Deep Learning for Prognosis Using Task-fMRI: A Novel Architecture and Training Scheme

Ge Shi
geshi@ucdavis.edu
University of California, Davis
Davis, California, USA

Jason Smucny
jsmucny@ucdavis.edu
University of California, Davis
Davis, California, USA

Ian Davidson
davidson@cs.ucdavis.edu
University of California, Davis
Davis, California, USA

ABSTRACT

Most existing brain imaging work focuses on resting-state fMRI (rs-fMRI) data where the subject is at rest in the scanner typically for disease diagnosis problems. Here we analyze task fMRI (t-fMRI) data where the subject performs a multi-event task over multiple trials. t-fMRI data allows exploring more challenging applications such as prognosis of treatment but at the cost of being more complex to analyze. Not only do multiple types of trials exist but the trials of each type are repeated a varying number of times for each subject. This leads to a multi-view (multiple types of trials) and multi-instance (multiple trials of each type of each subject) setting. We propose a deep multi-model architecture to encode multi-view brain activities from t-fMRI data and a multi-layer perceptron ensemble model to combine these view models and make subject-wise predictions. We explore domain adaptation transfer learning between models to address unbalanced views and a novel way to make predictions out of multi-instance embeddings. We evaluate our model's performance on subject-wise cross-validations to accurately determine performance. The experimental results show the proposed method outperforms published methods on the AX-CPT fMRI data for the prognosis problem of predicting treatment improvement in recent-onset childhood schizophrenia. To our knowledge, this is the first data-driven study of the aforementioned task on voxelwise t-fMRI data of the whole brain.

CCS CONCEPTS

• **Applied computing** → **Imaging**; • **Computing methodologies** → **Learning settings**.

KEYWORDS

neuroimaging, task-fMRI, deep multi-view multi-instance learning

ACM Reference Format:
Ge Shi, Jason Smucny, and Ian Davidson. 2022. Deep Learning for Prognosis Using Task-fMRI: A Novel Architecture and Training Scheme. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539362

1 INTRODUCTION

Analysis of functional magnetic resonance imaging (fMRI) data is most frequently performed for patients in "resting state" (absence of a task) during which the default mode network (DMN) [9] is the most active network. Since the brain is in a steady-state this facilitates many common machine learning methods as the underlying data can be converted to a graph (if correlations between voxels over time are used as edge-weights) or even a simple picture (if the voxel values over time are averaged). Such analysis of the DMN has been useful for exploring disease diagnosis such as Post traumatic stress disorder (PTSD) [13], Alzheimer's Disease [14] and even Traumatic Brain Injury (TBI) [4].

Table 1: Differences between rs-fMRI and AX-CPT t-fMRI.

	Events During Trial	Number of Trials
rs-fMRI	None	One
AX-CPT	CueA, ProbeX, CueB, ProbeY	Varies by subject and trial-type

However, important problems such as prognosis (the forecast of the effectiveness of a treatment for a disease) can not be easily determined by analyzing resting state fMRI (rs-fMRI) and the DMN [9]. Instead task fMRI (t-fMRI) data is used when the subject is performing a multi-event task inside the scanner. Since the brain is performing a task, its behavior is dynamic and hence the simple representations such as graphs or pictures for rs-fMRI data are inappropriate. Such t-fMRI data, unlike the rs-fMRI, is understudied in machine learning contexts ([22, 25]). In this paper, we focus on the executive network elicited by the AX-version of the Continuous Performance Task (AX-CPT) (Figure 1) but our work can be generalized to other tasks which involve multi-event task with repetitive trials. The application of our work is to forecast the symptomatic improvement due to treatment for recent-onset schizophrenia in children by analyzing the baseline AX-CPT t-fMRI data before the treatment is applied.

Analyzing t-fMRI data is challenging and differs from rs-fMRI in several ways as shown in Table 1. For example, in the AX-CPT task each trial consists of several events with each trial-type being repeated different number of times to each subject (see Figure 1). However, only one clinical evaluation (label) is given to each subject. This leads to a novel multi-view (a view for each type of trial) and multi-instance (an instance for each trial taken by the subject) setting. In our work, we propose a deep learning architecture as shown in Figure 3 to address the challenges in t-fMRI data (see section 2). We build a model for each trial-type (`CueA`⟶`ProbeX`, `CueA`⟶`ProbeY`, `CueB`⟶`ProbeX`, `CueB`⟶`ProbeY`), which we will denote as: AX, BX, AY, BY. Since trial-types are not performed with equal frequencies between each other and subjects, we leverage transfer learning to make the training on rarer types possible. We

KDD '22, August 14–18, 2022, Washington, DC, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3539362

then use a multi-layer perceptron (MLP) ensemble model to combine the models to make subject-level[1] predictions.

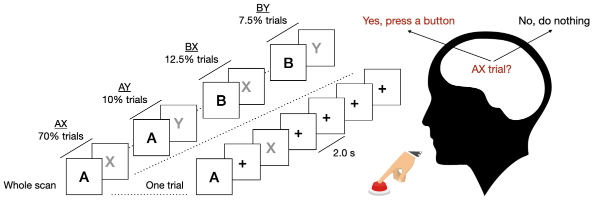


Figure 1: An illustration of the AX-CPT task. Each trial is started with a cue ('A' or 'B') and followed by some `Rest` frames ('+') and then a probe ('X' or 'Y'). The subject is expected to press a button only for the combination where a `CueA` is followed by a `ProbeX`. This task elicits an executive reasoning network in the brain. There are 4 types of trials (`CueA`→`ProbeX`, `CueA`→`ProbeY`, `CueB`→`ProbeX`, `CueB`→`ProbeY`). Each type of trials repeat for varying number of times across trial-types and subjects.

The main contributions of this work are as follows:

- We explore beyond diagnosis to prognosis applications by analyzing t-fMRI data in a novel multi-view multi-instance (MVMI) setting. We show our method (section 3) can be used to address prognosis problem in a small data situation.
- We propose a novel deep learning architecture (Figure 3) combining multiple trial-type models to make an ensemble prediction that is easily adaptable to t-fMRI data with repeated independent trials.
- To improve performance limitations caused by data scarcity, we utilize domain adaptation transfer learning to inject knowledge from the trial-type model with more training instances to other trial-type models with scarce training instances.
- We empirically show that our best practice reaches the subject-wise accuracy of 75.6% ± 3.2% in comparison with the previous best practice 72.6% ([26]) despite the latter method using more data [2]

The rest of this paper is organized as follows. In section 2, we explain the challenges of the t-fMRI data we use which motivated the MVMI learning setting. We explored the deep model architecture and technical details in section 3. Then we show our experimental setup and results in section 4. Finally, a discussion of our approach (in section 5), related work (in section 6) and future work is demonstrated (in section 7).

2 DATA CHALLENGES AND MVMI SETTING

We now overview the challenges of working with t-fMRI data in the context of the AX-CPT task. However, these insights as is our approach, are applicable to other multi-event multi-trial task settings.

We sketch the correspondences between medical terminologies and multi-view multi-instance setting in Table 2. In the following paragraphs, we use the terminologies: subject, scan, trial-type, trial, event, frame. A subject is a participant performing the tasks while the brain activity is recorded. Each subject has a label (response to treatment or not). A scan is the whole fMRI sequence of one subject. A trial is a snippet of a scan from the beginning of a cue to the last frame before the next cue. An event is either a "Cue", a "Probe" with "Rest" frames (delay time) between any cues and probes. A frame is a 3-dimensional (3D) picture of the brain consisting of voxels. In our data BOLD (blood-oxygen-level-dependent) measurements are taken at the voxel level.

2.1 AX-CPT fMRI Data

The AX-version of the Continuous Performance Test is a clinical test to: i) evaluate a human's ability to maintain a goal in short-term memory and ii) process a perceptive context. The AX-CPT and associated task parameters have been described in detail [26]. Briefly, subjects repeat a series of clinical trials with each trial consisting of a cue event followed by a probe event with some delay time between the two events. There are 4 types of cue and probe combinations: (`CueA, ProbeX`), (`CueB, ProbeX`), (`CueA, ProbeY`) and (`CueB, ProbeY`) which defines four types of trials (see Figure 1).

The subjects are instructed to make a target response (pressing a button) if and only if the probe letter `X` was preceded by the cue letter `A`. For all other trial-types (AY, BX, BY), the subject should not press the button. The target sequence trials (AX) are frequent (70% occurrences) and set up a prepotent tendency to make a target response when the probe letter X occurs. A typical `CueA`→`ProbeX` clinical trial consist of the following frames (`CueA, Rest, ProbeX, Rest, Rest, Rest`), where `CueA` and `ProbeX` can be replaced with `CueB` and `ProbeY` for other types of trials. The time resolution between two frames in a scan is 2 seconds. Each subject performs the clinical trials many times in varying proportions of trial-types. In our work, we propose a classification problem with the label being "response to treatment".

2.2 Challenges in AX-CPT Data

Given the above, there are 3 main challenges in t-fMRI data analysis which though similar to existing machine learning settings is different enough to warrant a new approach.

- **Multi-view Problem**. There are four trial-types (AX, BX, AY, BY) for AX-CPT t-fMRI scans in descending order of proportion. Each trial-type reveals a different aspect of brain activities and hence can be seen as a different view of the subject.
- **Multi-instance Problem**. As each trial-type occurs multiple times for a subject and one subject-wise clinical evaluation (label) is given to the whole scan, this creates a multi-instance setting.
- **Small Data Entity Problem**. Most t-fMRI settings involve at most 100 subjects. In our data, we have only 51 scans from different subjects, which is a small data problem.

Unsuitability of Existing Data Modelling Approaches. Since with t-fMRI data, the brain is not in a consistent state (i.e. resting) but instead performing different functions, rs-fMRI representation schemes are not suitable. For example, the co-activation matrix ([7]) or brain anatomical region graphs ([11]) methods require one stable

[1] No instances from the same subject exists in both the validation set and training set.
[2] The data used in ([26]) are collected from two scanners with different lengths of trials.

time-invariant brain networks. Similarly, though the AX-CPT fMRI data is sequential in nature, the dependencies between trials are very weak and each trial is short, hence it is not appropriate to learn LSTM or RNN temporal models models.

A Different MVMI Learning Setting. Here we justify our multi-view multi-instance (MVMI) setting that simultaneously incorporates both multi-view learning and multi-instance learning. We treat each trial-type as a view of the subject and each view is a bag of trials of the same trial-type performed by the subject. By splitting the long subject-wise scan into trials, we also mitigate the small data problem by enlarge the training examples. We sketch the correspondences between medical terminologies and multi-view multi-instance setting in Table 2. Though this setting is superficially similar to existing works, the existing works are not directly applicable. Unlike classic multi-view learning [33], our data doesn't have a fixed set of features but a varying sized bag of instances to represent a view; Unlike with classic multi-instance learning [5], our goal is not to make instance-level prediction but to make bag-level predictions.

One work [21] has the most similar setting as ours, however their labels are assigned to each instance in the bag and it's an non-deep method. The classic definitions and related works of multi-view learning and multi-instance learning can be found in section 6.

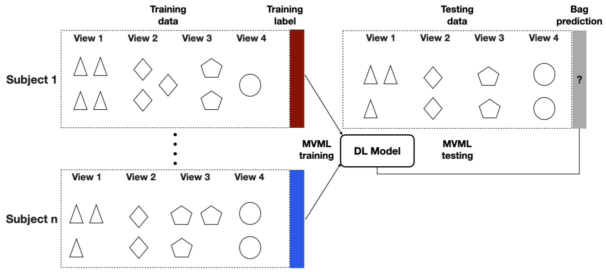


Figure 2: Our multi-instance multi-view setting. Unlike previous work, the number of instances per view is not constant and there is a single label per subject not per instance.

3 OUR APPROACH

We begin by overviewing the entire approach and then going into greater detail in each sub-section:

- For each trial type, we create a different trial-type model (see subsection 3.3). Each of them is a deep CNN model trained from all instances of the same trial type.
- Due to the varying amount of trials of trial-types, we explore transfer learning to transfer knowledge from the dominant trial-type model (AX) to other trial-type models (see subsection 3.4).
- Each view can be considered as being an expert and we explore a mixture of experts style ensemble method to fuse the different views. This dynamically combines the multiple views in an instance-specific manner. (see subsection 3.5)

Table 2: The correspondences between medical terminologies (first line) and our multi-view multi-instance setting (second line).

scan/subject	trial-type	trial	frames	voxel
data example	view	instance	channels	pixel

3.1 Our Problem Formulation

We formalize the problem of multi-view multi-instance (MVMI) learning as follows. Assume the input features is $\mathcal{X}$ and the label space is $\mathcal{Y}$, then the dataset is defined as $\mathcal{D} = \{(X_n, Y_n)|n = 1...N\}$, where $X_n \in \mathcal{X}$ and $Y_n \in \mathcal{Y}$. As this is a multi-view multi-instance setting, each training example consists of V views when each view being a bag of instances rather than just one instance. For example, the n-th example X_n in $\mathcal{D}$ is a set of V views, $X_n = \{\mathcal{B}_{n,v}|v = 1...V\}$. Each view $\mathcal{B}_{n,v}$ is a multiset (bag) of $M_{n,v}$ instances, $\mathcal{B}_{n,v} = \{\boldsymbol{x}_{n,v,m}|m = 1...M_{n,v}\}$, where $M_{n,v}$ (varies by n and v) denotes the number of instances in the v-th view of the n-th example. The goal of the MVMI is to estimate a predictor $f(\cdot, \boldsymbol{\theta}) : \mathcal{X} \rightarrow \mathcal{Y}$, with a hypothesis $\boldsymbol{\theta}$.

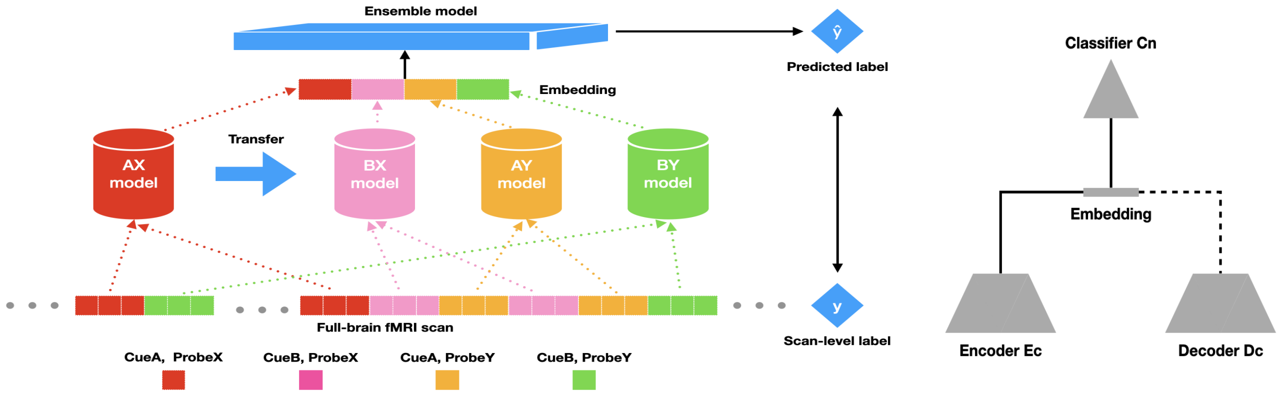
3.2 Converting t-fMRI Data to Our Setting

Each fMRI scan of one subject consists of a series of 3D frames and one binary label of prognosis. Assume the n^{th} scan of N scans is $s_n \in \mathbb{R}^{W\times H\times D\times T}$. We split each whole scan into trial instances of length L for each of the V trial-types, where $V = 4$ given trial-types (AX, BX, AY, BY) and $L = 6$ (3 frames for after each cue and 3 frames for after each probe). Each L consecutive 3D frames of size (W, H, D) beginning with each "Cue" are concatenated into a training instance $\boldsymbol{x}_{n,v,m} \in \mathbb{R}^{W\times H\times D\times L}$ for each trial-type model. The bag of all trial instances from s_n are put together to form a training bag X_n. Given the training bag X_n and subject-wise label y_n of the n^{th} subject out of the total N subjects, the goal of the model is to estimate the conditional probability $p(y_n|X_n)$, where $X_n = \{\boldsymbol{x}_{n,v,m}|v = 1...V, m = 1...M_{n,v}\}$.

3.3 Model Architecture

The deep learning architecture we have created for this situation is shown in Figure 3a. We build four 3D CNN models with the same architecture, called the trial-type model or view model, to train and evaluate the instances of each trial-type. Each trial-type model model has two heads sharing the same backbone feature extraction network (see Figure 3b), one reconstruction head and one classification head. Assume the feature extraction network is Ec and it extracts n_f features from a trial instance. Ec and the reconstruction head Dc forms an autoencoder to minimize the reconstruction loss $\mathcal{L}_{rec}(\boldsymbol{x}_{n,v,m}, Dc(Ec(\boldsymbol{x}_{n,v,m})))$. Ec and the classification head Cn forms a binary classifier and minimize the calssification loss $\mathcal{L}_{bin}(y_n, Cn(Ec(\boldsymbol{x}_{n,v,m})))$. It can give predictions independently based on a trial instance $\boldsymbol{x}_{n,v,m}$ from one view v. We treat each trial instance $\boldsymbol{x}_{n,v,m}$ and the corresponding subject-wise label y_n as an independent data pair $(\boldsymbol{x}_{n,v,m}, y_n)$ to feed into the model for learning the binary classification task.

Each trial-type model has 3 hidden blocks with a convolutional layer, a batch normalization layer, a max-pooling layer, a ReLU



(a) Architecture

(b) Trial-type model

Figure 3: An overview of our architecture. (a) The long scans are splitted into four types of trials (AX, BX, AY, BY) and a model is built for each. Then an ensemble model is learned to concatenate embeddings extracted by the trial type models as input to train an ensemble predictor. The final predictions are supposed to match the subject-level labels. (b) The trial-type model are trained with both classification and reconstruction tasks. The embeddings of each training instance is extracted by the trial-type model for the training of the ensemble model.

layer, and 2 linear feedforward layers before the output layer. In the training of trial-type model, the autoencoder (Dc, Ec) and the classification head Cn of the trial-type models are trained simultaneously with reconstruction loss $\mathcal{L}_{rec}$ and binary classification loss $\mathcal{L}_{bin}$. The training terminates when each trial-type model reaches its best performance based on validation set performance. In the training of our ensemble model (see Figure 3b), the heads of the deep learners (Dc and Cn) are not used, instead, the embeddings are used. The encoder Ec acts as a feature extractor to encode the multi-instances of each type and feed the embedding into the ensemble model for training the ensemble model. To address the unbalanced view issue in it, we randomly sample the instances of each trial-type from the same subject instead of specifically designing a multi-instance fusion method. However, in the validation phase, we use the mean pooling method to do multi-instance fusion which is shown in subsection 4.3.

3.4 Training Using Transfer Learning

Since we only have a limited number (51) of subjects, this is a small data problem. To address this problem, we investigate transferring the learned knowledge from the source task model with plenty of training examples (AX) to the target models with limited training examples (AY, BX, BY). This is motivated due to the features important for prognosis will be similar between views.

We tried different transfer learning schemes and presents the results of our best practices in subsection 4.4. We found transferring both the convolutional layers and the feedforward layers is the best practice. Further, we studied the direction of transfer and found the most success when transferring from abundant training instance trial-type models to the trial-type models with scarce training instances. Table 3 summarizes these experiments.

3.5 Combining Trial Type Models

Once we have four well-trained trial-type models, we explored the approach to combining the knowledge from four trial-type models to make a subject-wise prediction. We consider the models as knowledge extractors and encode the instances using the intermediate output of the first linear layer after the convolutional layers of the models to ensemble the prediction. To learn instance-specific weights to facilitate dynamic trial-type model combination we fix the embedding vectors and train a MLP neural network with 3 linear layers and the leaky-relu activation function.

In the training phase, we randomly sample trials of all four trial-types from a scan. We construct the input feature vector by concatenation and to feed it into the ensemble model. This approach greatly enlarges the training set by randomly combining embedding vectors from different trial-types of the same subject. The number of combinations of the trial instances of different types of trials grows in polynomial with regarding the number of trial instances of a single time. In the validation phase, we use the multi-instance fusion methods in subsection 4.3 to obtain a vector representation out of all instances of a scan from each view. We create a single feature vector for each scan by concatenation and hence get a unique prediction for a scan. In Figure 4, We empirically showed that an ensemble network on the embedding vectors can combine the knowledge learned from multiple trial-type models and surpass the performance of every single trial-type model.

4 EXPERIMENTS AND RESULTS

In this section we discuss the experimental settings and results addressing the following questions:

- How do we collect and preprocess the data? What is our general experimental setup and validation method? (See subsection 4.1) This section focuses on reproducibility and can be skipped when first reading the paper.

- Since we have four trial-type models to train, what is the best strategy to transfer knowledge between trial-type models? (See subsection 4.2 and Table 3)
- Can multi-instance pooling methods improve the performance? (See subsection 4.3 and Table 4)
- Can an ensemble model help in combining views? (See subsection 4.4 and Table 4)
- Which frames are important to the prediction of the trained model on each type of event? (See subsection 4.5 and Table 5)

4.1 Data and Experimental Setup.

The data used is freely available on request (website will be disclosed upon publication) but not publicly available due to privacy requirements. Code will also be made available to document and reproduce experimental results.

Data Collection. Our functional images were acquired with a gradient-echo T2* blood oxygenation level-dependent (BOLD) contrast technique and were performed in a 1.5 T scanner (GE Healthcare). fMRI data were preprocessed using SPM8 [12] (Wellcome Department of Imaging Neuroscience, London). Briefly, images were slice-timing corrected, realigned, normalized to the template using and smoothed Gaussian kernel. All individual fMRI runs had small translational and rotational within-run movement and framewise displacement. All subjects had at least two fMRI runs surviving the criteria described in [26]. The minimum length of the scans is 560 and the maximum length of the scans is 1120. The time resolution of each frame is 2 seconds. The size of a 3D frame is $80 \times 96 \times 72$. Each frame is paired with a binary flag indication of out-of-range displacement caused by subjects' movement. There is a great displacement in the frame which may cause data analyzing error if the flag value is 1, otherwise, it's 0. In both the train and evaluation process, we remove the entire trial instance if any of the frames in it has a flag 1. The ratio of the one type of trial over trials is 70% AX, 10% AY, 12.5% BX, 7.5% BY.

Data Preprocessing. The difference between our work and the work [26] in data pre-processing is we converted the human brain BOLD measurements into grayscale 3D images. The BOLD measurements of voxels in human brain gray matter are first min-max clipped to the range of $[-100, 100]$, and then be applied with a modified sigmoid function to normalize to the pixel values within the range of $[0.25, 1]$. All the other pixels in the image coordinates that are not brain gray matters (such as white matters or skulls) are assigned value 0. Given the raw BOLD measurement of a voxel $I \in \mathbb{R}+$, the pre-processing can be expressed by the following equations,

$$x = \max(-100, \min(I, 100)) \tag{1}$$

$$\sigma_\beta(x) = \frac{1}{1 + \beta \times \epsilon^{-x}}, \quad -100 \leq x \leq 100 \tag{2}$$

where β is a hyper-parameter to choose. The choice of β should consider to maximize the entropy of the voxel values thus the resulted gray scale images carry the greatest amount of information from the raw BOLD measurement. We choose $\beta = 2.5$ through empirical studies.

Experimental Setup For all trial-type models, we use the sum of binary cross-entropy loss over trial instances as our binary classification loss function and the sum of mean square loss over trial instances as our reconstruction loss function. In all of experiments, stochastic gradient descent (SGD) is adopted as the training optimizer. The SGD optimizer has a base learning rate of 1×10^{-2} and momentum of 0.95. We decay the learning rate of each parameter group by 0.5 every 50 epochs with early stopping when the loss stops decreasing on the validation set.

Validation method As for validation, a lot of critiques towards existing machine learning studies on medical imaging is they split the data example from the same subject to both training set and validation set. There is a total of 51 subjects in our study and each subject can complete a task hundreds of times, which poses a threat to data leakage on subjects. In our experiments, we used five-folds cross-validation on *subject* level predictions to prevent data leakage and hence report inflated performance results. There is no shared subjects, trials or frames between the training set and the evaluation set. To evaluate the performance of the trial-type models (AX, BX, AY, BY), we did mean pooling to find the most popular prediction among the trial instances of each trial-type from a single scan. For the ensemble results, we concatenate the embedding encoded by the trial-type models as input to the ensemble model to make a subject-wise prediction. Our best practice reaches the accuracy of 75.6% ± 3.2% in comparison with the previous best practice 72.6% in [26].

4.2 Transfer Learning Schemes

Amongst all the trial-types, the AX trials account for 70% occurrences which is much greater than the occurrences of (BX, AY, BY) trials. This is so because this trial type is when the subject must press a response button and is the conditioning trial. Hence we use the trial-type model AX as our source model and transfer its knowledge to other trial-type models. One popular parameter transfer learning [38] practice of deep CNNs is to use the pre-trained model parameters on the source domain, freeze some low-level hidden layers without updating the parameters, and fine-tune the other layers on the target domain.

We first train AX model on the AX trial instances to create the source model. Then, we use parameters of this source model to initialize the parameters of the target model and try different transfer learning schemes by freezing a subset of the three convolutional layers and fine-tuning the feed-forward parameters. We show the results of varying transfer schemes in Table 3. The performance of fine-tuning all the parameters initialized by trial-type AX model is better than training from scratch without initialization or freezing some convolutional layers preceding fine-tuning the other layers. Based on these results, we use AX model as the source model to initialize all the other trial-type models and fine-tune on all the parameters for the following experiments.

4.3 Multi-instance Fusion For Prediction

A challenge with our multi-instance setting is that since each subject performs each task many times, there are many instances of each trial type for each subject on which to make a prediction. For

Table 3: Transfer learning schemes and results. We present results transferring the AX source model using a variety of schemes (rows) to other models (columns). If a layer is not frozen, it is fine-tuned. The base accuracy of the AX model is 72.6% and the base accuracy for the other models is in the first row. In all cases, the target model is initialized with the source model's parameters.

Schemes	Model AY (Target)	Model BX (Target)	Model BY (Target)
Train from scratch	62.7%	66.7%	54.9%
Fine-tune all parameters	**68.6**%	**70.5**%	**62.7**%
Freeze Only Conv1	68.6%	68.6%	60.8%
Freeze Only Conv1 & 2	64.8%	64.8%	56.8%
Freeze Only Conv1 & 2 & 3	58.8%	60.8%	54.9%

Table 4: Various methods of multi-instance fusion and multi-view combination for prediction. The three rows denotes the multi-instance (MI) fusion methods: prediction majority voting, max pooling, mean pooling. Columns 2 to 5 denote this fusion methods applied across instances for a single trial-type. Columns 6 to 8 columns denote the multi-view (MV) fusion method applied across model types.

Fusion methods	AX model	AY model	BX model	BY model	Majority voting	Weighted average	Ensemble model
Vote aggregation	66.7%	60.7%	62.7%	58.4%	64.6%	66.7%	68.6%
Max pooling	74.5%	64.6%	68.6%	58.4%	68.6%	72.6%	72.6%
Mean pooling	72.6%	68.6%	70.5%	62.7%	68.6%	72.6%	**75.6**%

example, if a subject sees 70 AX, 12 BX, 10 AY, and 8 BY trials there will be 100 instances on which to make a prediction.

To fuse the predictions of many instances to get an overall level prediction, there are two popular ways: voting method and pooling methods. The voting method is to count the majority prediction of all instances in a bag as the overall prediction whilst the pooling method [30] uses a pooling layer M to pool the embeddings of all instances into one embedding. In the voting method, the whole network including the feedforward layers is used as an expert model to directly give predictions of instances. In the pooling method, the fused embedding acts as the input to the feed-forward layers for the prediction of a bag. In this paper, we use two popular MI pooling methods: max pooling and mean pooling.

$$\text{max pooling: } M(\boldsymbol{x}_{i|i=1\dots m}) = \max_i \boldsymbol{x}_{i|i=1\dots m} \tag{3}$$

$$\text{mean pooling: } M(\boldsymbol{x}_{i|i=1\dots m}) = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_{i|i=1\dots m} \tag{4}$$

We show our experimental results in Table 4 and the best performance method is the mean pooling method which we use going forward.

4.4 Trial-type Models Ensemble

The result of the previous step was to find a single embedding for each trial. Here we explore how to combine them. The high-level view of our proposed work is to use multiple trial-type models to learn task-related features and then combine the learned knowledge to get a final subject-wise prediction. We design a MLP ensemble network that takes in the last linear layer before the output layer in the trial-type models as inputs (we call them embedding vectors of trials), and dynamically ensemble the learned knowledge of the embedding vectors. These vectors are then passed to 3 fully connected layers (512, 64, and 8 respectively) of leaky-relu units. The output is a one-hot encoding representation for the subject-wise predicted label.

We compare the above approach with two baselines "majority voting" and "weighted average" described in Table 4. From the four trained trial-type models, we can get the subject-wise prediction through the MI fusion methods in subsection 4.3 from each trial-type models. The "majority voting" means treating each trial-type model as an expert to vote for the most popular prediction. The "weighted average" means to weighted average the prediction of each trial-type model by the number of trials of each type to get the subject-wise prediction. The results are shown in Table 4. We find that the trial-type model performances are the bottlenecks of the performance of "majority voting" and "weighted average". We repeated the experiments from trial-type model training to ensemble model evaluation 10 times with random seeds to present the error bars in Figure 4. We found the ensemble model can raise the performance of each trial-type model by a large margin and we adopt the ensemble model for the following experiments.

4.5 Frame Importance

Recall the input to each trial-type model is six frames (3 associated with the cue and 3 with the probe). To understand which frames are important for the model to make predictions, we designed the blocking studies on the input frames. At evaluation time, we block one of the frames (from cue or probe) by making all pixel values to be zeros in the 3D image and perform the cross-validation experiment using such input. A blocked frame which causes a large drop in accuracy is indicative of an important frame.

The results of cues and probes are shown in Table 5. To illustrate the notation, the "Cue" and "Probe" columns are specifying the three frames in each type of event. The "Block #index" means, we

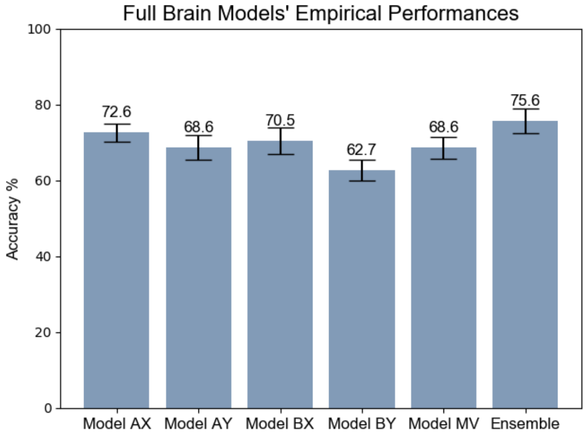


Figure 4: Performances of trial-type models and an ensemble model. The four trial-type models (AX, BX, AY, BY) are evaluated by mean pooling across all the instances of the same model type. The performance of 'Model MV' is evaluated on the average prediction across the four trial-type models. 'Ensemble' is the ensemble model on the embedding vectors from trial-type models.

make all of the pixels in this indexed frame of this event to be zeros and leave the other two frames unchanged. The numbers shared by multiple columns are accuracies of the controlled trial-type models in same the row. Compared with the baseline performance, the frames that are most significant to the prediction for each trial type are frame 2 of CueA in AX , frame 3 of CueB in BX, frame 2 of CueA in AY, frame 3 of CueB in BY.

5 DISCUSSION

We did exhaustive experiments on the effects of every sub-module. In the experiment in Table 3, we found re-using the learned parameters from trial-type model A to initialize the other trial-type models greatly raises the performances of other trial-type models. However, the performances deteriorate when parameters of more layers from bottom to top are frozen. One possible reason is the lower-level features are more transferable than the higher-level features. It implies the whole brain activities change a lot on seeing different events but the local activities are very reproducible. The result of Figure 4 shows the ensemble model easily breaks the performance bottleneck of each trial-type model, which indicates the combination of different types of trials is meaningful. The result of Table 4 suggests utilizing the mean pooling to fuse the multi-instances of a trial-type is much better than voting aggregation method where every noisy instance may directly contribute to the final prediction. The performance mean pooling method also surpasses the performance of max pooling method, especially in the trial-type models with fewer training instances. One possible reason is the maximum values of embedding may overfit the dataset by some out-of-distribution instance. Table 5 reveals how the trial-type models put attention on the time series. The peak BOLD signals of the human brain seeing cues and probes are variant between 4-8 seconds after the event occurs. The frames that the models focus thus roughly align with the peak BOLD response as expected.

Our work opens up a gate to study t-fMRI data and there are some future works waiting for people to explore. First of all, the explanation aspect of the work is in great need. Following ([1]), we can explore the following questions: What anatomical regions are important to the model's prediction? What are the variations between the different types of trials? What are salient patterns for the true positive, true negative results? Why are some of the scans classified incorrectly? Can we generate neurobiological relevance fingerprinting like [3] did? Moreover, to study the generalizability of the model and make it more robust to machine noises is essential. One crucial limitation for the deeply learned model to be widely used in real clinical practices is the data collected from different sources may introduce a high volume of noises. The models trained on the data collected from one machine may not apply to the data collected from other machines. Adversarial domain adaptation [8] is a decent approach to learn a model resisting domain discrepancy. Furthermore, it's worth trying to explore more complex ensemble methods and transfer learning techniques. Different from the strict causal relationship in real life time-vary data such as natural languages or movies, the trials are independent of each other. The former trials have little impact on the later trials. In this work, we ensemble the models by types of trials. A more straightforward idea is to ensemble the models by trials directly. This may lead to other challenges such as overfitting to the trial type with the most occurrences. Maximum Mean Discrepancy (MMD) [10] is a kernel two-sample test to evaluate the distance between the source and target for transfer learning, which helps the design of more advanced transfer learning methods [20].

6 RELATED WORK

The Deep Learning Studies on fMRI Data. Medical imaging analysis has seen considerable development over the last several decades. Thanks to the rapid progress, particularly convolutional neural networks (CNNs) [15], towards medical imaging analysis [31]. Impressive performance comparable to human experts on image classification, object detection, segmentation, registration, and other tasks [18] has occurred. As one of the most popular modalities, most of the previous works are on resting-state fMRI (rs-fMRI) data. [23] used convolutional neural networks to classify Alzheimer's brain from the normal healthy brain. [2] proposed an unsupervised matrix tri-factorization to discover an underlying network that consists of cohesive spatial regions (nodes) and relationships between those regions (edges) for brain imaging data. Such works on rs-fMRI focus on exploring the intrinsically functionally segregation or specialization of brain regions/networks [19] but are limited on identifying spatiotemporal brain patterns that are functionally involved in specific task performance.

The t-fMRI Studies. Recently, the t-fMRI analysis [28] is attracting more and more attention for its ability to connect human activities to brain functioning. In the work of [24], the subjects in the study are asked to read a chapter from a novel while the fMRI scans recording their brain activities are conducted. They fine-tuned a pre-trained BERT model to map the natural language to brain fMRIs. [22] used time-varying persistence diagrams to

Table 5: Frame importance by the blocking experiment. Recall the input into the deep learner for each probe and cue is three temporally adjacent frames. Here we aim to determine the most important by blocking (assign zeros to all voxels) various frames during the prediction of the performances of the trial-type models' changes. The blocked frame that most decreases performance is the most important.

Trial-types	w/o blocking	Cue			Probe		
		Block 1	Block 2	Block 3	Block 1	Block 2	Block 3
AX	72.6%	66.7%	**60.8**%	64.8%	**64.8**%	64.8%	66.7%
BX	70.5%	64.8%	62.8%	**60.8**%	66.7%	**64.8**%	68.6%
AY	68.6%	60.8%	**56.9**%	58.8%	**60.8**%	62.7%	66.7%
BY	62.7%	56.9%	54.9%	**52.9**%	58.8%	**56.9**%	60.8%

represent the human brain activities when the subjects are watching the movie. [25] studies deep image reconstruction by decoding fMRI into the hierarchical features of a pre-trained deep neural network (DNN) for the same input image. The studies in schizophrenia diagnosis utilizing cognitive control tasks suffered from either a small sample size or modest classification performance [37]. All these t-fMRI settings are different from the AX-CPT setting for they don't have multiple types of repeated independent clinical trials to result in one combined evaluation. Instead, their tasks are sequence-to-sequence guided by the inputs such as a series of images and natural languages.

The AX-CPT t-fMRI Studies. The AX-CPT task is a clinical test on reactive and proactive control processes to identify human cognitive control deficits [16]. With modest classification accuracy, the first schizophrenia diagnosis study [35] on the fMRI scans conducted while the cohort subjects completed the AX-CPT task suggests an application to discriminate disorganization levels among the patients. [27] began the studies on the prognosis of treatment of schizophrenia by analyzing the task-fMRI data. The task-fMRI scans of 82 subjects with psychotic disorders were collected and small regions of interest (ROI) were extracted from the scans for the study. The following work of [26] compared machine and naive deep learning-based algorithms for the prediction of clinical improvement in psychosis with the same task-fMRI data. It achieved ROI voxelwise accuracy of 62.4% using a logistic regression model and 72.6% using a multi-layer perceptron model which we used as the baseline for our work. These works highly rely on hand-crafted regions of interest segmentation and they are also analyzing the task-fMRI data on the average activation of some selected keyframes in the scans, which may contribute to a great amount of information loss. In our work, we use the same source of data [3]as the two above works ([26, 27]) on prognosis but only the 51 scans in the 1st protocol are included. Different from the above works, we don't need any handcrafted ROI segmentation.

The Multi-view Learning and Multi-instance Learning. Multi-view learning ([33]) and multi-instance learning ([5]) are prevalent in practice; for example, the text contents and the links are two views of a web page; the gene sub-sequences can seen as the multiple instances in a bag of a chromosome. Our approach fits the general multi-view multi-instance learning definition but still shows explicit differences from the other latest works in this scope. [6, 17, 21, 29] are non-deep matrix factorization or graph representation methods which are not applicable in very high dimensional feature space. In contrast to our multi-view multi-instance setting where a bag of instances represents one view of an example, [32, 34, 36] study multi-instance learning on bag of instances where each instance has multiple views. All these subtle differences invalidate their approaches to be used in our setting.

7 CONCLUSION

In this study, we first bring forward some intrinsic challenges of existing methods on task-fMRI data. These challenges put limits on the utilization of some existing methods such as co-activation matrix, multi-view co-training methods, and variations of recurrent neural networks. Through scrutiny, we come up with a novel multi-view multi-instance learning setting that perfectly fits the task. Then, we proposed a deep learning architecture that is able to extract task-specific features from different types of trials through trial-type models and concatenate the features to make subject-wise predictions through an ensemble model. The CNN-based trial-type models are trained on varied numbers of repeated trials. Transfer learning is used between different trial-type models, which enables the knowledge injection between them. Transferring the parameters from the most frequent trial-type model to other trial-type models by initialization and fine-tuning has a tremendous impact on the performance. Our deep architecture involving multiple trial-type models and an ensemble model is an adaptable new paradigm in task-fMRI analysis with multiple types of repeated trials.

ACKNOWLEDGMENTS

This work was supported in part by NSF Grant IIS-1910306 titled: "Explaining Unsupervised Learning: Combinatorial Optimization Formulations, Methods and Applications" and NIH grant R01MH122139 titled: "Pathophysiological Biomarkers of Treatment Response in Early Psychosis"

[3] The data is freely available after requests but can not be publicly posted due to privacy concerns.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] Zilong Bai, Peter Walker, Anna Tschiffely, Fei Wang, and Ian Davidson. 2017. Unsupervised network discovery for brain imaging data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 55–64.

[3] Moritz Böhle, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. 2019. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Frontiers in aging neuroscience* (2019), 194.

[4] Valerie Bonnelle, Timothy E Ham, Robert Leech, Kirsi M Kinnunen, Mitul A Mehta, Richard J Greenwood, and David J Sharp. 2012. Salience network integrity predicts default mode network function after traumatic brain injury. *Proceedings of the National Academy of Sciences* 109, 12 (2012), 4690–4695.

[5] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition* 77 (2018), 329–353.

[6] Yukuo Cen, Xu Zou, Jianwei Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Representation learning for attributed multiplex heterogeneous network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1358–1368.

[7] Nicolas A Crossley, Andrea Mechelli, Petra E Vértes, Toby T Winton-Brown, Ameera X Patel, Cedric E Ginestet, Philip McGuire, and Edward T Bullmore. 2013. Cognitive relevance of the community structure of the human brain functional coactivation network. *Proceedings of the National Academy of Sciences* 110, 28 (2013), 11583–11588.

[8] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.

[9] Michael D Greicius, Ben Krasnow, Allan L Reiss, and Vinod Menon. 2003. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences* 100, 1 (2003), 253–258.

[10] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13, 1 (2012), 723–773.

[11] Elseline Hoekzema, Susana Carmona, J Antoni Ramos-Quiroga, Vanesa Richarte Fernandez, Rosa Bosch, Juan Carlos Soliva, Mariana Rovira, Antonio Bulbena, Adolf Tobena, Miguel Casas, et al. 2014. An independent components and functional connectivity analysis of resting state fMRI data points to neural network dysregulation in adult ADHD. *Human brain mapping* 35, 4 (2014), 1261–1272.

[12] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. 2008. Identifying natural images from human brain activity. *Nature* 452, 7185 (2008), 352–355.

[13] Anthony P King, Stefanie R Block, Rebecca K Sripada, Sheila Rauch, Nicholas Giardino, Todd Favorite, Michael Angstadt, Daniel Kessler, Robert Welsh, and Israel Liberzon. 2016. Altered default mode network (DMN) resting state functional connectivity following a mindfulness-based exposure therapy for posttraumatic stress disorder (PTSD) in combat veterans of Afghanistan and Iraq. *Depression and anxiety* 33, 4 (2016), 289–299.

[14] Walter Koch, Stephan Teipel, Sophia Mueller, Jens Benninghoff, Maxmilian Wagner, Arun LW Bokde, Harald Hampel, Ute Coates, Maximilian Reiser, and Thomas Meindl. 2012. Diagnostic power of default mode network resting state fMRI in the detection of Alzheimer's disease. *Neurobiology of aging* 33, 3 (2012), 466–478.

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[16] Tyler A Lesh, Andrew J Westphal, Tara A Niendam, Jong H Yoon, Michael J Minzenberg, J Daniel Ragland, Marjorie Solomon, and Cameron S Carter. 2013. Proactive and reactive cognitive control and dorsolateral prefrontal cortex dysfunction in first episode schizophrenia. *NeuroImage: Clinical* 2 (2013), 590–599.

[17] Bing Li, Chunfeng Yuan, Weihua Xiong, Weiming Hu, Houwen Peng, Xinmiao Ding, and Steve Maybank. 2017. Multi-view multi-instance learning based on joint sparse representation and multi-view dictionary learning. *IEEE transactions on pattern analysis and machine intelligence* 39, 12 (2017), 2554–2560.

[18] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42 (2017), 60–88.

[19] Nikos K Logothetis. 2008. What we can do and what we cannot do with fMRI. *Nature* 453, 7197 (2008), 869–878.

[20] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2017. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*. PMLR, 2208–2217.

[21] Cam-Tu Nguyen, Xiaoliang Wang, Jing Liu, and Zhi-Hua Zhou. 2014. Labeling complicated objects: Multi-view multi-instance multi-label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28.

[22] Bastian Rieck, Tristan Yates, Christian Bock, Karsten Borgwardt, Guy Wolf, Nicholas Turk-Browne, and Smita Krishnaswamy. 2020. Uncovering the topology of time-varying fMRI data using cubical persistence. *Advances in neural information processing systems* 33 (2020), 6900–6912.

[23] Saman Sarraf and Ghassem Tofighi. 2016. Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks. *arXiv preprint arXiv:1603.08631* (2016).

[24] Dan Schwartz, Mariya Toneva, and Leila Wehbe. 2019. Inducing brain-relevant bias in natural language processing models. *Advances in neural information processing systems* 32 (2019).

[25] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. 2019. Deep image reconstruction from human brain activity. *PLoS computational biology* 15, 1 (2019), e1006633.

[26] Jason Smucny, Ian Davidson, and Cameron S Carter. 2021. Comparing machine and deep learning-based algorithms for prediction of clinical improvement in psychosis with functional magnetic resonance imaging. *Human brain mapping* 42, 4 (2021), 1197–1205.

[27] Jason Smucny, Tyler A Lesh, and Cameron S Carter. 2019. Baseline frontoparietal task-related BOLD activity as a predictor of improvement in clinical symptoms at 1-year follow-up in recent-onset psychosis. *American Journal of Psychiatry* 176, 10 (2019), 839–845.

[28] Jason Smucny, Ge Shi, and Ian Davidson. 2022. Deep Learning in Neuroimaging: Overcoming Challenges With Emerging Approaches. *Frontiers in Psychiatry* 13 (2022). https://doi.org/10.3389/fpsyt.2022.912600

[29] Keyao Wang, Jun Wang, Carlotta Domeniconi, Xiangliang Zhang, and Guoxian Yu. 2020. Differentiating isoform functions with collaborative matrix factorization. *Bioinformatics* 36, 6 (2020), 1864–1871.

[30] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. 2018. Revisiting multiple instance neural networks. *Pattern Recognition* 74 (2018), 15–24.

[31] Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, Olivier Colliot, et al. 2020. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical image analysis* 63 (2020), 101694.

[32] Yuying Xing, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Zili Zhang, and Maozu Guo. 2019. Multi-view multi-instance multi-label learning based on collaborative matrix factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 5508–5515.

[33] Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. 2021. Deep multi-view learning methods: A review. *Neurocomputing* 448 (2021), 106–129.

[34] Yuanlin Yang, Guoxian Yu, Carlotta Domeniconi, and Xiangliang Zhang. 2021. Deep Multi-type Objects Muli-view Multi-instance Multi-label Learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM, 486–494.

[35] Jong H Yoon, Danh V Nguyen, Lindsey M McVay, Paul Deramo, Michael J Minzenberg, J Daniel Ragland, Tara Niendham, Marjorie Solomon, and Cameron S Carter. 2012. Automated classification of fMRI during cognitive control identifies more severely disorganized subjects with schizophrenia. *Schizophrenia research* 135, 1-3 (2012), 28–33.

[36] Ye Yuan, Guangxu Xun, Kebin Jia, and Aidong Zhang. 2018. A multi-view deep learning framework for EEG seizure detection. *IEEE journal of biomedical and health informatics* 23, 1 (2018), 83–94.

[37] Ling-Li Zeng, Huaning Wang, Panpan Hu, Bo Yang, Weidan Pu, Hui Shen, Xingui Chen, Zhening Liu, Hong Yin, Qingrong Tan, et al. 2018. Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. *EBioMedicine* 30 (2018), 74–85.

[38] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proc. IEEE* 109, 1 (2020), 43–76.