3**	3.8 ± 1.5	7.0 ± 2.7	6.8 ± 2.4
	Number of parameters	301,220*	301,220*	301,220*	265,820,324	9860*	9860*
	Fully preprocessed	*2.1 ± 0.9*	2.3 ± 1.0	**2.0 ± 1.1**	3.6 ± 1.4	6.7 ± 2.2	6.4 ± 2.0
	Number of parameters	301,220*	301,220*	301,220*	291,207,332	9860*	9860*
Without a Fc layer	*Raw volume*	*28.4 ± 6.1*	29.8 ± 6.3	**28.3 ± 6.0**	31.3 ± 6.5	39.6 ± 7.4	36.2 ± 7.1
	STC	*4.7 ± 2.3*	4.9 ± 2.2	**4.2 ± 1.8**	7.0 ± 2.7	11.6 ± 4.1	9.5 ± 3.8
	MC	*4.5 ± 2.1*	4.8 ± 2.1	**4.0 ± 1.1**	7.1 ± 2.5	10.8 ± 3.7	9.4 ± 3.6
	SS	*3.5 ± 1.8*	3.8 ± 1.4	**3.2 ± 1.8**	5.9 ± 1.8	9.0 ± 3.0	8.8 ± 2.8
	STC+MC	*4.2 ± 2.3*	4.3 ± 1.7	**3.9 ± 2.1**	6.9 ± 2.1	9.9 ± 3.6	9.1 ± 4.1
	STC+SS	*3.8 ± 2.0*	4.1 ± 2.0	**3.7 ± 2.2**	6.3 ± 1.9	9.2 ± 3.1	8.6 ± 2.9
	MC+SS	*3.9 ± 1.9*	4.0 ± 2.1	**3.7 ± 1.7**	6.0 ± 2.1	9.1 ± 3.5	8.7 ± 2.7
	STC+MC+SS	*3.4 ± 1.2*	3.7 ± 1.2	**2.9 ± 1.5**	5.8 ± 1.8	8.9 ± 3.0	8.6 ± 2.3
	Number of parameters	14,884*	14,884*	14,884*	8,312,356	292*	292*
	Fully preprocessed	*3.2 ± 1.3*	3.4 ± 1.4	**2.4 ± 1.2**	5.7 ± 1.6	8.6 ± 2.7	8.4 ± 2.5
	Number of parameters	14,884*	14,884*	14,884*	9,105,700	292*	292*

The pooling (i.e., 2 × 2 × 2) and stride (i.e., 2 × 2 × 2) sizes were fixed.
The lowest error rate is shown in bold and the second-lowest in italics for each row.
The 3D-CNN model with Conv layers, with stride, and without pooling is the model shown in Fig. 1(c).
Avg, average; STC, slice timing correction; MC, motion correction; SS, spatial smoothing; NOR, spatial normalization; 3D-CNN, 3D convolutional neural network; Conv, convolutional layer; Fc, fully connected layer.
* , the number of parameters for the fully preprocessed fMRI volume data and raw/minimally preprocessed fMRI volume data are equal because the sizes of feature maps for each channel at the Conv3 were 4 × 6 × 3 (=72) and 6 × 6 × 2 (=72) for the fully preprocessed fMRI volume data and raw/minimally preprocessed fMRI volume data, respectively. Also, the number of parameters for the 3D-CNN models (a) with stride only and without a pooling layer and (b) without stride and with a pooling layer are same due to the same down-sampling operations of 2 × 2 × 2 voxels in both cases.

46.8%), whereas the 1D-fcDNN and SVM models demonstrated a performance level similar to that for chance (i.e., a 75% error rate).

Table S3 presents the error rates from the input patterns of the average PSC BOLD intensities from the four highly task-related ROIs only using the SVM, LDA, RF, and LR classifiers in comparison to the error rates from the whole-brain voxel-level and whole-brain ROI-level input patterns. The error rates clearly demonstrated that the whole-brain voxel-level input patterns outperformed the ROI-based input patterns and that the whole-brain ROI-based input patterns outperformed the input patterns based only on the highly task-related brain regions.

Fig. 9(c) presents the average error rates obtained from the baseline machine learning classifiers (i.e., SVM, LDA, RF, and LR) using the input patterns obtained from fMRI volume data smoothed using Gaussian smoothing kernels of various sizes. The smoothing kernel size of 8 mm produced the lowest error rates; as the smoothing kernel size moved further from 8 mm (either higher or lower), the greater the error rates became. For example, the SVM classifier had error rates of 4.8%, 4.1%, and 5.2% for kernel sizes of 4, 8, and 15 mm, respectively.

3.6. Performance in online classification of the sensorimotor tasks using real-time fMRI

Table 3 summarizes the results for the online five-class classification (i.e., four sensorimotor tasks and RS; chance level = 20%) obtained from each of the three classifier models and the three participants. Similar to the offline classification results for the original 12 subjects, the 3D-CNN model (average accuracy of 78.5%) outperformed the 1D-fcDNN (26.7%) and SVM (21.5%; Bonferroni-corrected $p = 8.6 \times 10^{-4}$ from one-way ANOVA) models using the whole-brain voxel-level input patterns. Fig. 10(a) presents the online classification results using the whole-brain voxel-level patterns for the first participant as an example. Fig. 10(b) shows the offline classification results using the 116 AAL region-based patterns of the participant as well as the transition of the eight brain networks while the subject was performing the tasks during real-time fMRI acquisition based on the input patterns and AMs of the trained 3D-CNN classifiers. As consistently found from the classification of four tasks using non-real-time data, the classification performance using whole-brain ROI-based input patterns was compromised compared to the use of whole-brain voxel-level input patterns. Interestingly, there were apparent changes and/or shifts in the active brain networks when the four tasks alternated with resting periods. For example, while both the RH and LH tasks clearly showed dominant motor activations in the contralateral hemisphere in the input patterns, the LH task also showed increased activations in multiple functional networks including the DA, VA, FP, and VIS, particularly in the first task block compared to the second task block. In the AMs of the corresponding output nodes for these two motor tasks, dominant activations are evident only in the contralateral sensorimotor networks. For the VS task, the DA, FP, and SM networks, as well as the VIS network, were dominant, particularly in the second task block compared to the first block of the input patterns. However, these multiple functional networks were substantially suppressed except for the VIS network in the AMs. The AD task clearly showed significantly small activations across all of the functional networks in the input patterns, in which increased activations at the beginning of the first block were substantially decreased. In the AMs of the output node of the AD task, only the DA and AUD networks showed increased intensities which could have enhanced the classification performance. Strong negative activations in the input patterns and virtual non-existence of substantially active functional network in the AMs across all of the networks during the RS periods compared to the task periods were also evident.

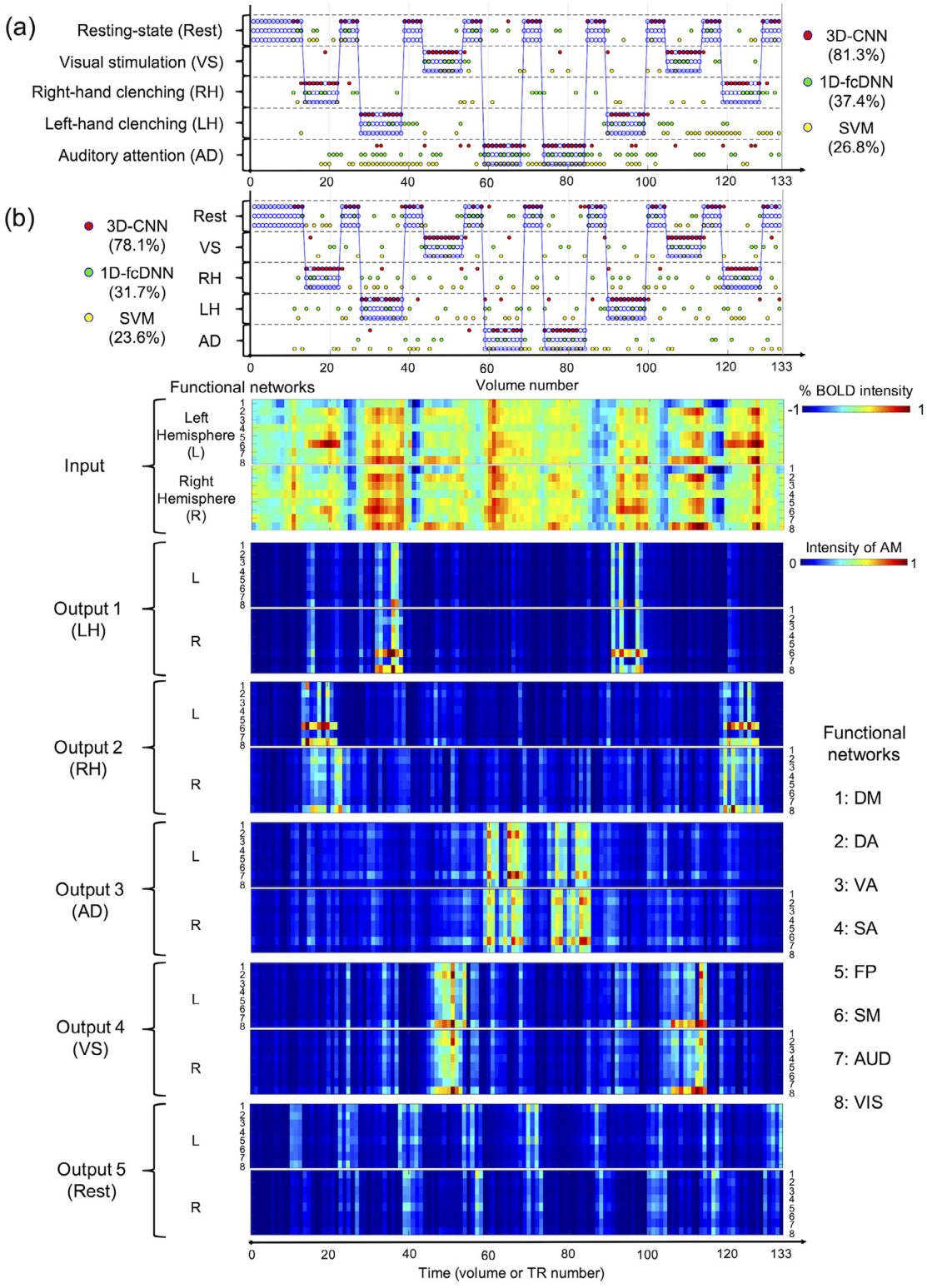

Fig. 10. (a) Results of the online classification of fMRI 3D volume data acquired in real-time from one participant using each of the three classifier models (i.e., 3D-CNN, 1D-fcDNN, and SVM). The white circles indicate a target task based on the experimental paradigm (Fig. 3(b)), and the colored circles indicate the classified labels from each of the three trained models. The first five volumes were not classified to allow the T1 effect to equilibrate and the next five volumes after that were not classified to reserve volumes for the definition of the baseline BOLD intensity and to normalize the subsequent fMRI volumes (with arbitrary intensity values) into PSC intensity. Thus, online classification began with the 11th fMRI volume. (b) Results from offline classification using the input patterns defined from the whole-brain ROIs using the 116 AAL regions: correctly and incorrectly classified samples (top row) and the changes of (i) the percentage BOLD intensities of the input patterns and (ii) intensities of activation maps (AMs) for each of the five output nodes across eight functional networks during the task performance of the subject (bottom row) are shown. 3D-CNN, 3D convolutional neural network; 1D-fcDNN, 1D fully connected deep neural network; SVM, support vector machine; PSC, percentage signal change; BOLD, blood-oxygenation-level-dependent; ROI, region-of-interest; AAL, Automated Anatomical Labeling; DM, default mode network; DA, dorsal attention network; VA, ventral attention network; SA, salience network; FP, fronto-parietal network; SM, sensorimotor network; AUD, auditory network; VIS, visual network; TR, repetition time.

Table 3
Online five-class classification performance for the four sensorimotor tasks and RS for three participants using the pretrained 3D-CNN, 1D-fcDNN, and SVM models based on whole-brain voxel-level patterns. The average accuracy (± standard error) across all three participants and the number of correctly classified volumes for each task are also reported. Statistical significance was established using one-way ANOVA, with the *F*-score and the corresponding *p*-value reported.

Subject number	*Model*	*Number of correctly classified volumes (a total number of volumes)*						*Accuracy (%)*
		LH	*RH*	*AD*	*VS*	*RS*	*Total*	
1	*3D-CNN*	16 (21)	15 (19)	19 (21)	16 (19)	34 (43)	100 (123)	81.3
	1D-fcDNN	7 (21)	8 (19)	14 (21)	9 (19)	8 (43)	46 (123)	37.4
	SVM	2 (21)	2 (19)	17 (21)	5 (19)	7 (43)	33 (123)	26.8
2	*3D-CNN*	8 (21)	17 (18)	18 (20)	16 (18)	34 (43)	93 (120)	77.5
	1D-fcDNN	4 (21)	11 (18)	10 (20)	5 (18)	1 (43)	31 (120)	25.8
	SVM	3 (21)	7 (18)	8 (20)	4 (18)	4 (43)	26 (120)	21.7
3	*3D-CNN*	8 (19)	13 (18)	16 (18)	19 (21)	35 (43)	91 (129)	76.6
	1D-fcDNN	3 (19)	4 (18)	3 (18)	4 (21)	6 (43)	20 (129)	16.8
	SVM	1 (19)	3 (18)	3 (18)	5 (21)	7 (43)	19 (129)	15.9
Model		*3D-CNN*	*1D-fcDNN*	*SVM*		*F-score, p-value*		
Mean Accuracy (%)		78.5 ± 1.4	26.7 ± 5.9	21.5 ± 3.1		$F = 62.7$, $p = 8.6 \times 10^{-4}$		

3D-CNN, 3D convolutional neural network; 1D-fcDNN, 1D fully connected deep neural network; SVM, support vector machine; LH, left-hand clenching; RH, right-hand clenching; AD, auditory attention; VS, visual stimulation; RS, resting-state; AVOVA, analysis-of-variance.

Fig. 11. Classification error rates using the HCP dataset from the three classifier models (i.e., 3D-CNN, 1D-fcDNN, and SVM) (i) for fully preprocessed and minimally preprocessed fMRI volume data (i.e., MC, MC+SS, and MC+NOR+SS), (ii) for each of the three classification test scenarios (i.e., 7 tasks, 23 conditions, and 7 selected conditions as target classes), and (iii) for each of the four atlases using the ROI-based input patterns in comparison to the whole-brain voxel-level fMRI data. The statistical significance was assessed using one-way ANOVA and McNemar tests (*, corrected $p < 0.05$; **, corrected $p < 0.01$; ***, corrected $p < 0.001$). 3D-CNN, 3D convolutional neural network; 1D-fcDNN, 1D fully connected deep neural network; SVM, support vector machine; MC, motion correction; SS, spatial smoothing; NOR, spatial normalization; ANOVA, analysis-of-variance; ns, not significant.

3.7. Classification performance using HCP data

Fig. 11 summarizes the error rates for the fully preprocessed and minimally preprocessed HCP fMRI volume data for each of the three classification test scenarios (i.e., seven tasks, 23 conditions, and seven selected conditions across the seven tasks) using the 3D-CNN, 1D-fcDNN, and SVM models, as well as the error rates for each of the four atlases for ROI-based classification. Overall, the 3D-CNN model had lower error rates across all scenarios than did the 1D-fcDNN and SVM models. For example, using the fully preprocessed volumes, the average error rates (%; ± SE) for the 3D-CNN, 1D-fcDNN, and SVM models when classifying the seven tasks were 22.4 (± 2.5), 27.5 (± 3.4), and 33.1 (± 3.5), respec-

tively, and when classifying the seven selected conditions, they were 8.4 (± 1.1), 12.5 (± 1.5), and 17.7 (± 1.6), respectively. Moreover, the error rates using the whole-brain voxel-level input patterns were consistently lower than the error rates using the ROI-based input patterns, which is in line with the results from our sensorimotor dataset.

Using the minimally preprocessed fMRI volume data, it is notable that the 3D-CNN model was superior to the 1D-fcDNN and SVM models. The average error rates (%; ± SE) for the 3D-CNN, 1D-fcDNN, and SVM models were 37.8 (± 3.7), 59.2 (± 4.1), and 65.9 (± 4.3), respectively, for the seven-task classification and 30.2 (± 3.3), 44.1 (± 3.6), and 46.5 (± 3.8), respectively, for the seven conditions selected as target classes using motion-corrected whole-brain voxel-level input patterns. These error rates were compromised for all classifiers using the ROI-based approach. The error rates were also substantially reduced using spatial smoothing followed by motion correction from all three classifiers. For instance, the error rate for motion-corrected and additionally spatially smoothed voxel-level volume data was 15.6 (± 2.3), which was substantially lower than the error rate for motion-corrected volume data (i.e., 30.2 ± 3.3) using the 3D-CNN model for the classification of the seven selected conditions.

4. Discussion

4.1. Summary of the study

In this study, we investigated the efficacy of a 3D-CNN model for the classification of whole-brain BOLD fMRI volumes in comparison to the 1D-fcDNN, SVM, and the three alternative classifiers (i.e., LDA, RF, and LR). The 3D-CNN model exhibited a classification accuracy that was superior to that of the 1D-fcDNN, SVM, and the three alternative classifier models for fully preprocessed fMRI 3D volume data with spatial normalization, raw and minimally preprocessed fMRI 3D volume data without spatial normalization, and online classification. For example, the classification error rates (± SE) for the 3D-CNN, 1D-fcDNN, and SVM models using our sensorimotor data based on 3-fold CV were 2.4% (± 1.0), 4.2% (± 1.3), and 10.1% (± 2.0) when the data were fully preprocessed (Bonferroni-corrected $p = 2.5 \times 10^{-4}$ from one-way ANOVA) and 4.4% (± 0.9), 12.2% (± 1.5), and 18.7% (± 2.3) when the data was only spatially smoothed (Bonferroni-corrected $p = 5.3 \times 10^{-5}$). The online classification accuracy was 78.5% (± 1.4), 26.7% (± 5.9), and 21.5% (± 3.1) using data acquired via real-time fMRI (Bonferroni-corrected $p = 8.6 \times 10^{-4}$). Compared to our sensorimotor dataset, classification using the HCP dataset to which only motion correction had been applied produced higher error rates for all three classifier models (i.e., 3D-CNN, 1D-fcDNN, and SVM) than did the fMRI volume data to which spatial smoothing was also applied. This was possibly because the raw HCP volume data were noisier than our sensorimotor volume data despite motion correction and spatial smoothing substantially recovering task-related activations (Fig. S4).

4.2. Comparison of the 3D-CNN model with alternative classifier models

The superior performance of the 3D-CNN model in comparison to the 1D-fcDNN, SVM, and the three alternative classifier models is possibly because the 3D-CNN model is able to handle shifted and scaled neuronal activations in local functional networks via down-sampling operations such as the use of a pooling layer and/or stride during convolution operations (Figs. 5–7). The spatial layout of the neuronal activations in the local brain regions is shifted and scaled due to imperfections in the spatial normalization algorithms, differences in slice timing, and head motion, while there is also individual variability in hemodynamic responses caused by differences in physiological responses derived from the neuronal activity measured using BOLD signals across runs, sessions, and/or subjects (Aguirre et al., 1998; Calhoun et al., 2017; Dohmatob et al., 2018; Handwerker et al., 2004; Tavor et al., 2016). Thus, the down-sampling operations in the 3D-CNN may have processed these shifted and scaled neuronal activations in a variety of spatial layouts across the whole brain obtained from the learned 3D convolution filters/kernels. In addition, the 3D-CNN provided richer information on the neuronal activations across the whole brain than did the 1D-fcDNN, which is potentially vulnerable to the spatial misalignment of voxels (Fig. 1(b) and 6(a)). This ability of the 3D-CNN to represent important neuronal activations across the whole brain via down-sampling following 3D convolution operations may have led to the robust classification results from the minimally preprocessed and raw fMRI 3D volume data in offline analysis and from the real-time fMRI 3D volume data in online classification. The slightly lower classification error rates for our four sensorimotor tasks using the 1D-fcDNN model that incorporated a cross-entropy cost function rather than an MSE cost function (Table S1) is in line with a previous study (Golik et al., 2013). In that study, the multi-layer neural networks trained using MSE appeared to reach a suboptimal local minimum where the gradient vanished, unlike neural networks trained using cross-entropy. By employing the baseline machine learning classifiers (i.e., SVM, LDA, RF, and LR), our classification results suggested that a Gaussian smoothing kernel size of 8 mm was best suited to reduce potential non-neuronal artifacts in our sensorimotor dataset (Fig. 9(c)), which is in line with a previous report (Mikl et al., 2008). However, we believe that this smoothing kernel size may not necessarily be optimal for other fMRI datasets. Thus, the smoothing kernel size can be set as a hyper-parameter in classifiers in order to optimize the classification performance for a specific dataset.

4.3. Splitting data into training and testing sets using k-fold cross-validation

We evaluated several scenarios for dividing the 12 subjects into training and testing groups for the classifier models via *k*-fold CV. Overall, the average classification performance was better with LOOCV (i.e., 12-fold) than with *k*-fold CV (i.e., $k = 3$, 4, and 6) for both fully preprocessed and minimally preprocessed data (Table 1). It was also notable that the standard deviation (STD) of the classification accuracy was overall greater for LOOCV than for *k*-fold CV. These results are consistent with a previous study that has reported that classification performance using LOOCV is substantially more variable than when using ordinary *k*-fold CV (e.g., 5-fold) due to the overly positive estimation of classification performance using the former approach (Varoquaux et al., 2017). Thus, it was recommended to reserve at least 20% of the data for testing and use only 80% for training and validation (Varoquaux et al., 2017). Using our dataset of 12 subjects, LOOCV used data from 11 subjects for training and validation (91.7%) and the data from the one remaining subject for testing (8.3%; cf., 83.3% of data for training and validation and 16.7% for testing with 6-fold CV; 75.0% for training and validation and 25.0% for testing with 4-fold CV; and 66.7% for training and validation and 33.3% for testing with 3-fold CV). Thus, the performance of the classifiers was possibly less vulnerable to over-estimation with 4-fold and 3-fold CV than with LOOCV. Overall, the SE of the error rates for the *k*-fold schemes, including LOOCV, was greater for the 1D-fcDNN and SVM models than for the 3D-CNN model (Table 1). This may also suggest that the 3D-CNN is a robust classifier and better suited to fMRI volume classification because it minimizes potential bias in the adopted training and testing data.

4.4. Feature representation using the 3D-CNN and 1D-fcDNN models

Based on our interpretation of the feature maps from the 3D-CNN models and the weight feature maps from the 1D-fcDNN, richer information was observed in the feature maps of the 3D-CNN model, with slightly shifted patterns in local areas, compared to the mostly overlapping weight feature maps from the 1D-fcDNN model (Fig. 6). The feature maps from the LH and RH tasks produced patterns in the visual area, possibly because the subjects were instructed how to perform the motor task via text instructions on MR-compatible goggles. The extracted weight features in the first hidden layer of the 1D-fcDNN model

parcellated the whole brain into multiple regions that included areas potentially unrelated to the specific target task (Jang et al., 2017). These parcellated regions were pruned in the hidden layers until the resulting features were highly task-specific in the output layer (Jang et al., 2017). In contrast, the extracted feature maps of the Conv1 layer of the 3D-CNN represented brain regions that are important for learning the target concept, which in the present study was the classification of the four tasks (Fig. 5). These feature maps appear to merge together in the Conv layers, while the dimensions of the feature maps were reduced due to the stride during the convolution process by compressing task-related activations. The feature maps of the Conv layers of the 3D-CNN model trained using *k*-fold CV (i.e., k = 3, 4, and 6) were similar but noisier across the whole brain compared to the feature maps using LOOCV (data not shown). This may be due to the reduced number of training samples used with *k*-fold CV compared to LOOCV. The fact that our 3D-CNN model extracted markedly task-specific information from the 3D whole-brain volumes to learn the target concept in classifying the four sensorimotor tasks was also clearly shown in both (a) the class saliency maps, which represented the task-specific regions in the input 3D fMRI volumes used to maximize the class score at the output layer, and (b) the class activation maps, which were obtained from the linear combination of highly task-specific 3D feature maps at the last Conv layer (Fig. 8).

4.5. Systematic evaluation of performance depending on the architecture of the 3D-CNN model

In our study, a systematic evaluation of the performance of the 3D-CNN model for single fMRI volume classification was conducted depending on whether stride was applied during the convolution operation and/or whether a pooling layer was used (Table 2). The importance of the Conv layer and Fc layer was also evaluated. From the consequent sensorimotor fMRI volume classification results, it was found that (a) down-sampling followed by a convolutional layer was crucial to enhancing the classification performance, (b) the various down-sampling operations (i.e., stride or average/max pooling) did not substantially affect the performance, though max-pooling-based down-sampling appeared to be slightly better than average pooling or stride-based down-sampling, and (c) the convolution-operation-based filtering of activation patterns was important to enhancing classification performance. Nonetheless, only down-sampling operations using average/max pooling led to error rates of less than 10%, indicating that our adopted sensorimotor fMRI volume data consisted of relatively distinct patterns of activation for each of the four classes. The fact that the error rates were only slightly lower when applying the Fc layer also indicates that our adopted sensorimotor tasks were not too complicated to classify and/or that the extracted features across the Conv layers were sufficiently distinct. Otherwise, classification performance would have been more clearly enhanced with the use of a non-linear transformation layer (i.e., the Fc layer).

We adopted a stride of 2 during the convolution operation without a following pooling layer and the application of stride had a clear advantage compared to the application of a pooling layer in the context of computational time and resources while maintaining a suitable performance level. Recent studies have suggested using a large stride without a pooling layer in favor of CNN architecture consisting of only repeated Conv layers (Pu et al., 2016; Radford et al., 2015; Springenberg et al., 2014). For example, Springenberg et al. (2014) reported that a Conv layer followed by max-pooling can simply be replaced with a Conv layer with a greater stride size without degrading performance as evaluated using several image recognition benchmark tests (Springenberg et al., 2014). In particular, the removal of pooling layers is important for training generative models such as variational autoencoders (VAEs) or generative adversarial networks (GANs) (Pu et al., 2016; Radford et al., 2015).

4.6. Classification based on parcellated ROIs defined from atlases

Our four adopted sensorimotor tasks appeared relatively straightforward to classify compared to the tasks/conditions in the HCP dataset. Nonetheless, the ROI-based input patterns (i.e., mean activations within ROIs) did not enhance classification performance compared to the whole-brain voxel-level input patterns for the 3D-CNN, 1D-fcDNN, and SVM models. The performance for the AD task was particularly low for all three classifier models and all four atlases (Fig. S2(a)). This could be because the primary auditory cortex is a small area (i.e., Heschl's gyrus), meaning that the corresponding activations are measured with a small number of voxels (Warrier et al., 2009; Yousry et al., 1997); a misalignment of these voxels across fMRI volumes may prevent the correct classification of the AD task. For instance, an example volume for the AD task was misclassified as the VS task after applying parcellated ROI-based classification from the AAL 116, Shen 268, and HCP 360 atlases (Fig. S2(b)) in all three classifier models. Moreover, not all of the voxels in the highly task-specific ROIs for our sensorimotor tasks (e.g., the primary motor areas and primary somatosensory areas for the LH and RH tasks; the primary visual area for the VS task) demonstrated higher BOLD intensities. For instance, the hand area is a part of the primary motor/somatosensory region; consequently, the mean BOLD intensity of the corresponding ROIs was blurred in comparison to the voxel-wise BOLD intensity.

It has also been reported that the variability of BOLD intensity levels and spatial layouts across regions, runs, and subjects is possibly due to variability in physiological hemodynamic responses (Aguirre et al., 1998; Handwerker et al., 2004). As a result, the average BOLD intensity in an ROI would represent a lower BOLD intensity level compared to the BOLD intensities of highly task-specific voxels in the corresponding ROI, which may have degraded the classification performance using the atlas-based approach in comparison to the whole-brain voxel-level approach. Thus, we believe that the whole-brain voxel-level approach is advantageous in terms of enhancing classification performance in multi-task decoding unless fine-grained functional and/or probabilistic atlases that accommodate a variety of tasks are available (Craddock et al., 2012; Poldrack, 2007). On the other hand, the number of parameters and computational time could be drastically reduced with the use of an ROI-based approach compared to a whole-brain voxel-level approach.

4.7. Application of the 3D-CNN model to the online classification of fMRI volumes via real-time fMRI

The results from the online classification of fMRI volumes using the rtfMRI method were particularly interesting in several respects. First, DNNs were applied for the first time in the present study to the online classification of whole-brain fMRI volumes in a real-time fMRI framework. The test phase of our 3D-CNN model for a single fMRI volume was completed significantly faster than the TR (i.e., < 1 s) once the model had been trained offline. We demonstrated the feasibility of the 3D-CNN model for use in neuroscientific and neurobiological applications based on its superior classification capability in a real-time task-classification setup and visualized relevant brain networks using the AMs of the trained 3D-CNN (Fig. 10b). The potential applications of our 3D-CNN in real-time fMRI settings include the real-time decoding of brain states (LaConte, 2011) and consequent real-time fMRI based neurofeedback from decoded brain states (Kim et al., 2019b; Lee et al., 2012; Linden and Turner, 2016; Sitaram et al., 2017; Stoeckel et al., 2014; Sulzer et al., 2013; Thibault et al., 2018; Weiskopf, 2012), while it also holds great promise for a broad range of basic neuroscientific and (pre-)clinical applications (Kim and Birbaumer, 2014; Linden and Turner, 2016; Ruiz et al., 2014; Sitaram et al., 2017; Stoeckel et al., 2014; Sulzer et al., 2013; Thibault et al., 2018; Weiskopf, 2012), including brain–machine interfaces (Goebel et al., 2010; Lee et al., 2009b; Weiskopf et al., 2004; Yoo et al., 2004). It is worth noting that the fMRI volumes used to train the classifier models were acquired from the 12

subjects approximately seven years earlier using an MRI scanner that was around a year old. The fMRI volumes acquired during this online classification experiment were thus preprocessed using spatial smoothing and linear detrending to remove thermal noise and drift artifacts, respectively. The resulting classification performance was higher for the 3D-CNN model (an average accuracy of 78.5% for five-class classification) than for the 1D-fcDNN (26.7% ± 5.9) and SVM (21.5% ± 3.1) models.

One of the main contributions of our study is that we have proposed a 3D-CNN model for the classification of 3D fMRI volumes without using decomposition to generate input samples (Zhao et al., 2017b; Zhao et al., 2018), thus our model is readily available for complex or noisy whole-brain 3D volume data such as raw fMRI data in real-time. Based on our findings as a proof-of-concept, we believe that more complicated and insightful neuroscientific research can be conducted in future studies by utilizing and modifying our implementation and trained model, with the sample data publicly available (https://github.com/bsplku/3dcnn4fmri).

4.8. Potential weaknesses and strengths of our 3D-CNN model

Fine-tuning the hyperparameters, such as the number of Conv layers (along with the size and number of filters for each Conv layer) and Fc layers (along with the number of hidden nodes for each Fc layer) in the nested *k*-fold CV framework (Varoquaux et al., 2017), may further enhance the classification performance and the feature representation capacity of the 3D-CNN model. The number of weight parameters in our 3D-CNN (301,220) is significantly smaller than the number of weight parameters in the adopted 1D-fcDNN (6,842,604). Thus, the 3D-CNN model is inherently less vulnerable to potential overfitting than the 1D-fcDNN, though the sparsity level of the weight parameters in the 1D-fcDNN model can be systematically optimized to reduce overfitting (Jang et al., 2017; Kim et al., 2019a; Kim et al., 2016). We used the dropout regularization scheme for the Conv3 and Fc layers. Dropout regularization for the remaining Conv layers (Srivastava et al., 2014), L1/L2 regularization (Zou and Hastie, 2005) for the Fc layers, and/or batch normalization (Ioffe and Szegedy, 2015) may further enhance the performance of our 3D-CNN model.

4.9. Related works

In a more recent study, a recurrent neural network (RNN) with long short-term memory (LSTM) achieved remarkable advances in distinguishing brain states during the different task events available in the HCP dataset (Li and Fan, 2019). In order to extract the functional features for brain state decoding while reducing the dimensionality of 3D whole-brain activations to generate patterns with less spatial variation than the whole-brain 3D fMRI volumes, intrinsic functional networks (FNs) have been obtained by applying a collaborative sparse brain decomposition method to the resting-state fMRI runs in HCP data (Li et al., 2017). As a result, informative FNs for brain-decoding models have been identified from brain activation patterns for working memory tasks in the HCP data, and the LSTM-based RNN model substantially improved brain decoding performance by adaptively capturing temporal dependency within functional time series data (Li and Fan, 2018, 2019). There have also been recent efforts to construct advanced 3D-CNN architectures for the analysis of brain networks (Kawahara et al., 2017; Wachinger et al., 2018). For example, BrainNetCNN is composed of novel edge-to-edge, edge-to-node, and node-to-graph Conv filters that utilize the topological locality of structural brain connectivity networks from diffusion tensor images; it offers superior predictions of neurodevelopmental scores than a fully connected neural network (Kawahara et al., 2017). In another study, DeepNAT, a 3D patch-based CNN model, was proposed for the automatic segmentation of brain tissue such as gray matter and white matter from T1-weighted MRI (Wachinger et al., 2018). We believe that it is possible that these novel CNN architectures and/or 3D patch-based sampling and analysis can be applied to 3D fMRI volumes in future research.

5. Conclusion

To the best of our knowledge, this is the first study that has reported the classification of a single fMRI volume using a 3D-CNN model without using decomposition to generate input samples. Using a sensorimotor dataset for four tasks and a public HCP dataset, our proposed 3D-CNN model demonstrated superior performance compared to the 1D-fcDNN and SVM models. This is due to the ability of the 3D-CNN model to extract task-relevant information using 3D convolution filters and to handle shifted and scaled neuronal activations using down-sampling operations via stride and/or pooling operations, as illustrated by the 3D convolution filters and corresponding feature maps. Thus, this increases its classification accuracy when using both minimally processed fMRI volumes without spatial normalization and fMRI volumes acquired from real-time fMRI for online classification. This is an important finding because minimally preprocessed fMRI volumes and fMRI volumes acquired in real-time can be seriously affected by the substantial misalignment between fMRI volumes, confounded by the spatial variability of activation patterns across sessions and/or subjects. We believe that the proposed 3D-CNN model can be applied to data acquired from a range of neurocognitive tasks (in addition to the sensorimotor and HCP datasets analyzed in the present study) for the classification of target classes and for regression in the attempt to capture a target concept such as age (Cole et al., 2017; Levi and Hassner, 2015) and emotion levels (Kim et al., 2019a; Kragel and LaBar, 2016; Sitaram et al., 2011).

Author statement

H. Vu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing-Original Draft, Writing-Review & Editing, Visualization

H.C. Kim: Conceptualization, Investigation, Writing-Original Draft, Writing-Review & Editing

M. Jung: Software, Validation, Formal analysis, Data Curation, Writing-Review & Editing

J. H. Lee: Conceptualization, Methodology, Software, Writing-Original Draft, Writing-Review & Editing, Supervision, Project administration, Funding acquisition

Declaration of Competing Interest

The authors have no conflicts of interest regarding this study, including financial, consultant, institutional, or other relationships. The sponsors had no involvement in the study design, data collection, analysis or interpretation of the data, manuscript preparation, or the decision to submit for publication.

Funding source

This work was supported by a National Research Foundation (NRF) grant from the Ministry of Science, ICT and Future Planning of Korea (NRF-2017R1E1A1A01077288).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2020.117328.

References

Agosta, F., Valsasina, P., Absinta, M., Riva, N., Sala, S., Prelle, A., Copetti, M., Comola, M., Comi, G., Filippi, M., 2011. Sensorimotor functional connectivity changes in amyotrophic lateral sclerosis. Cereb. Cortex 21, 2291–2298.

Aguirre, G.K., Zarahn, E., D'esposito, M., 1998. The variability of human, BOLD hemodynamic responses. NeuroImage 8, 360–369.

Barch, D.M., Burgess, G.C., Harms, M.P., Petersen, S.E., Schlaggar, B.L., Corbetta, M., Glasser, M.F., Curtiss, S., Dixit, S., Feldt, C., 2013. Function in the human connectome: task-fMRI and individual differences in behavior. Neuroimage 80, 169–189.

Bell, P.T., Shine, J.M., 2015. Estimating large-scale network convergence in the human functional connectome. Brain Connect. 5, 565–574.

Biswal, B., Zerrin Yetkin, F., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn. Resonanc. Med. 34, 537–541.

Bressler, S.L., Menon, V., 2010. Large-scale brain networks in cognition: emerging methods and principles. Trends Cogn. Sci. 14, 277–290.

Buckner, R.L., 2012. The serendipitous discovery of the brain's default network. Neuroimage 62, 1137–1145.

Calhoun, V.D., Wager, T.D., Krishnan, A., Rosch, K.S., Seymour, K.E., Nebel, M.B., Mostofsky, S.H., Nyalakanai, P., Kiehl, K., 2017. The impact of T1 versus EPI spatial normalization templates for fMRI data analyses. Hum. Brain Mapp. 38, 5331–5342.

Chenji, S., Jha, S., Lee, D., Brown, M., Seres, P., Mah, D., Kalra, S., 2016. Investigating default mode and sensorimotor network connectivity in amyotrophic lateral sclerosis. PLoS One 11.

Cole, J.H., Poudel, R.P., Tsagkrasoulis, D., Caan, M.W., Steves, C., Spector, T.D., Montana, G., 2017. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage 163, 115–124.

Craddock, R.C., James, G.A., Holtzheimer III, P.E., Hu, X.P., Mayberg, H.S., 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. Hum. Brain Mapp. 33, 1914–1928.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.

Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10, 1895–1923.

Dohmatob, E., Varoquaux, G., Thirion, B., 2018. Inter-subject registration of functional images: do we need anatomical images? Front. Neurosci. 12, 64.

Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., Mok, V.C., Shi, L., Heng, P., 2016. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. J. Mag. 35, 1182–1195.

Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. Proc. Natl. Acad. Sci. 201602413.

Fox, M.D., Corbetta, M., Snyder, A.Z., Vincent, J.L., Raichle, M.E., 2006. Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. Proc. Natl. Acad. Sci. 103, 10046–10051.

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171.

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., 2013. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage 80, 105–124.

Goebel, R., Zilverstand, A., Sorger, B., 2010. Real-time fMRI-based brain–computer interfacing for neurofeedback therapy and compensation of lost motor functions. Imaging in Medicine 2 (4), 407–415.

Golik, P., Doetsch, P., Ney, H., 2013. Cross-entropy vs. squared error training: a theoretical and experimental comparison. Interspeech 1756–1760.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Gordon, E.M., Laumann, T.O., Adeyemo, B., Huckins, J.F., Kelley, W.M., Petersen, S.E., 2014. Generation and evaluation of a cortical area parcellation from resting-state correlations. Cereb. Cortex 26, 288–303.

Güçlü, U., van Gerven, M.A., 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. 35, 10005–10014.

Handwerker, D.A., Ollinger, J.M., D'Esposito, M., 2004. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. Neuroimage 21, 1639–1651.

Hazlett, H.C., Gu, H., Munsell, B.C., Kim, S.H., Styner, M., Wolff, J.J., Elison, J.T., Swanson, M.R., Zhu, H., Botteron, K.N., Collins, D.L., Constantino, J.N., Dager, S.R., Estes, A.M., Evans, A.C., Fonov, V.S., Gerig, G., Kostopoulos, P., McKinstry, R.C., Pandey, J., Paterson, S., Pruett, J.R., Schultz, R.T., Shaw, D.W., Zwaigenbaum, L., Piven, J., 2017. Early brain development in infants at high risk for autism spectrum disorder. Nature 542, 348–351.

Horikawa, T., Kamitani, Y., 2017. Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features. Front. Comput. Neurosci. 11 (4).

Huang, H., Hu, X., Zhao, Y., Makkie, M., Dong, Q., Zhao, S., Guo, L., Liu, T., 2018. Modeling task fMRI data via deep convolutional autoencoder. IEEE Trans. Med. Imaging 37.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Jain, S., Huth, A., 2018. Incorporating context into language encoding models for fMRI. bioRxiv, 327601.

Jang, H., Plis, S.M., Calhoun, V.D., Lee, J.-H., 2017. Task-specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: Evaluation using sensorimotor tasks. NeuroImage 145, 314–328.

Kanazawa, A., Sharma, A., Jacobs, D., 2014. Locally scale-invariant convolutional neural networks. arXiv preprint arXiv:1412.5104.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732.

Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G., 2017. BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage 146, 1038–1049.

Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S.V., McDermott, J.H., 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 98, e616 630–644.

Khaligh-Razavi, S.-M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. 10, e1003915.

Kim, D.-Y., Yoo, S.-S., Tegethoff, M., Meinlschmidt, G., Lee, J.-H., 2015. The inclusion of functional connectivity information into fMRI-based neurofeedback improves its efficacy in the reduction of cigarette cravings. J. Cogn. Neurosci. 27, 1552–1572.

Kim, H.-C., Bandettini, P.A., Lee, J.-H., 2019a. Deep neural network predicts emotional responses of the human brain from functional magnetic resonance imaging. NeuroImage 186, 607–627.

Kim, H.-C., Tegethoff, M., Meinlschmidt, G., Stalujanis, E., Belardi, A., Jo, S., Lee, J., Kim, D.-Y., Yoo, S.-S., Lee, J.-H., 2019b. Mediation analysis of triple networks revealed functional feature of mindfulness from real-time fMRI neurofeedback. NeuroImage 195, 409–432.

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.

Kim, S., Birbaumer, N., 2014. Real-time functional MRI neurofeedback: a tool for psychiatry. Curr. Opin. Psychiatry 27, 332–336.

Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., Biller, A., 2016. Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. Neuroimage 129, 460–469.

Kragel, P.A., LaBar, K.S., 2016. Decoding the nature of emotion in the brain. Trends Cogn. Sci. 20, 444–455.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 1097–1105.

LaConte, S.M., 2011. Decoding fMRI brain states in real-time. NeuroImage 56, 440–454.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.

Lee, J.-H., Kim, J., Yoo, S.-S., 2012. Real-time fMRI-based neurofeedback reinforces causality of attention networks. Neurosci. Res. 72, 347–354.

Lee, J.-H., Marzelli, M., Jolesz, F.A., Yoo, S.-S., 2009a. Automated classification of fMRI data employing trial-based imagery tasks. Med. Image Anal. 13, 392–404.

Lee, J.-H., Ryu, J., Jolesz, F.A., Cho, Z.-H., Yoo, S.-S., 2009b. Brain–machine interface via real-time fMRI: preliminary study on thought-controlled robotic arm. Neurosci. Lett. 450, 1–6.

Levi, G., Hassner, T., 2015. Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 34–42.

Li, H., Fan, Y., 2018. Brain decoding from functional MRI using long short-term memory recurrent neural networks. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 320–328.

Li, H., Fan, Y., 2019. Interpretable, highly accurate brain decoding of subtly distinct brain states from functional MRI using intrinsic functional networks and long short-term memory recurrent neural networks. Neuroimage 202, 116059.

Li, H., Satterthwaite, T.D., Fan, Y., 2017. Large-scale sparse functional networks from resting state fMRI. Neuroimage 156, 1–13.

Linden, D.E., Turner, D.L., 2016. Real-time functional magnetic resonance imaging neurofeedback in motor neurorehabilitation. Curr. Opin. Neurol. 29, 412.

Luo, W., Li, Y., Urtasun, R., Zemel, R., 2016. Understanding the effective receptive field in deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 4898–4906.

Manly, B.F., 2018. Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapman and Hall/CRC.

Maturana, D., Scherer, S., 2015. Voxnet: a 3d convolutional neural network for real-time object recognition. In: Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 922–928.

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153–157.

Menon, V., 2015. Salience Network: Brain Mapping: An Encyclopedic Reference. Elsevier Inc.

Mikl, M., Mareček, R., Hluštík, P., Pavlicová, M., Drastich, A., Chlebus, P., Brázdil, M., Krupa, P., 2008. Effects of spatial smoothing on fMRI group inferences. Magn. Resonanc. Imaging 26, 490–503.

Nie, D., Zhang, H., Adeli, E., Liu, L., Shen, D., 2016. 3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 212–220.

Norouzi, M., Ranjbar, M., Mori, G., 2009. Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE, pp. 2735–2742.

Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015. Deep Face Recognition. BMVC, p. 6.

Pinto, N., Cox, D.D., DiCarlo, J., 2008. Why is real-world visual object recognition hard? PLoS Comput. Biol. 4, e27.

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.A., Calhoun, V.D., 2014. Deep learning for neuroimaging: a validation study. Front. Neurosci. 8, 229.

Poldrack, R.A., 2007. Region of interest analysis for fMRI. Soc. Cogn. Affect. Neurosci. 2, 67–70.

Power, J.D., Cohen, A.L., Nelson, S.M., Wig, G.S., Barnes, K.A., Church, J.A., Vogel, A.C., Laumann, T.O., Miezin, F.M., Schlaggar, B.L., 2011. Functional network organization of the human brain. Neuron 72, 665–678.

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., Carin, L., 2016. Variational autoencoder for deep learning of images, labels and captions. Adv. Neural Inf. Process. Syst. 2352–2360.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Ramasangu, H., Sinha, N., 2014. Cognitive state classification using transformed fMRI data. In: Proceeding of the 2014 International Conference on Signal Processing and Communications (SPCOM). IEEE, pp. 1–5.

Raschka, S., 2018. Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808.

Ruiz, S., Buyukturkoglu, K., Rana, M., Birbaumer, N., Sitaram, R., 2014. Real-time fMRI brain computer interfaces: self-regulation of single brain regions to networks. Biol. Psychol. 95, 4–20.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823.

Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S., 2014. CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 806–813.

Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248.

Shen, X., Tokoglu, F., Papademetris, X., Constable, R.T., 2013. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. Neuroimage 82, 403–415.

Shirer, W.R., Ryali, S., Rykhlevskaia, E., Menon, V., Greicius, M.D., 2012. Decoding subject-driven cognitive states with whole-brain connectivity patterns. Cereb. Cortex 22, 158–165.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: visualizing image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Simonyan, K., Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. Adv. Neural Inf. Process. Syst. 568–576.

Sitaram, R., Lee, S., Ruiz, S., Rana, M., Veit, R., Birbaumer, N., 2011. Real-time support vector classification and feedback of multiple emotional brain states. Neuroimage 56, 753–765.

Sitaram, R., Ros, T., Stoeckel, L., Haller, S., Scharnowski, F., Lewis-Peacock, J., Weiskopf, N., Blefari, M.L., Rana, M., Oblak, E., 2017. Closed-loop brain training: the science of neurofeedback. Nat. Rev. Neurosci. 18, 86.

Song, S., Zhan, Z., Long, Z., Zhang, J., Yao, L., 2011. Comparative study of SVM methods combined with voxel selection for object category classification on fMRI data. PloS one 6, e17191.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2014. Striving for simplicity: the all convolutional net. arXiv preprint arXiv:1412.6806.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Steimke, R., Nomi, J.S., Calhoun, V.D., Stelzel, C., Paschke, L.M., Gaschler, R., Goschke, T., Walter, H., Uddin, L.Q., 2017. Salience network dynamics underlying successful resistance of temptation. Soc. Cogn. Affect. Neurosci. 12, 1928–1939.

Stoeckel, L.E., Garrison, K.A., Ghosh, S.S., Wighton, P., Hanlon, C.A., Gilman, J.M., Greer, S., Turk-Browne, N.B., deBettencourt, M.T., Scheinost, D., 2014. Optimizing real time fMRI neurofeedback for therapeutic discovery and development. NeuroImage: Clin. 5, 245–255.

Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A.s.D.N., 2014. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage 101, 569–582.

Sulzer, J., Haller, S., Scharnowski, F., Weiskopf, N., Birbaumer, N., Blefari, M.L., Bruehl, A.B., Cohen, L.G., DeCharms, R.C., Gassert, R., 2013. Real-time fMRI neurofeedback: progress and challenges. NeuroImage 76, 386–399.

Tavor, I., Jones, O.P., Mars, R., Smith, S., Behrens, T., Jbabdi, S., 2016. Task-free MRI predicts individual differences in brain activity during task performance. Science 352, 216–220.

Thibault, R.T., MacPherson, A., Lifshitz, M., Roth, R.R., Raz, A., 2018. Neurofeedback with fMRI: a critical systematic review. Neuroimage 172, 786–807.

Thomas Yeo, B., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 106, 1125–1165.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15, 273–289.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.-M.H., 2013. The WU-Minn human connectome project: an overview. Neuroimage 80, 62–79.

Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., 2012. The Human Connectome Project: a data acquisition perspective. Neuroimage 62, 2222–2231.

Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. NeuroImage 145, 166–179.

Vossel, S., Geng, J.J., Fink, G.R., 2014. Dorsal and ventral attention systems: distinct neural circuits but collaborative roles. Neuroscientist 20, 150–159.

Wachinger, C., Reuter, M., Klein, T., 2018. DeepNAT: deep convolutional neural network for segmenting neuroanatomy. NeuroImage 170, 434–445.

Wang, X., Liang, X., Zhou, Y., Wang, Y., Cui, J., Wang, H., Li, Y., Nguchu, B.A., Qiu, B., 2018. Task state decoding and mapping of individual four-dimensional fMRI time series using deep neural network. arXiv preprint arXiv:1801.09858.

Warrier, C., Wong, P., Penhune, V., Zatorre, R., Parrish, T., Abrams, D., Kraus, N., 2009. Relating structure to function: Heschl's gyrus and acoustic processing. J. Neurosci. 29, 61–69.

Weiskopf, N., 2012. Real-time fMRI and its application to neurofeedback. NeuroImage 62, 682–692.

Weiskopf, N., Mathiak, K., Bock, S.W., Scharnowski, F., Veit, R., Grodd, W., Goebel, R., Birbaumer, N., 2004. Principles of a brain-computer interface (BCI) based on real–time functional magnetic resonance imaging (fMRI). IEEE Trans. Biomed. Eng. 51, 966–970.

Wen, D., Wei, Z., Zhou, Y., Li, G., Zhang, X., Han, W., 2018. Deep learning methods to process fMRI data and their application in the diagnosis of cognitive impairment: a brief overview and our opinion. Front. Neuroinf. 12, 23.

Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J., 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. 111, 8619–8624.

Yoo, S.-S., Fairneny, T., Chen, N.-K., Choo, S.-E., Panych, L.P., Park, H., Lee, S.-Y., Jolesz, F.A., 2004. Brain–computer interface using fMRI: spatial navigation by thoughts. Neuroreport 15, 1591–1595.

Yousry, T., Fesl, G., Buttner, A., Noachtar, S., Schmid, U., 1997. Heschl's gyrus-Anatomic description and methods of identification on magnetic resonance imaging. Int. J. Neuroradiol. 3, 2–12.

Yuan, R., Di, X., Taylor, P.A., Gohel, S., Tsai, Y.-H., Biswal, B.B., 2016. Functional topography of the thalamocortical system in human. Brain Struct. Funct. 221, 1971–1984.

Zanto, T.P., Gazzaley, A., 2013. Fronto-parietal network: flexible hub of cognitive control. Trends Cogn. Sci. 17, 602–603.

Zhang, C., Yao, L., Song, S., Wen, X., Zhao, X., Long, Z., 2018. Euler elastica regularized logistic regression for whole-brain decoding of fMRI data. J. Mag. 65, 1639–1653.

Zhao, Y., Dong, Q., Chen, H., Iraji, A., Li, Y., Makkie, M., Kou, Z., Liu, T., 2017a. Constructing fine-granularity functional brain network atlases via deep convolutional autoencoder. Med. Image Anal. 42, 200–211.

Zhao, Y., Dong, Q., Zhang, S., Zhang, W., Chen, H., Jiang, X., Guo, L., Hu, X., Han, J., Liu, T., 2017b. Automatic recognition of fMRI-derived functional networks using 3-D convolutional neural networks. IEEE Trans. Biomed. Eng. 65, 1975–1984.

Zhao, Y., Ge, F., Liu, T., 2018. Automatic recognition of holistic functional brain networks using iteratively optimized convolutional neural networks (IO-CNN) with weak label initialization. Med. Image Anal. 47, 111–126.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 67, 301–320.