Received: 18 September 2019 | Revised: 20 November 2019 | Accepted: 27 November 2019

DOI: 10.1002/hbm.24891

RESEARCH ARTICLE

WILEY

Decoding and mapping task states of the human brain via deep learning

Xiaoxiao Wang[1] | Xiao Liang[1] | Zhoufan Jiang[1] | Benedictor A. Nguchu[1] | Yawen Zhou[1] | Yanming Wang[1] | Huijuan Wang[1] | Yu Li[1] | Yuying Zhu[1] | Feng Wu[1] | Jia-Hong Gao[2,3] | Bensheng Qiu[1]

[1]Centers for Biomedical Engineering, University of Science and Technology of China, Hefei, China

[2]MRI Research Center and Beijing City Key Lab for Medical Physics and Engineering, Peking University, Beijing, China

[3]McGovern Institute for Brain Research, Peking University, Beijing, China

Correspondence
Benching Qiu, Centers for Biomedical Engineering, University of Science and Technology of China, Hefei, China.
Email: bqiu@ustc.edu.cn

Jia-Hong Gao, MRI Research Center and Beijing City Key Lab for Medical Physics and Engineering, Peking University, Beijing, China.
Email: jgao@pku.edu.cn

Funding information
Beijing Municipal Science and Technology Commission, Grant/Award Number: Z171100000117012; Fundamental Research Funds for the Central Universities of China, Grant/Award Number: WK2070000033; National Natural Science Foundation of China, Grant/Award Numbers: 31421003, 81430037, 81701665, 81727808, 81790650, 81790651, 91432301

Abstract

Support vector machine (SVM)-based multivariate pattern analysis (MVPA) has delivered promising performance in decoding specific task states based on functional magnetic resonance imaging (fMRI) of the human brain. Conventionally, the SVM-MVPA requires careful feature selection/extraction according to expert knowledge. In this study, we propose a deep neural network (DNN) for directly decoding multiple brain task states from fMRI signals of the brain without any burden for feature handcrafts. We trained and tested the DNN classifier using task fMRI data from the Human Connectome Project's S1200 dataset (*N* = 1,034). In tests to verify its performance, the proposed classification method identified seven tasks with an average accuracy of 93.7%. We also showed the general applicability of the DNN for transfer learning to small datasets (*N* = 43), a situation encountered in typical neuroscience research. The proposed method achieved an average accuracy of 89.0 and 94.7% on a working memory task and a motor classification task, respectively, higher than the accuracy of 69.2 and 68.6% obtained by the SVM-MVPA. A network visualization analysis showed that the DNN automatically detected features from areas of the brain related to each task. Without incurring the burden of handcrafting the features, the proposed deep decoding method can classify brain task states highly accurately, and is a powerful tool for fMRI researchers.

Abbreviations: 2D, two dimensional; 3D, three dimensional; 4D, four dimensional; ABIDE, Autism Brain Imaging Data Exchange; ADNI, Alzheimer's Disease Neuroimaging Initiative; BA, Brodmann area; BN, batch normalization; BOLD, blood-oxygenation-level dependent; CNN, convolutional neural network; COPE, contrast of parameter estimate; DBN, deep belief network; DNN, deep neural network; fMRI, functional magnetic resonance imaging; FN, false negative; FP, false positive; GLM, general linear model; HCP, Human Connectome Project; HRF, hemodynamic response function; LSTM, long short-term memory; M1, primary motor cortex; M2, secondary motor cortex; MNI, Montreal Neurological Institute; MVPA, multi-voxel pattern analysis; RBM, restricted Boltzmann machine; ReLU, rectified linear unit; RNN, recurrent neural network; ROC, receiver operating characteristic; ROI, region of interest; S1, primary somatosensory area; SGD, stochastic gradient descent; SVM, support vector machine; TN, true negative; TP, true positive; WM, working memory.

Xiaoxiao Wang and Xiao Liang contributed equally to this work.

KEYWORDS
brain decoding, deep learning, functional brain mapping, functional magnetic resonance imaging, Human Connectome Project, transfer learning

1 | INTRODUCTION

For years, researchers have been attempting to decode and identify functions of the human brain based on functional brain imaging data (Dehaene et al., 1998; Haynes & Rees, 2006; Jang, Plis, Calhoun, & Lee, 2017; Poldrack, Halchenko, & Hanson, 2009; Rubin et al., 2017). The most popular among these brain-decoding methods is the support vector machine (SVM)-based multi-voxel pattern analysis (MVPA), a supervised technology that incorporates information from multiple variables at the same time (Kim & Oertzen, 2018; Kriegeskorte & Bandettini, 2007; Kriegeskorte, Goebel, & Bandettini, 2006; Norman, Polyn, Detre, & Haxby, 2006). Despite its popularity, the SVM struggles to perform well on high-dimensional raw data, and requires the expert use of design techniques for feature selection/extraction (LeCun, Bengio, & Hinton, 2015; Vieira, Pinaya, & Mechelli, 2017). Thus, we explore in this study an open-ended brain decoder that uses whole-brain neuroimaging data on humans.

In recent years, the deep neural network (DNN), a series of model-free machine learning methods, has performed well in abstracting representations of high-dimensional data (LeCun et al., 2015). The hierarchical structure of a DNN with a nonlinear activation function enables the learning of a more complex output function than those that can be learned using traditional machine learning methods, and one that can be trained end to end. DNNs have already yielded remarkable results in medical image analyses (Cichy & Kaiser, 2019; Shen, Wu, & Suk, 2017; Vieira et al., 2017). Considering these characteristics, a DNN classifier may be suited for classifying brain states directly from a massive whole-brain fMRI time series without requiring feature selection.

Deep learning methods are effective if massive amounts of data are available for training. However, under controlled conditions, most typical neuroimaging studies have collected data from only tens to hundreds of subjects, with the purpose of identifying minor differences between different states (Horikawa & Kamitani, 2017) or groups thereof (Vieira et al., 2017). An applicable brain decoder is supposed to be able to identify these differences even with a limited amount of data. Transfer learning is widely used for training DNNs with limited medical data (Sharif Razavian, Azizpour, Sullivan, & Carlsson, 2014). It takes advantage of similar data within big datasets (Ciompi et al., 2015; Kermany et al., 2018; Wen, Shi, Chen, & Liu, 2018). Recent large fMRI projects, such as the Human Connectome Project (HCP; Van Essen, et al., 2013) and BioBank (Miller et al., 2016), allow us to access massive amounts of fMRI data. It is, therefore, now possible to directly train a DNN decoder by means of big fMRI data and generalize the DNN decoder for common fMRI studies.

In this study, we propose a DNN classifier that effectively decodes and maps an individual's ongoing brain task state by reading 4D fMRI signals related to the task. We illustrate the generalizability of this DNN for typical neuroimaging studies by testing the decoder on the classification of task sub-types.

2 | METHODS

2.1 | HCP datasets

The HCP S1200 minimally preprocessed 3T data release, which contains imaging and behavioral data from a large population of young healthy adults (Van Essen, et al., 2013), was used in this study. We employed data of 1,034 participants of the HCP who had performed seven tasks: emotion, gambling, language, motor, relational, social, and working memory (WM). Further details of the recruitment process, imaging data acquisition, behavior collection, and MRI preprocessing can be found in previous papers (Barch, et al., 2013; Van Essen, et al., 2012; Van Essen, et al., 2013).

2.2 | Preparation of fMRI time series for deep learning

We analyzed the HCP volume-based preprocessed fMRI data, which had already been normalized to the Montreal Neurological Institute's (MNI) 152 space. Most of the seven tasks were constituted by control conditions (e.g., 0-back places in the WM task and shape stimuli in the emotion task) and task conditions (e.g., 2-back in the WM task and fear stimuli in the emotion task). In each task, only one condition was selected for the next step. For tasks (emotion, language, gambling, social, and relational tasks) with only two conditions, the condition that showed a greater association with the task had priority over the other. WM and motor tasks contained more than one task condition, and we randomly chose one (2-back body for WM and right hand for motor) from the list (Table 1).

For each task, an input sample was a continuous BOLD series that covered the entire block and 8 s past the block, including the post-signal of the hemodynamic response function (HRF). Furthermore, each BOLD volume was cropped from $91 \times 109 \times 91$ to $75 \times 93 \times 81$ to exclude the area that was not part of the brain. Thus, the input data varied from $27 \times 75 \times 93 \times 81$ to $50 \times 75 \times 93 \times 81$ ($\text{time} \times x \times y \times z$, TR = 0.72 s). A total of 34,938 fMRI 4D data items were obtained across all tasks and subjects.

TABLE 1 Details of the selected BOLD time series for each task

Task	Candidate conditions	Selected condition	Duration of the block (s)
Emotion	Fear, shape	Fear	18
Gambling	Reward, loss	Loss	28
Language	Story, math	Present story	20
Motor	Right hand, left hand, right foot, left loot, tongue	Right hand	12
Relational	Relational, match	Relational	16
Social	Mental, random	Mental	23
Working memory (WM)	2-back places, 0-back places, 2-back body, 0-back body, 2-back tools, 0-back tools, 2-back faces, 0-back faces	2-back places	27.5

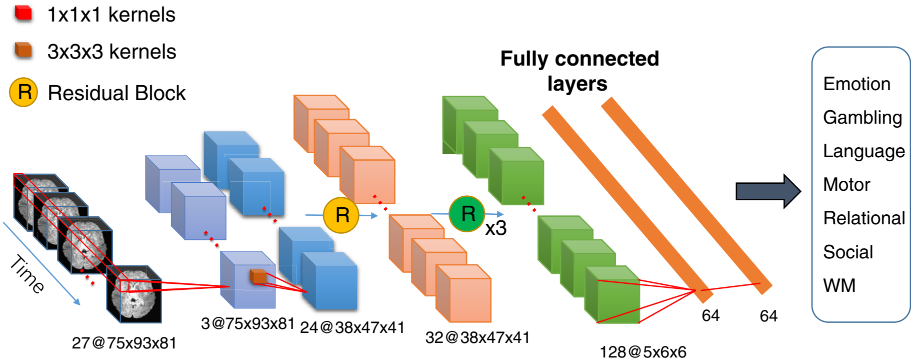

FIGURE 1 The proposed deep neural network. The network consists of five convolutional layers and two fully connected layers. The model takes fMRI scans as input and provides labeled task classes as output

2.3 | The DNN

Figure 1 shows a flow diagram of our proposed network that consists of five convolutional layers and two fully connected layers. In this experiment, $27 \times 75 \times 93 \times 81$ data were generated via the aforementioned preprocessing and data augmentation steps. In the first layer, we used $1 \times 1 \times 1$ convolutional filters, which have been widely used in recent structural designs of convolutional neural networks (CNNs) because these filters increase nonlinearity without changing the receptive fields of the convolutional layer (Hu, Shen, & Sun, 2017; Iandola et al., 2016; Simonyan & Zisserman, 2014). These filters can generate temporal descriptors for each voxel of the volume of the fMRI, and their weights can be easily learnt by DNNs during training. Therefore, after adopting this type of filter, the time dimension of the data was reduced from 27 to 3. Following this, a convolutional layer and four residual blocks were stacked to extract the high-level features. Our residual block is formed by replacing the 2D convolutional layer in the original residual block (He, Zhang, Ren, & Sun, 2016) with a 3D convolutional layer (Maturana & Scherer, 2015). The output channels of the four residual blocks are in *multiples of two*—32, 64, 64, and 128, respectively. We adopted a stride of two in the second convolutional layer and the last three residual blocks. These layers were designed in such a way that their dimensions could be quickly reduced to balance the consumption of GPU memory. For ease of network visualization analysis, we used a full convolution in the last convolutional layer instead of the pooling operation in CNNs used in common. Two fully connected layers were used after a stack of convolutional layers; the first had 64 channels and the second performed seven-way classification (one for each class). In our models, the rectified linear unit (ReLU) function (Krizhevsky, Sutskever, & Hinton, 2012) and batch normalization (BN) layer (Ioffe & Szegedy, 2015) were applied after each convolutional layer, whereas the softmax function was employed in the last fully connected layer.

Big data played an important role in training the DNNs. Despite the remarkable success of DNNs, their application to a limited amount of data is still a problem. Data augmentation is an efficient way to generate more samples, and has been widely used in applications (Ciompi et al., 2015; Donahue et al., 2014; Wachinger, Reuter, & Klein, 2018). The main purpose of data augmentation is to increase variations in the data where this can prevent overfitting and improve the invariance of the neural network. Contrary to traditional images, the input images in this experiment were already aligned with the standard MNI152 template; therefore, performing data augmentation in the spatial domain was considered redundant. Considering the varied durations of the input data, we applied data augmentation in the temporal domain to improve the generalizability of the neural networks in this situation. A fragment of k continuous TRs (k = 27 in our experiments) was randomly split from each input data item in every epoch of the training stage (Figure 2a). To avoid fluctuations in the reported accuracy, only the fragment consisting of the first k TRs of each data was used in validation and testing stages.

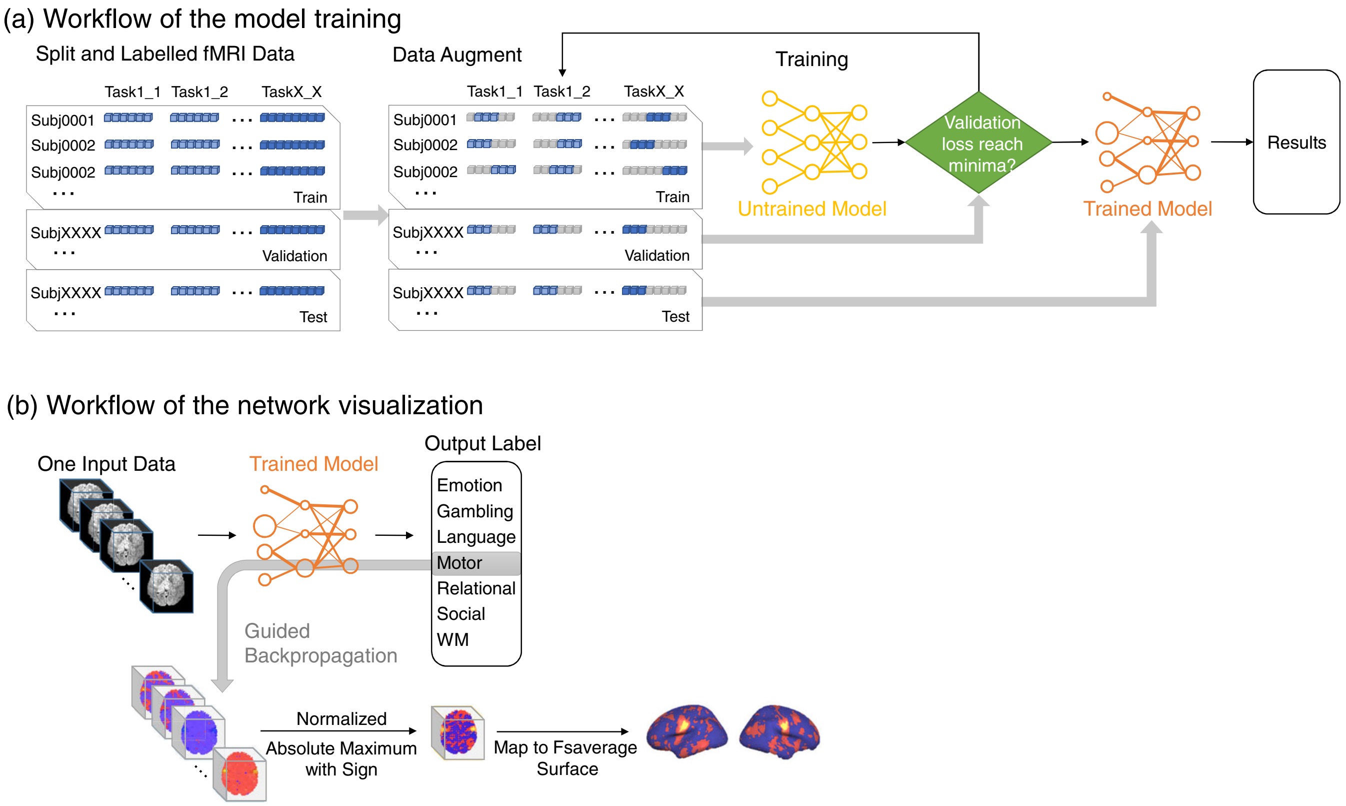

FIGURE 2 Workflows of model training and network visualization. (a) The proposed model automatically learns features of the labeled fMRI time series and stops training when the loss of validation reaches a minimum. Thus, no feature handcrafting is required for model training. The workflow of transfer learning is similar, except that the untrained model is replaced by the trained model. (b) The classification of each data item is back propagated to the network layers to obtain a visualization of parts important to the classification. The visualized data, which have the same size as the input data, are then reduced in the time dimension and mapped into the fsaverage surface. A motor task data is chosen for the illustration

The implementation of our proposed network was based on the PyTorch framework (https://github.com/pytorch/pytorch). The design was constructed from scratch but initially utilized weights suggested by He, Zhang, Ren, and Sun (2015). To guarantee effectiveness, we used Adam with the standard parameters (β_1 = 0.9 and β_2 = 0.999) (Kingma & Ba, 2014). Due to memory constraints on the graphics board, the batch size was set to 32. The initial learning rate was set to 0.001, and gradually decayed by a factor of 10 each time the validation loss plateaued after 15 epochs. To avoid overfitting, we used the early stopping approach, and stopped training when the validation loss reached a minimum.

Our validation strategy employed a fivefold cross-validation across subjects. Prior to training, the subjects' data were categorized into subsets as follows: training set (70%), validating set (10%), and testing set (20%; Figure 2a). The sample of training/validation/testing was later altered for each of fivefolds. Applying the SVM-MVPA to tens of thousands of data items is time consuming. A comparison between the SVM-MVPA and the proposed method was thus not applied to the entire dataset, but to the Test–Retest task-fMRI group data in Section 2.4.

2.4 | Transfer learning

An important advantage of deep learning methods, CNNs in particular, compared with traditional methods, is their reusability, which means that the trained CNN can be directly reused on similar tasks. We used a transfer learning strategy for the trained CNN to validate the general use characteristics of the proposed model. The workflow of transfer training is largely similar to that of the initial training (Figure 2a), except that it starts with a model where the first four layers are trained and the output layer is untrained. We employed the TEST dataset of the TEST–RETEST task-fMRI group from the HCP (N = 43). We trained the deep model to classify two WM task sub-states—0bk-body and 2bk-body. A subject-wise fivefold cross validation was applied with 60% (100 samples of 25 subjects) used for training, 20% (36 samples of nine subjects) for validation, and 20% (36 samples of nine subjects) for testing (172 samples in total are comparable in size to commonly used fMRI research datasets). For further validation, we trained the deep model to classify four motor task sub-states—left foot, left hand, right foot, and tongue movement—using fivefold cross validation with 60% (400 samples of 25 subjects) used for training, 20% (144 samples of nine subjects) for validation, and 20% (144 samples of nine subjects) for testing (688 samples altogether). As in the previous scheme, an input sample was a continuous BOLD series that covered the entire block and 8 s past the block, including the post-signal of the HRF.

For a comparison with the proposed deep learning method, the SVM-MVPA method was also used to analyze the TEST–RETEST dataset using The Decoding Toolbox (Hebart, Gorgen, & Haynes,

2014) in MATLAB (MathWorks, Natick, MA). The run-wise beta images of each subject were obtained through a GLM with separate regressors embedded in the HCP standard FEAT scripts for each task condition. The resulting beta images were then taken as inputs to the SVM-MVPA. A searchlight analysis was also applied: A sphere with a radius of three voxels "searchlight" moved through each brain using a multi-class classification SVM function (fitcecoc, the Statistics and Machine Learning Toolbox of MATLAB) with a linear kernel. The F1 score (see Section "2.6 Assessments") for each condition was calculated as the resulting map. Fivefold cross-validation was also employed. The classifier was trained on data from four-fifths of the subjects and tested on data from the remaining one-fifth.

To evaluate the applicability of the DNN of fMRI studies using small sample sizes, we trained the deep classifiers on data from the 43 subjects of the HCP TEST scans: N = 1, 2, 4, 8, 17, 25, 34. To avoid variance in accuracy, all tests were applied to the RETEST data of all 43 subjects in the HCP Test–Retest dataset. The deep learning was stopped after 120 epochs. Searchlight and whole-brain SVM-MVPA methods were also used for comparison.

2.5 | Performance evaluation

To assess the performance of the model in classifying different tasks, some useful parameters were computed. The $F1$ score was computed for each task condition as a function of the TP, FP, and FN: $F1 = (2 \times TP)/(2 \times TP + FP + FN)$. Here, TP is the true positive, FP is the false positive, and FN is the false negative for each label. The receiver operating characteristic (ROC) curve was also calculated for each label by the one-versus-rest approach, with the parameter sensitivity and specificity denoted by: *sensitivity* = TP/(TP + FN) and *specificity* = TN/(TN + FP), where TN is the true negative equal to the sum of the TPs of the rest of the labels. Accuracy was defined as the ratio of the correct predictions to the total number of classifications: *accuracy* = (TP + TN)/(TP + FP + TN + FN).

2.6 | Network visualization analysis

Guided back-propagation (Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014), a widely used deep network visualization method, was applied to produce pattern maps of each classification and task-weighted representation of the input fMRI 4D time series. During standard back-propagation, the partial derivative of a ReLU unit is copied backward if the input to it is positive, and is otherwise set to zero. In guided-back-propagation, the partial derivative of a ReLU unit is copied backward if both the input to it and the partial derivative are positive. Thus, guided back-propagation maintain paths that have a positive influence on the class score and outputs data features that the CNN detects rather than those it does not. As shown in Figure 2b, after feeding data to the trained networks, 27 × 75 × 93 × 81 prediction gradients were produced with respect to the input data. Then, the signed value with an absolute maximum in the time domain for each voxel was drawn out and built up in a 3D task pattern map, which was then normalized to its maximum value. Finally, the pattern map was mapped into the fsaverage surface. In addition, Cohen's d effect for the normalized pattern maps of the test group was

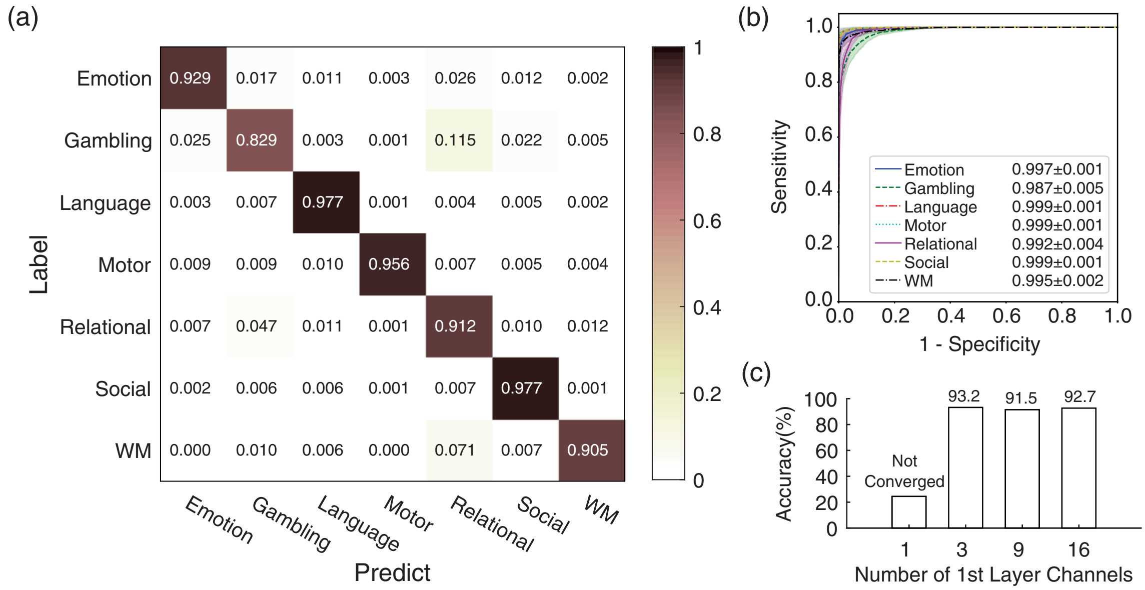

FIGURE 3 Results of deep learning classification on the HCP S1200 task fMRI dataset. (a) The average confusion matrix normalized to the number of labels in the fivefold cross-validation, with the top two confusions caused by gambling vs. relational and relational versus WM. The mean (±*SD*) accuracy of classification on the seven tasks was 93.7% (±1.9%) with a chance level of 14.29%. (b) The mean (solid lines) and *SD* (shadow envelopes) of the ROC curves for each label in the fivefold cross-validation. The legend shows the mean ± *SD* of the AUC of the ROC for the seven tasks. (c) The classification performance (accuracy in %) of the proposed network following various settings of the number of channels in the first layer (N_{Ch1}), which was three in the proposed model. The model failed to converge within 30 epochs when $N_{Ch1} = 1$

calculated as the mean of the pattern maps of each task divided by their *SD* (Cohen, 1998). Analysis was conducted in AFNI (Cox, 1996), Freesurfer (Fischl, 2012), HCP Connectome Workbench (https://www.humanconnectome.org/software/connectome-workbench), and MATLAB (MathWorks, Natick, MA). For a comparison between the traditional GLM map and the pattern map, we also obtained the Cohen's effect of contrast of parameter estimate (COPE) from the fMRI analysis package of the HCP task.

3 | RESULTS

3.1 | The deep model's performance in general task classification

The training session required approximately 72 hr for the 30 epochs with an NVIDIA GTX 1080Ti board, and the proposed model successfully distinguished seven tasks with an accuracy of 93.7 ± 1.9%

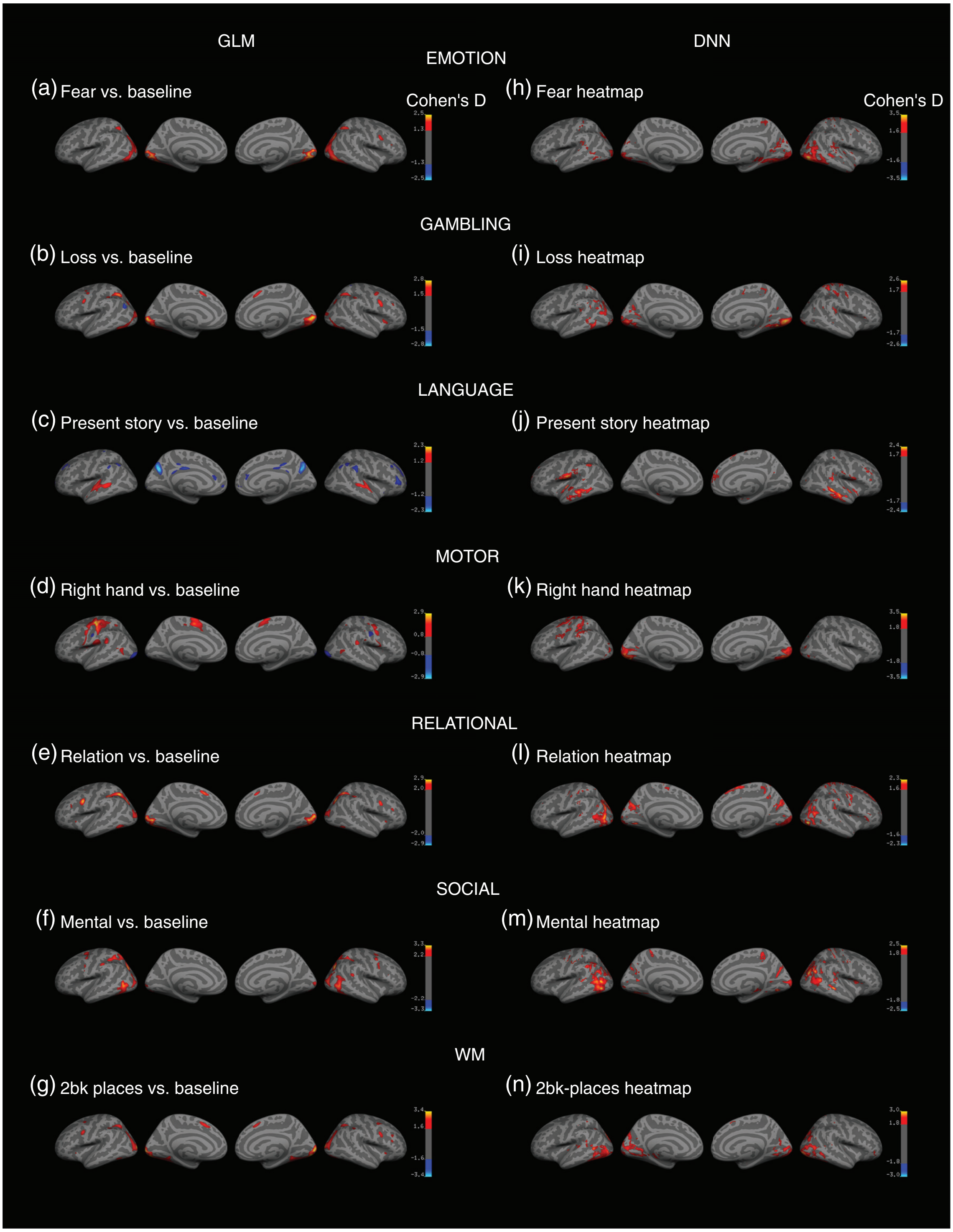

FIGURE 4 Cohen's *d* effect size for the HCP group average (left column) and DNN heatmaps (right column) on the HCP S1200 dataset

(mean ± *SD*). An analysis of F1 scores showed that the classifier performed differently across the seven tasks: emotion (94.0 ± 1.6%), gambling (83.7 ± 4.6%), language (97.6 ± 1.1%), motor (97.3 ± 1.6%), relational (89.8 ± 3.2%), social (96.4 ± 1.0%), and WM (91.9 ± 2.3%, mean ± *SD*). The average confusion matrix showed that the top two confusions were caused by gambling versus relational and WM versus relational (Figure 3a). Figure 3b illustrates the ROC curves, according to which the motor, language, and social task have the largest area under the curve (AUC), while gambling has the smallest.

Upon validation of the choice of key hyper-parameter—the number of 1x1x1 kernel channels (N_{Ch1})−−the model recorded accuracy values of 93.2, 91.5, and 92.7% with N_{Ch1} = 3, 9, and 27, respectively (Figure 3c). With N_{Ch1} = 1, the model could not converge within 30 epochs.

3.2 | Visualization of learnt patterns

To identify the voxels contributing most to each classification, we produced pattern maps by using guided back-propagation (Springenberg et al., 2014). Figure 4 shows group statistical maps of the effect size of Cohen's *d* for the GLM analysis on the task COPE (Figure 4a–g), and the Cohen's *d* on the DNN pattern maps (Figure 4h–n). As shown in the illustrations, the Cohen's *d* on the DNN pattern maps was similar to that on the GLM COPEs for emotion, language, motor, social, and WM tasks. For example, with the language condition, a large effect size was aberrant in the bilateral Brodmann 22 area in the GLM COPEs (Figure 4c) and DNN pattern maps (Figure 4j). In the same fashion, both maps (Figure 4d,k) revealed similar effects in the Brodmann 4 and bilateral Brodmann 18 areas following the right-hand movement condition in the motor task. For further details on annotations, see Table S1.

3.3 | Transfer learning of WM task sub-types on small datasets

Following fivefold cross-validation, the proposed DNN reached an average accuracy of 89.0 ± 2.0% (Figure 5a) and an average AUC of ROC 0.931 ± 0.032 (Figure 5b) in the tests. As shown in Figure 5c, the accuracy of the DNN was significantly higher than that of SVM-MVPA whole-brain (t[8] = 9.14, *p* = .000017; mean ± *SD* = 55.6

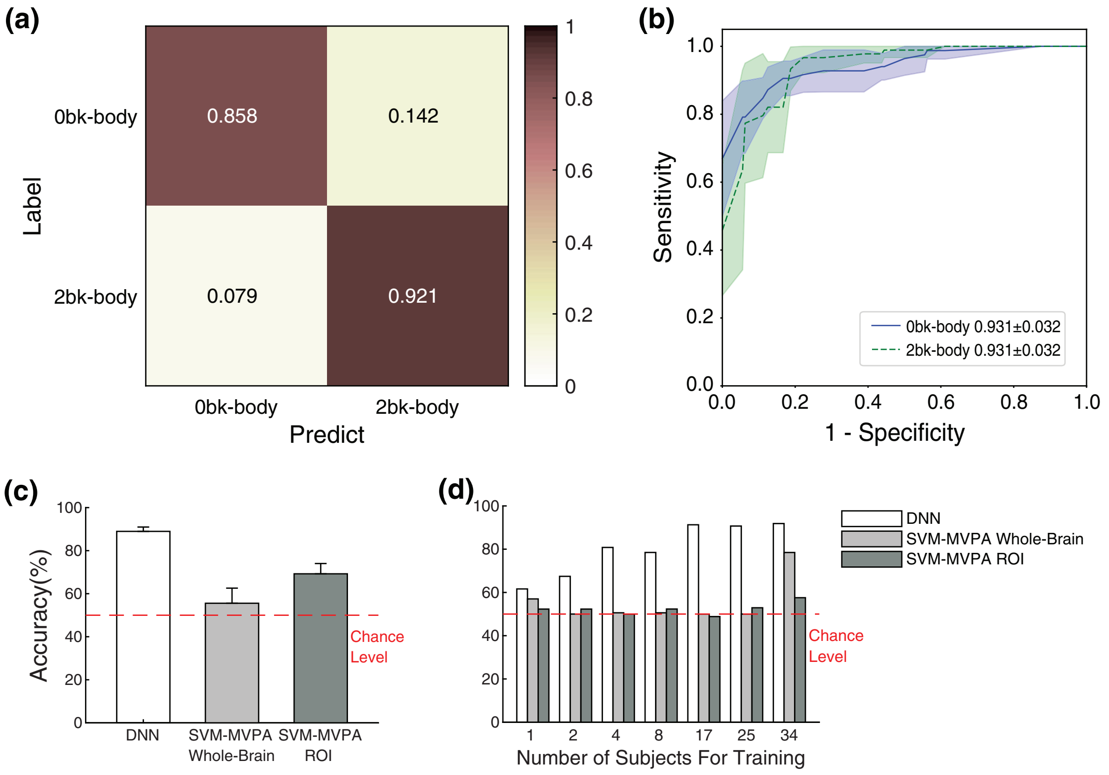

FIGURE 5 Results of transfer learning for classification of the working memory task (0bk-body vs. 2bk-body). (a) The average confusion matrix normalized to the number of instances of each label in fivefold cross-validations. This yielded an average accuracy of 89.0 ± 2.0% (mean ± *SD*) in terms of classifying the two tasks (chance level = 50%). (b) The mean (solid lines) and *SD* (shadow envelopes) of the ROC curves for each label in fivefold cross validation. The mean ROC area and *SD* are labeled in the legend. (c) Accuracy of fivefold cross-validation classification on the working memory task on a small dataset. The accuracy of the DNN (89% ± 2%) was significantly higher than that of the SVM-MVPA whole-brain (t[8] = 9.14, *p* = .000017; mean ± *SD* = 55.6 ± 7.9%) and SVM-MVPA ROI (t[8] = 7.59, *p* = .000064; mean ± *SD* = 69.2 ± 5.4%) method. (d) The performance of the three methods across different numbers of subjects for training (N_{Subj}). N_{Subj} = 2 was enough for the DNN to learn the classification, whereas the SVM-MVPA whole-brain and SVM-MVPA ROI methods needed N_{Subj} = 34

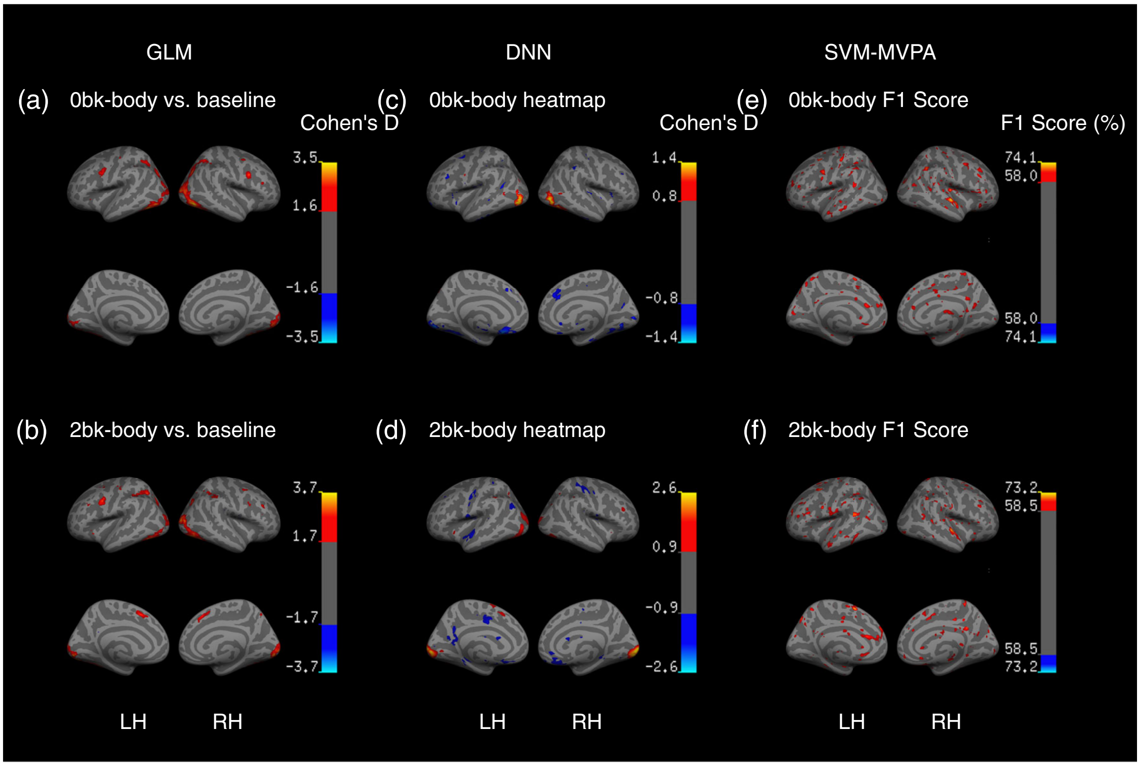

FIGURE 6 Visualization of brain task-related maps during the working memory task via GLM, DNN, and SVM-MVPA. (a, b) Cohen's *d* for the GLM beta maps. (c, d) Cohen's *d* for the DNN pattern maps, which showed similar localizations of the fusiform and lateral occipital areas, and dissimilar localizations of lateral and medial orbitofrontal areas, compared with those of the GLM beta maps. (e, f) The F1 score of the SVM-MVPA searchlight method. It shows that the searchlight failed to localize any functional cluster related to the task but reported widespread scatters all over the brain

± 7.9%) and SVM-MVPA ROI (t[8] = 7.59, *p* = .000064; mean ± *SD* = 69.2 ± 5.4%) through a two sample *t* test.

We then validated the amount of data needed for learning. The results showed that N_{Subj} = 2 was enough for the DNN to learn the classification (accuracy = 67.4%), whereas SVM-MVPA whole-brain and SVM-MVPA ROI needed N_{Subj} = 34, yielding accuracy values = 91.9, 78.5, and 57.6%, respectively (Figure 5d).

Finally, we visualized the DNN pattern maps and found that the Cohen's *d* reached its highest value in the Brodmann area 38 (fusiform) and Brodmann area 18/19 (extrastriate visual areas) (Figure 6c,d), which were similar to the results of the GLM COPEs (Figure 6a,b). Moreover, the SVM-MVPA searchlight method reported widespread activity scatters, rather than activity clusters, all over the brain (Figure 6e,f). Refer to Table S2 for further details on the annotations of the maps.

3.4 | Transfer learning multiple sub-types of motor task using small datasets

Following fivefold cross-validation, the proposed DNN reached an average accuracy of 94.7 ± 1.7% (Figure 7a) and an average AUC of ROC 0.996 ± 0.005 (Figure 7b). The average confusion matrix showed that the top confusion was caused by left foot versus right foot (Figure 7a). Figure 7c shows that the accuracy of the DNN (94.7 ± 1.7%) was significantly higher than that of SVM-MVPA whole-brain (t[8] = 3.59, *p* = .0071; mean ± *SD* = 81.6 ± 7.1%) and SVM-MVPA ROI (t[8] = 8.77, *p* = .000022; mean ± *SD* = 68.6 ± 5.7%) through a two sample *t* test.

We then validated the amount of data needed for learning. All three methods reported higher than chance-level accuracy across all N_{Subj}. N_{Subj} = 8 was enough for the DNN (80.3%) to outperform the ordinary SVM-MVPA whole-brain (41.7%) and SVM-MVPA ROI (56.3%) methods in terms of accuracy (Figure 7d).

Finally, we visualized the DNN pattern maps and found that Cohen's *d* reached the highest values in the corresponding motor topological areas, which was similar to the results of the GLM COPEs and the SVM-MVPA searchlight method (Figure 8). Refer to Table S2 for further details on the annotations of the maps.

4 | DISCUSSION

4.1 | Summary

In this study, we proposed a general deep learning framework for decoding and mapping ongoing brain task states from whole-brain fMRI signals of humans. After training and testing it using data from

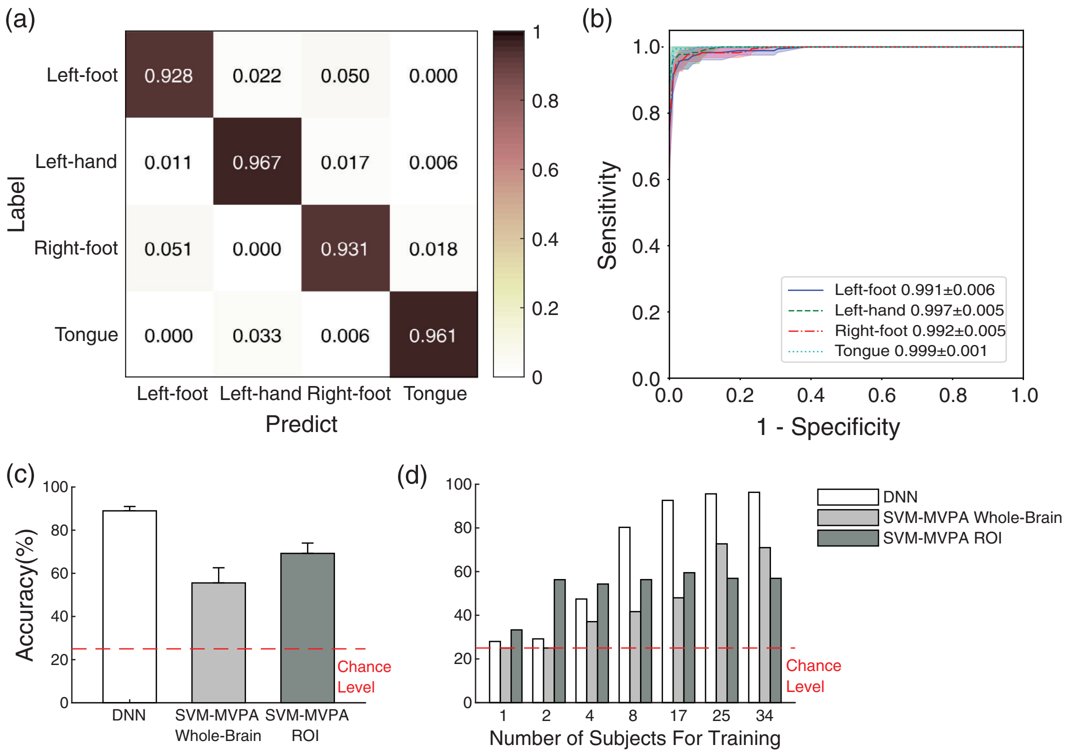

FIGURE 7 Results of transfer learning of classification on motor tasks (left foot, left hand, right foot, and tongue). (a) The average confusion matrices normalized to the number of instances of each label in the fivefold cross-validation, with the top confusion caused by left foot vs. right foot. It reported an average accuracy of 94.7 ± 1.7% (mean ± *SD*) on the four tasks (chance level = 25%). (b) The mean (solid lines) and *SD* (shadow envelopes) of ROC curves for each label in the fivefold cross validation. The mean ROC area and *SD* are labeled in the legend. (c) Accuracy of fivefold cross-validation classification on the motor task on a small dataset. The accuracy of the DNN (94.7 ± 1.7%) was significantly higher than that of SVM-MVPA whole-brain (t[8] = 3.59, *p* = .0071; mean ± *SD* = 81.6 ± 7.1%) and SVM-MVPA ROI (t[8] = 8.77, *p* = .000022; mean ± *SD* = 68.6 ± 5.7%) methods. (d) The performance of the three methods across different numbers of subjects for training (N_{Subj}). All conditions reported higher than chance-level accuracy. N_{Subj} = 8 was enough for the DNN to outperform the ordinary SVM-MVPA methods

the HCP, the proposed DNN classifier achieved an average accuracy of 93.7% and an average area under the ROC curve of 0.996 on a seven-class classification task. The DNN was able to transfer-learn a new classification task using small fMRI datasets and yielded higher accuracy than SVM-MVPA methods. Moreover, a network visualization analysis showed that the DNN automatically detected and located features in areas of the brain that have been reported to have significant effects in the traditional GLM method.

4.2 | Deep learning as a research tool

Deep learning is capable of automatic data-driven feature learning and has deeper models than earlier methods. Analogous to the brain's sensory network, DNNs perform complex computations through deep stacks of simple intra-layer neural circuits. Thus, researchers have widely used DNN models to understand the human brain network, especially sensory brain networks (Eickenberg, Gramfort, Varoquaux, & Thirion, 2017; Guclu & van Gerven, 2015; Horikawa & Kamitani, 2017; Rajalingham et al., 2018; Yamins & DiCarlo, 2016). At the same time, DNNs are capable of discovering complex structures within high-dimensional input data, and can transform these structures into abstract levels (LeCun et al., 2015). These important features allow researchers to efficiently model complex systems without the burden of model/prior knowledge selection, especially in cases where too many features exist, as when analyzing medical images (Shen et al., 2017). Thus, DNNs are widely used by researchers for medical image analysis, such as brain image segmentation (Havaei et al., 2017; Wachinger et al., 2018; Zhang et al., 2015), neurology and psychiatric diagnostics (Hosseini-Asl, Keynton, & El-Baz, 2016; Meszlenyi, Buza, & Vidnyanszky, 2017; Plis et al., 2014; Vieira et al., 2017), brain state decoding (Jang et al., 2017), and brain computer interfaces (Schirrmeister et al., 2017).

A variety of deep methods have been applied to fMRI data, such as the autoencoder (Kim, Calhoun, Shim, & Lee, 2016), deep belief network (DBN; Jang et al., 2017; Plis et al., 2014), long short-term memory (LSTM) recurrent neural network (RNN; Li & Fan, 2019), and 2D CNN (Meszlenyi et al., 2017). Although the autoencoder is known to be efficient, especially when the dataset is small, it overemphasizes some relationships while neglecting others, that is, it loses information. DBNs have been criticized for a number of shortcomings, such as the computational cost associated with training and loss of

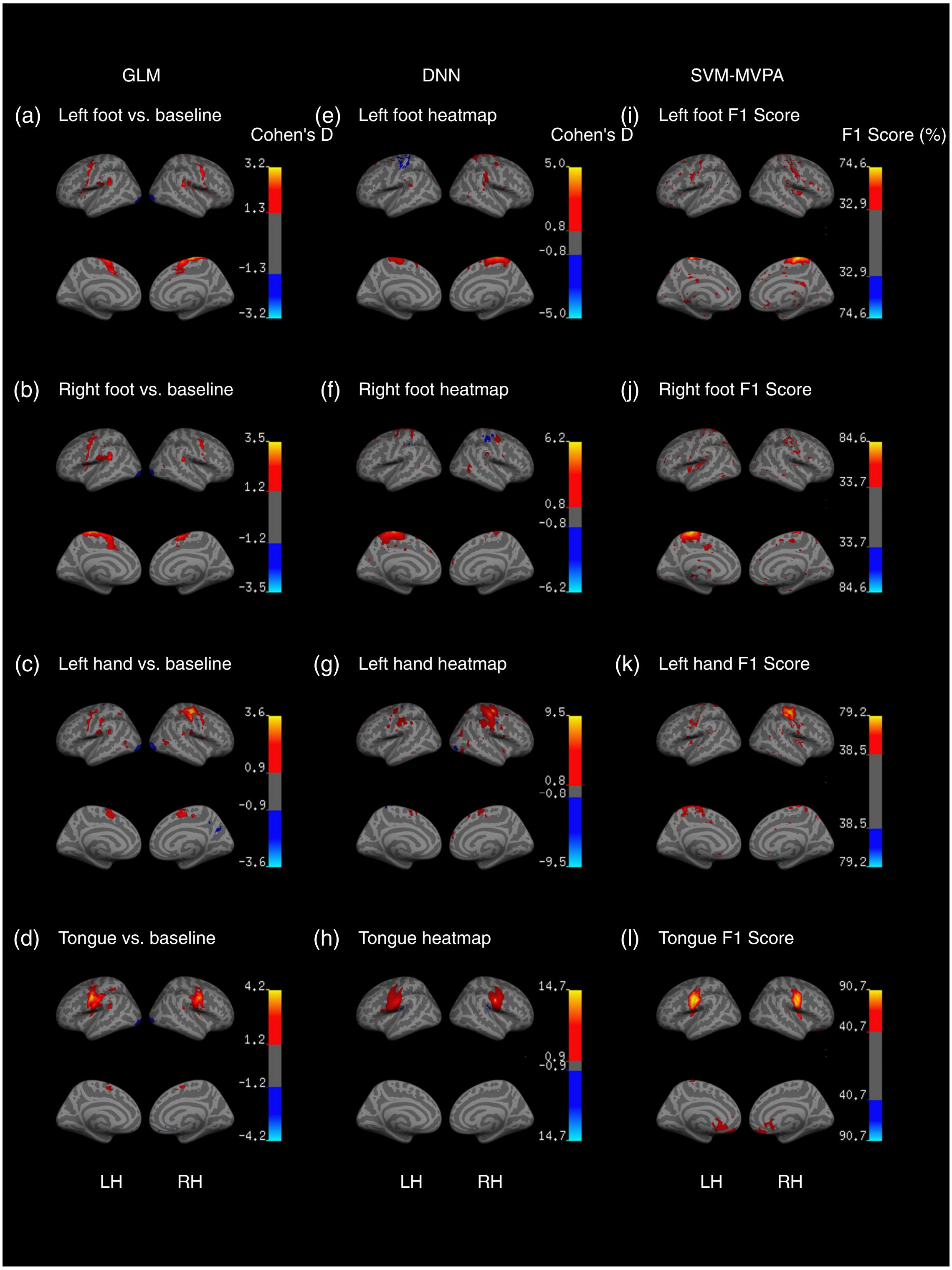

FIGURE 8 Visualization of brain task-related maps during motor tasks via GLM, DNN, and SVM-MVPA. (a–d) Cohen's *d* effect sizes for the GLM beta maps. (e–h) Cohen's *d* effect sizes for DNN pattern maps. (i–l) The F1 score of the SVM-MVPA searchlight method. Collectively, the three methods identified similar brain activity maps

spatial information in learning, which may significantly affect their performance and interpretability in medical image analysis (Voulodimos, Doulamis, Doulamis, & Protopapadakis, 2018). The RNN with LSTM, a deep learning method for sequence modeling, ignores spatial information within the input data (Hochreiter & Schmidhuber, 1997). The 2D CNN cannot encode the 3D nature of fMRI data. Thus, both Li and Fan (2019) and Meszlenyi et al. (2017) methods require functional network-based features as inputs. Our study represents a significant departure from these studies, however, by directly targeting fMRI volume through the 3D CNN. The proposed 3D CNN, which makes use of the spatial structure of the input data, is efficient in capturing spatial relationships of the brain activity. As end-to-end learning methods, CNNs have the unique capability of learning features automatically and avoids the design of a feature extractor. On the contrary, CNNs heavily rely on manually labeled training data, but this is not a problem for neuroimaging research because almost all neuroimaging data are carefully labeled with diagnostics, task states, and questionnaires. Moreover, because the CNN requires scant handcrafting of features by experts, it is easily usable by data scientists on neuroimaging data.

We used an NVIDIA GTX 1080Ti GPU in our experiments. The initial training took a long time (72 hr for 30 epochs) while transfer learning took much less time (9 hr for 120 epochs on the two-class classification task, and 21 hr for 120 epochs in the four-class classification task). The proposed CNN was composed of three convolutional layers and two fully connected layers with 3,981,852 parameters. Given these layers and their hyperparameters, we could make countless possible combinations of network architectures. We evaluated the impact of the number of 1 × 1 × 1 channels (Figure 3c), and found that three channels provided enough information to distinguish between task states. The proposed model was implemented on the PyTorch library: a free and open-source software and among the most popular deep learning platforms. Researchers interested in reusing the proposed model on other platforms can refer to the Open Neural Network Exchange created by Facebook and Microsoft.

4.3 | Visualization of learnt patterns

The proposed method also offers researchers the opportunity to investigate decisions of the neural network. A challenge of applying deep models to neuroimaging research is the black-box characteristic of this approach: No one knows exactly what the deep network is doing. In recent years, a method for tracing consecutive layers of weights back to the original image inputs has been proposed, and has achieved good performance in natural image recognition (Springenberg et al., 2014). Researchers have employed various methods for the analysis of the processes of DNNs (Bach et al., 2015; Yamins & DiCarlo, 2016). Guclu and van Gerven (2015) employed a DNN model to predict the responses of each voxel and found a gradient in the feature complexity aligning with the ventral pathway. Through linear predictive models, Eickenberg et al. (2017) generalized human visual cortical activity maps elicited by visual stimulation. Jang et al. (2017) proposed a ROI-wise task-specific activity map by extracting the weights of the nodes in the output layer of a deep network.

We employed guided back-propagation, a widely used network visualization method, to visualize features of the data detected by the CNN for the classification of each entered data item. The visualized voxels with values other than zero comprised features important for classification. There is a criticism where good decoding performance is not a guarantee that patterns of brain activity are learned (Ritchie, Kaplan, & Klein, 2019), for a decoder may learn from nuisance or latent variables (Riley, 2019)—for example, the different visual responses to different stimulus images or patterns of response key-pressing across the seven tasks. The guided back-propagation allows scientists to intuitively locate and investigate features the DNN detected in every entered fMRI data item. In this work, the similarity between the pattern maps and the GLM maps (Figures 4, 6, and 8) suggest that the proposed DNN decoded states from task-related brain activity patterns, not from nuisance variables. Furthermore, correlated with the β maps of the GLM, the pattern maps showed potential for localizing state-related areas of the brain. However, the statistical property of guided back-propagation remains unclear, and we should be cautious until further investigations on its reliability and statistical properties.

4.4 | Transfer learning helps model construction with small samples

Transfer learning is a machine learning method that learns from networks trained on a related but different task from the given one. By taking advantage of transferred knowledge, it eliminates the need for big training data (Rawat & Wang, 2017). Hosseini-Asl et al. (2018) pre-trained a 3D convolutional autoencoder to capture anatomical shape variations in brain MRI scans and fine-tuned it for AD classification on images from 210 subjects. Gao et al. (2019) pre-trained a 2D-CNN for classification on ImageNet, a database containing >14 million natural images, and fine-tuned it to decode 2D fMRI slices. The proposed method transfer-learns in a more direct way—transferring knowledge learnt from a big fMRI dataset to limited fMRI datasets. We believe that the proposed DNN can transfer-learn a related but different decoding task using fMRI data from as few as four subjects (Figure 5d). Although our deep learning framework was trained and validated using the HCP S1200 dataset, the consistent internal properties of human hemodynamic responses make fMRI data reasonably consistent across scanners and sites. Nowadays big datasets, such as BioBank, HCP, and OpenfMRI, provide comprehensive neuroimaging scans across a wide range of ages and diseases, and provide the opportunity for pretraining on big data and transfer learning on small fMRI datasets.

4.5 | Transfer learning to the WM task

We evaluated the generalizability of our deep learning framework in transfer learning to WM data of 43 subjects. WM refers to a brain function for the temporary storage and manipulation of information for cognitive processing (Baddeley, 1992). We chose the WM because researches have shown that it is not processed in a single brain site, but stored and processed in widely distributed brain regions (Christophel, Klink, Spitzer, Roelfsema, & Haynes, 2017; Mencarelli et al., 2019), ranging from the sensory (Pasternak, Lui, & Spinelli, 2015; Sreenivasan, Curtis, & D'Esposito, 2014) to prefrontal (Durstewitz, Seamans, & Sejnowski, 2000; Riley & Constantinidis,

2015) and parietal (Xu & Jeong, 2016) cortices. This distributed nature of the WM makes it impossible to decode from a single ROI, as shown in this work, and poses a major obstacle to ROI selection in the MVPA. We proposed a machine-learning framework that automatically abstracted the activity patterns of the brain, affording a powerful tool to decode comprehensive brain functions. Moreover, by using guided back-propagation, we showed that the proposed model detected features from areas of the brain that have been reported to be related to the WM function: BA 32 (anterior cingulate cortex, Owen, McMillan, Laird, and Bullmore (2005)), BA 38 (fusiform, Downing, Jiang, Shuman, and Kanwisher (2001); Kanwisher, McDermott, and Chun (1997)), and BA 18/19 (extrastriate visual cortex, Grill-Spector, Kourtzi, and Kanwisher (2001)). Its performance in classifying two tasks provided more evidence that the model learnt from task-related brain activity, rather than nuisance variables, because the stimuli were consistent, with merely the task altered, between 0-back and 2-back.

4.6 | Transfer learning to the motor task

We evaluated the generalizability of our deep learning framework in transferring learning to multi-class motor data of 43 subjects. Motor-related information was encoded in the primary motor cortex, premotor cortex, and supplementary motor area around the central sulcus. The topological nature of the motor area makes it the first cortex to be decoded in the human brain (Dehaene et al., 1998). In our experiment, the SVM-MVPA was good at single-label classification (high F1 scores for each task in Figure 8) but delivered poor performance at multi-class classification (low accuracy in Figure 7d). The proposed method showed its potential in multi-class classification over the SVM-MVPA method. Cognitive neuroscience has attended to particular brain functions, but researchers are now calling for models that generalize beyond specific tasks (Varoquaux & Poldrack, 2019; Yarkoni & Westfall, 2017). Brain systems are often engaged in a variety of brain functions (Varoquaux et al., 2018), and predictive investigations of general tasks can ultimately lead to a greater understanding of the human brain. The proposed method provides researchers with the choice of decoding and interpreting brain functions in an integrative way.

4.7 | Future work

Although we illustrated the deep model's ability to read the fMRI time series, researchers can modify the input layer and take a volume of brain features as input to the proposed deep model, such as the amplitude of low-frequency fluctuation (ALFF), fractional ALFF (fALFF), and regional homogeneity (ReHo) of resting-state fMRI as well as the fractional anisotropy (FA) and mean diffusivity (MD) of diffusion tensor imaging (DTI). The model is also applicable to multi-modal inputs to different channels, which are important for research in psychiatry and neurology because most of the open datasets used, such as ADNI (Alzheimer's Disease Neuroimaging Initiative), ABIDE (Autism Brain Imaging Data Exchange), BioBank, and SchizConnect. The proposed method can provide a basis for a brain-based information retrieval systems by classifying brain activity into different categories: brain-based disorder or psychiatric classification. Varieties of deep learning methods have shown their power in searching for biomarkers of psychiatric and neurologic diseases (Vieira et al., 2017), and the proposed method provides one more choice.

Activity classification can also benefit real-time fMRI neurofeedback (rt-fMRI-NF), a technology providing subjects with feedback stimuli from ongoing brain activity collected by an MRI scanner (Cox, Jesmanowicz, & Hyde, 1995; Sulzer et al., 2013). Recently, a data-driven and personalized MVPA rt-fMRI-NF method (Shibata, Watanabe, Sasaki, & Kawato, 2011), decoded neurofeedback (DecNef), was proposed, and has shown outstanding performance in both basic and clinical research (Thibault, MacPherson, Lifshitz, Roth, & Raz, 2018; Watanabe, Sasaki, Shibata, & Kawato, 2017). The proposed deep model has the potential to decode multiple brain states from whole-brain fMRI time series and to output these to feedback processing in real time. Moreover, the model can be fine-tuned to individual brain activity through transfer learning to build up a personalized rt-fMRI-NF.

4.8 | Conclusion

We proposed a method to classify and map an individual's ongoing brain function directly from a 4D fMRI time series. Our approach allows for the decoding of a subject's task state from a short fMRI scan without the burden of feature selection. This flexible and efficient brain-decoding method can be applied to both large-scale massive data and fine, small-scale data in neuroscience. Moreover, its characteristics of facility, accuracy, and generalizability allow the deep framework to be easily applied to a new population as well as a wide range of neuroimaging research, including internal mental state classification, psychiatric disease diagnosis, and real-time fMRI neurofeedback.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (91432301, 81701665, 81790650, 81790651, 81430037, 81727808, and 31421003), the Fundamental Research Funds for the Central Universities of China (WK2070000033), and the Beijing Municipal Science & Technology Commission (Z171100000117012).

CONFLICT OF INTEREST

The authors declare that the research reported here was conducted in the absence of any commercial or financial relationships that can be construed as potential conflicts of interest.

AUTHOR CONTRIBUTIONS

X.W. and X.L. analyzed the data and wrote the article. Z.J., B.A.N., Y.Z., Y.W., H.W., Y.L., Y.Z., and F.W. processed and analyzed the data. J.G. and B.Q. conceived of the study and contributed to writing the manuscript. All authors discussed the results and reviewed the manuscript.

DATA AVAILABILITY STATEMENT

All scripts described in this paper are available at https://github.com/ustc-bmec/Whole-Brain-Conv.

ORCID

Xiaoxiao Wang https://orcid.org/0000-0002-8498-7388

Jia-Hong Gao https://orcid.org/0000-0002-9311-0297

REFERENCES

Bach, S., Binder, A., Montavon, G., Klauschen, F., Muller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, *10*, e0130140. https://doi.org/10.1371/journal.pone.0130140

Baddeley, A. (1992). Working memory. *Science*, *255*, 556–559. https://doi.org/10.1126/science.1736359

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., … Consortium, W.U.-M.H. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, *80*, 169–189. https://doi.org/10.1016/j.neuroimage.2013.05.033

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J. D. (2017). The distributed nature of working memory. *Trends in Cognitive Sciences*, *21*, 111–124. https://doi.org/10.1016/j.tics.2016.12.007

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*, 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Ciompi, F., de Hoop, B., van Riel, S. J., Chung, K., Scholten, E. T., Oudkerk, M., … van Ginneken, B. (2015). Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. *Medical Image Analysis*, *26*, 195–202. https://doi.org/10.1016/j.media.2015.08.001

Cohen, J. (1998). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Lawrence Erlbaum Associates.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*, 162–173. https://doi.org/10.1006/cbmr.1996.0014

Cox, R. W., Jesmanowicz, A., & Hyde, J. S. (1995). Real-time functional magnetic resonance imaging. *Magnetic Resonance in Medicine*, *33*, 230–236. https://doi.org/10.1002/mrm.1910330213

Dehaene, S., Le Clec, H. G., Cohen, L., Poline, J. B., van de Moortele, P. F., & Le Bihan, D. (1998). Inferring behavior from functional brain images. *Nature Neuroscience*, *1*, 549–550. https://doi.org/10.1038/2785

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). DeCAF: A deep convolutional activation feature for generic visual recognition. *Proceedings of the 31st International Conference on Machine Learning*, p. 647–655.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*, 2470–2473. https://doi.org/10.1126/science.1063414

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. *Journal of Neurophysiology*, *83*, 1733–1750. https://doi.org/10.1152/jn.2000.83.3.1733

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194. https://doi.org/10.1016/j.neuroimage.2016.10.001

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*, 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Gao, Y., Zhou, B., Zhou, Y., Shi, L., Tao, Y., Zhang, J. (2019). Transfer learning-based behavioural task decoding from brain activity. *Proceedings of the 2nd international conference on healthcare science and engineering*, p. 71-81. https://doi.org/10.1007/978-981-13-6837-0_6.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*, 1409–1422. https://doi.org/10.1016/s0042-6989(01)00073-6

Guclu, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*, 10005–10014. https://doi.org/10.1523/jneurosci.5023-14.2015

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., … Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, *35*, 18–31. https://doi.org/10.1016/j.media.2016.05.004

Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews. Neuroscience*, *7*, 523–534. https://doi.org/10.1038/nrn1931

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 770–778. https://doi.org/10.1109/CVPR.2016.90

He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J. (2015). Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. *2015 IEEE International Conference on Computer Vision (Iccv)*. p. 1026–1034, https://doi.org/10.1109/Iccv.2015.123.

Hebart, M. N., Gorgen, K., & Haynes, J. D. (2014). The decoding toolbox (TDT): A versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, *8*, 88. https://doi.org/10.3389/fninf.2014.00088

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*, 15037. https://doi.org/10.1038/ncomms15037

Hosseini-Asl, E., Ghazal, M., Mahmoud, A., Aslantas, A., Shalaby, A. M., Casanova, M. F., … El-Baz, A. (2018). Alzheimer's disease diagnostics by a 3D deeply supervised adaptable convolutional network. *Frontiers in Bioscience (Landmark Edition)*, *23*, 584–596. https://doi.org/10.2741/4606

Hosseini-Asl, E., Keynton, R., & El-Baz, A. (2016). Alzheimer's disease diagnostics by adaptation of 3D convolutional network. *2016 IEEE International Conference on Image Processing (ICIP)*, p. 126–130. https://doi.org/10.1109/icip.2016.7532332

Hu, J., Shen, L., Sun, G. (2017). Squeeze-and-excitation networks. arXiv:1709.01507.

Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv:1602.07360.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning*, p. 448–456.

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. https://doi.org/10.1016/j.neuroimage.2016.04.003

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *17*, 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., … Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, *172*, 1122–1131 e9. https://doi.org/10.1016/j.cell.2018.02.010

Kim, B., & Oertzen, T. V. (2018). Classifiers as a model-free group comparison test. *Behavior Research Methods*, *50*, 416–426. https://doi.org/10.3758/s13428-017-0880-z

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. https://doi.org/10.1016/j.neuroimage.2015.05.018

Kingma, D.P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv:1412.6980.

Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*, *38*, 649–662. https://doi.org/10.1016/j.neuroimage.2007.02.022

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 3863–3868. https://doi.org/10.1073/pnas.0600244103

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems 25*, p. 1097–1105.

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

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. https://doi.org/10.1016/j.neuroimage.2019.116059

Maturana, D., & Scherer, S. (2015). *VoxNet: A 3D Convolutional Neural Network for real-time object recognition*. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, p. 922-928. https://doi.org/10.1109/IROS.2015.7353481.

Mencarelli, L., Neri, F., Momi, D., Menardi, A., Rossi, S., Rossi, A., & Santarnecchi, E. (2019). Stimuli, presentation modality, and load-specific brain activity patterns during n-back task. *Human Brain Mapping*, *40*, 3810–3831. https://doi.org/10.1002/hbm.24633

Meszlenyi, R. J., Buza, K., & Vidnyanszky, Z. (2017). Resting state fMRI functional connectivity-based classification using a convolutional neural network architecture. *Frontiers in Neuroinformatics*, *11*, 61. https://doi.org/10.3389/fninf.2017.00061

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., … Smith, S. M. (2016). Multimodal population brain imaging in the UKBiobank prospective epidemiological study. *Nature Neuroscience*, *19*, 1523–1536. https://doi.org/10.1038/nn.4393

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*, 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, *25*, 46–59. https://doi.org/10.1002/hbm.20131

Pasternak, T., Lui, L. L., & Spinelli, P. M. (2015). Unilateral prefrontal lesions impair memory-guided comparisons of contralateral visual motion. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*, 7095–7105. https://doi.org/10.1523/JNEUROSCI.5265-14.2015

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., … Calhoun, V. D. (2014). Deep learning for neuroimaging: A validation study. *Frontiers in Neuroscience*, *8*, 229. https://doi.org/10.3389/fnins.2014.00229

Poldrack, R. A., Halchenko, Y. O., & Hanson, S. J. (2009). Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological Science*, *20*, 1364–1372. https://doi.org/10.1111/j.1467-9280.2009.02460.x

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the Core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *38*, 7255–7269. https://doi.org/10.1523/JNEUROSCI.0388-18.2018

Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, *29*, 2352–2449. https://doi.org/10.1162/NECO_a_00990

Riley, M. R., & Constantinidis, C. (2015). Role of prefrontal persistent activity in working memory. *Frontiers in Systems Neuroscience*, *9*, 181. https://doi.org/10.3389/fnsys.2015.00181

Riley, P. (2019). Three pitfalls to avoid in machine learning. *Nature*, *572*, 27–29. https://doi.org/10.1038/d41586-019-02307-y

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*, *70*, 581–607. https://doi.org/10.1093/bjps/axx023

Rubin, T. N., Koyejo, O., Gorgolewski, K. J., Jones, M. N., Poldrack, R. A., & Yarkoni, T. (2017). Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition. *PLoS Computational Biology*, *13*, e1005649. https://doi.org/10.1371/journal.pcbi.1005649

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., … Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, *38*, 5391–5420. https://doi.org/10.1002/hbm.23730

Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit Worksh*. p. 806–813, https://doi.org/10.1109/CVPRW.2014.131.

Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, *19*, 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Shibata, K., Watanabe, T., Sasaki, Y., & Kawato, M. (2011). Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science*, *334*, 1413–1415. https://doi.org/10.1126/science.1212003

Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

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

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*, 82–89. https://doi.org/10.1016/j.tics.2013.12.001

Sulzer, J., Haller, S., Scharnowski, F., Weiskopf, N., Birbaumer, N., Blefari, M. L., … Sitaram, R. (2013). Real-time fMRI neurofeedback: Progress and challenges. *NeuroImage*, *76*, 386–399. https://doi.org/10.1016/j.neuroimage.2013.03.033

Thibault, R. T., MacPherson, A., Lifshitz, M., Roth, R. R., & Raz, A. (2018). Neurofeedback with fMRI: A critical systematic review. *NeuroImage*, *172*, 786–807. https://doi.org/10.1016/j.neuroimage.2017.12.071

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., & Consortium, W.U.-M.H. (2013). The WU-Minn human connectome project: An overview. *NeuroImage*, *80*, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., … Consortium, W.U.-M.H. (2012). The human connectome project: A data acquisition perspective. *NeuroImage*, *62*, 2222–2231. https://doi.org/10.1016/j.neuroimage.2012.02.018

Varoquaux, G., & Poldrack, R. A. (2019). Predictive models avoid excessive reductionism in cognitive neuroimaging. *Current Opinion in Neurobiology*, *55*, 1–6. https://doi.org/10.1016/j.conb.2018.11.002

Varoquaux, G., Schwartz, Y., Poldrack, R. A., Gauthier, B., Bzdok, D., Poline, J. B., & Thirion, B. (2018). Atlases of cognition with large-scale human brain mapping. *PLoS Computational Biology*, *14*, e1006565. https://doi.org/10.1371/journal.pcbi.1006565

Vieira, S., Pinaya, W. H., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience and Biobehavioral Reviews*, *74*, 58–75. https://doi.org/10.1016/j.neubiorev.2017.01.002

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, *2018*, 7068349–7068313. https://doi.org/10.1155/2018/7068349

Wachinger, C., Reuter, M., & Klein, T. (2018). DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*, *170*, 434–445. https://doi.org/10.1016/j.neuroimage.2017.02.035

Watanabe, T., Sasaki, Y., Shibata, K., & Kawato, M. (2017). Advances in fMRI real-time neurofeedback. *Trends in Cognitive Sciences*, *21*, 997–1010. https://doi.org/10.1016/j.tics.2017.09.010

Wen, H., Shi, J., Chen, W., & Liu, Z. (2018). Transferring and generalizing deep-learning-based neural encoding models across subjects. *NeuroImage*, *176*, 152–163. https://doi.org/10.1016/j.neuroimage.2018.04.053

Xu, Y. D., & Jeong, S. K. (2016). The contribution of human superior intraparietal sulcus to visual short-term memory and perception. In *Mechanisms of sensory working memory: Attention and perfomance* (Vol. XXV, pp. 33–42). San Diego: Academic Press. https://doi.org/10.1016/B978-0-12-801371-7.00004-1

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*, 356–365. https://doi.org/10.1038/nn.4244

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122. https://doi.org/10.1177/1745691617693393

Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., & Shen, D. (2015). Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*, *108*, 214–224. https://doi.org/10.1016/j.neuroimage.2014.12.061

SUPPORTING INFORMATION

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

How to cite this article: Wang X, Liang X, Jiang Z, et al. Decoding and mapping task states of the human brain via deep learning. *Hum Brain Mapp*. 2020;41:1505–1519. https://doi.org/10.1002/hbm.24891