OXFORD

Psychoradiology, 2023, **3**, 1–10

DOI: 10.1093/psyrad/kkad011
Advance access publication date: 7 July 2023
Research Article

Prediction of cognitive scores by joint use of movie-watching fMRI connectivity and eye tracking via Attention-CensNet

Jiaxing Gao[1], Lin Zhao[2], Tianyang Zhong[1], Changhe Li[1], Zhibin He[1], Yaonei Wei[1], Shu Zhang[3], Lei Guo[1], Tianming Liu[2], Junwei Han[1], Xi Jiang[4] and Tuo Zhang[1,*]

[1]School of Automation, Northwestern Polytechnical University, Xi'an 710072, China
[2]Cortical Architecture Imaging and Discovery Laboratory, Department of Computer Science and Bioimaging Research Center, The University of Georgia, Athens, GA 30602, USA
[3]School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
[4]School of Life Science and Technology, MOE Key Laboratory for Neuroinformation, University of Electronic Science and Technology of China, Chengdu 611731, China
*Correspondence: Tuo Zhang, tuozhang@nwpu.edu.cn

Abstract

Background: Brain functional connectivity under the naturalistic paradigm has been shown to be better at predicting individual behaviors than other brain states, such as rest and doing tasks. Nevertheless, the state-of-the-art methods have found it difficult to achieve desirable results from movie-watching paradigm functional magnetic resonance imaging (mfMRI) -induced brain functional connectivity, especially when there are fewer datasets. Incorporating other physical measurements into the prediction method may enhance accuracy. Eye tracking, becoming popular due to its portability and lower expense, can provide abundant behavioral features related to the output of human's cognition, and thus might supplement the mfMRI in observing participants' subconscious behaviors. However, there are very few studies on how to effectively integrate the multimodal information to strengthen the performance by a unified framework.

Objective: A fusion approach with mfMRI and eye tracking, based on convolution with edge-node switching in graph neural networks (CensNet), is proposed in this article.

Methods: In this graph model, participants are designated as nodes, mfMRI derived functional connectivity as node features, and different eye-tracking features are used to compute similarity between participants to construct heterogeneous graph edges. By taking multiple graphs as different channels, we introduce squeeze-and-excitation attention module to CensNet (A-CensNet) to integrate graph embeddings from multiple channels into one.

Results: The proposed model outperforms those using a single modality and single channel, and state-of-the-art methods.

Conclusions: The results indicate that brain functional activities and eye behaviors might complement each other in interpreting trait-like phenotypes.

Keywords: functional connectivity; naturalistic stimulus; eye movement; CensNet; attention

Introduction

Complex cognition could be a psychiatric trait that differentiates high-order species, including human beings, from others (Al-Aidroos *et al.*, 2012; Baars & Gage, 2010; Barack & Krakauer, 2021; Diamond, 2013; Gallistel & King, 2011; Harvey, 2022; Lezak *et al.*, 2004). Cognitive decline is usually an obvious manifestation of the progression of many psychiatric diseases, such as Alzheimer's disease, Parkinson's, and depression (LeMoult & Gotlib, 2019; Pick *et al.*, 2019; Stern, 2012; Sun *et al.*, 2020; Wolters *et al.*, 2019), and is a key determinant of a patient's quality of life and independence. Therefore, it has long been intriguing as to how cognition works in the human brain for decades in multiple disciplines, including cognitive neuroscience and psychoradiology (Baars & Gage, 2010; Bressler & Menon, 2010). To date, much of the current knowledge leads to a consensus that cognitive brain function comes from the large-scale brain organization, which is the orchestration of local and remote cortical areas by means of a densely connected brain network (Axer & Amunts, 2022; Bressler & Menon, 2010; Bullmore & Sporns, 2012; Thiebaut de Schotten & Forkel, 2022; Van Den Heuvel & Sporns, 2011). With the advent of brain imaging techniques, including functional magnetic resonance imaging (fMRI), it has become possible to record brain activity and estimate brain networks *in vivo*, granting them an "objective" observer of the outcomes of a participant's brain. On this basis, there is a growing interest in leveraging imaging-based whole-brain functional connectivity to predict non-brain-imaging phenotypes, including a variety of cognitive and behavioral measures (He *et al.*, 2020). Importantly, the patterns of the functional connectivity could serve as a "fingerprint" to identify individuals (Gao *et al.*, 2022). A well-performed predictive model could result in potential functional connectivity biomarkers to distinguish healthy statuses from abnormal ones, or monitor disease progress (Liu *et al.*, 2017).

Received: 5 April 2023; **Revised:** 14 June 2023; **Accepted:** 6 July 2023

Functional connectivity can manifest in various brain states. While resting and task states are commonly used paradigms (Deco *et al.*, 2011; Eickhoff *et al.*, 2020; Fox & Raichle, 2007; Grady *et al.*, 2021; Kannurpatti *et al.*, 2012; Malinen *et al.*, 2007; Yang *et al.*, 2020), movie-watching paradigms offer more immersive, life-like content (Barch *et al.*, 2013; Huijbers *et al.*, 2017; Li *et al.*, 2019; Sonkusare *et al.*, 2019), as they have been proposed to not only capture the common functional connectivity component by reducing inter-individual variability, but also to enhance the stability, amplification, and trait-like features of the remaining variability (Eickhoff *et al.*, 2020). Hence, a movie-watching paradigm outperforms resting and task states in high-order behavior prediction, including cognition, and could currently be the possible upper bound of non-brain-imaging phenotype prediction paradigms (Finn & Bandettini, 2021). Even so, the precision from fMRI using the movie-watching paradigm (mfMRI) still remains limited, partly due to the small dataset size (He *et al.*, 2020). Although recent studies have demonstrated that large-scale datasets can improve prediction performance by providing more training samples and a diverse range of inter-correlated phenotypic measures that are likely to be correlated with, but not identical to, a unique phenotype from a boutique study (He *et al.*, 2022), the fact is that improving the prediction performance by means of increasing dataset size is only possible for resting state and a large-scale collection of mfMRI datasets is challenging and costly.

To improve the performance of mfMRI predictions on small datasets, a couple of strategies are proposed. First, incorporating other physical measurements into the prediction method may enhance accuracy. A joint use of fMRI data on a participant's thoughts and their response to stimuli, along with behavioral measurements such as eye tracking, could provide a more comprehensive picture of the participant's cognitive and phenotypical measures. Eye tracking, in particular, is becoming increasingly popular due to its portability and low cost, and its features such as pupil size, fixation duration, and saccade patterns have been linked to cognitive and phenotypical measures (Carter & Luke, 2020; Hess & Polt, 1960; Lim *et al.*, 2020; Lohse & Johnson, 1996). Hence, eye-movement behavior might supplement the fMRI derived brain activities in monitoring participants' attention and task compliance and observing their subconscious traits (Beatty & Lucero-Wagoner, 2000; Einhäuser, 2017; Laeng & Alnaes, 2019; Laeng *et al.*, 2012; Mathôt, 2018; Son *et al.*, 2020). Second, increasing the number of video clips watched by the same group of participants may be an efficient way to improve prediction accuracy, given the limited number of participants. Third, an effective representation of a participant's relationship, where inter-individual variation is well described, is needed. The graph has been successfully used to this aim, where nodes represent individuals and edges represent cross-participant similarity. For example, individual brain activity features have been used as nodal features whereas demographic and behavioral measurement similarity have been used to define edges (Gao *et al.*, 2022), and graph convolution networks can embed brain activity features of a cohort of participants and estimate a mapping of these features to cognitive scores. This mapping can then be propagated to other nodes to predict their cognitive scores.

In short, although it has been demonstrated that a nonlinear model is suitable for the prediction of cognitive grade, a further improvement of prediction performance confronts two technical challenges: (i) how to incorporate complementary edge features with nodal features to generate both node and edge embeddings for graphs, and (ii) how to integrate embeddings of graphs with heterogeneous topologies on the same set of nodes for classification or regression tasks. To achieve this, we propose Attention-CensNet (A-CensNet), an extension of the convolution with edge-node switching graphic neural network (CensNet) (Jiang *et al.*, 2019). In A-CensNet, participants are represented with nodes, and mfMRI-derived functional connectivity is used to represent nodal features. Gaze trajectories derived from eye tracking and temporal variation in pupil size are used to measure the similarity between participants and create a set of heterogeneous edges. Each of these graphs is treated as an independent channel, and CensNet is used to learn both node and edge embeddings. The squeeze-and-excitation attention module (SENet) (Hu *et al.*, 2018) is then applied to combine the node-edge embeddings from multiple channels into a hybrid graph, on which the final round of node embedding is performed. It should be noted that the same cohort is exposed to different movie inputs, resulting in additional channels from mfMRI and eye-tracking data.

The following sections are structured as follows: first, we introduce the dataset and preprocessing steps. Next, we provide an overview of CensNet and SENet, followed by our proposed A-CensNet and its application to our task. We then present results from comparative and ablation studies on prediction accuracy (area under curve, AUC) to demonstrate the effectiveness of our multiple channel integration strategy and the superior performance of A-CensNet compared to other methods.

Related Studies

Linear regression, is widely used to fit brain features to cognitive scores. However, its performance has been dwarfed by nonlinear methods, especially in the scenarios where the relationship between samples and features is far from linear, including the personal trait prediction application (Finn & Bandettini, 2021; Gao *et al.*, 2022).

FNN (fully connected neural networks) (He *et al.*, 2020), a generic class of feedforward neural networks, is mainly composed of an input layer, hidden layer, and output layer; the hidden layer consists of several fully connected layers and nonlinear activation functions.

BrainNetCNN (Kawahara *et al.*, 2017), a convolutional neural network (CNN) framework, consists of three layers: the edge-to-edge, edge-to-node, and node-to-graph layers. After the first three layers, BrainNetCNN using a fully connected layer as same as FNN. Both FNN and BrainNetCNN have been used to predict individual phenotypes (He *et al.*, 2020), and have achieved better results than linear regression. However, in these methods, participants were taken as independent samples. Integration of multiple features was limited in the feature space.

Since the data structure can be described effectively by a graph in many applications, graph neural networks, such as graph convolutional networks (GCN) (Defferrard *et al.*, 2016), have been developed to implement deep representation in non-Euclidean domains to adapt to a more effective way in describing the relationships (edges) of objects (nodes). In Gao *et al.* (2022), eye tracking information and functional connectivity was used to define edges and nodes, respectively. This scheme of integrating heterogenous features was demonstrated to be effective in increasing personal trait prediction. Nevertheless, the potential of GCN for assigning heterogenous features to nodes and edges has been underestimated.

CensNet was proposed to embed both nodes and edges to a latent feature space (Jiang *et al.*, 2019), such that edges not only serve to construct graph topology but are also fully involved in feature embedding and fusion. Therefore, given our need in this study for multiple eye tracking information to yield multiple graphs of heterogenous edges, we adopt CensNet as the basic algorithm to investigate the possibility of feature integration on the graph-level.

Materials and Methods

Dataset

The Human Connectome Project (HCP) 7T release acquired movie-watching fMRI and resting-state fMRI data on a 7 Tesla Siemens Magnetom scanner (Griffanti *et al.*, 2014). Two of the four scan sessions, Movies 2 and 3, were selected for analysis. Imaging parameters were as follows: TR = 1000 ms, TE = 22.2 ms, flip angle = 45°, FOV = 208 × 208 mm, matrix = 130 × 130, spatial resolution = 1.6 mm^3, number of slices = 85, multiband factor = 5. The resting-state run consisted of 900 time points. During Movies 2 and 3, participants viewed four and five video clips, respectively, each separated by five 20 s of rest sessions. Eye tracking data were collected using an EyeLink S1000 system with a 1000 Hz sampling rate. The HCP offers numerous phenotypic measures from various domains. This work focuses on measures related to cognition, which has been a common interest in previous studies (Finn & Bandettini, 2021). After quality control, data from 81 participants were analyzed.

Preprocessing

The fMRI data underwent preprocessing using the minimal preprocessing pipeline for the HCP (Glasser *et al.*, 2013). This involved motion correction, distortion correction, high-pass filtering, nonlinear alignment to MNI template space, and regression of 24 framewise motion estimates as well as confound timeseries identified through independent components analysis (Griffanti *et al.*, 2014). The signals were then mapped to the grayordinate system, which consisted of 64 000 vertices on the cortical surface and 30 000 subcortical voxels for each individual. Within-participant cross-modal registration and cross-participant registration were used to warp the grayordinate vertices and volumetric voxels to the same space, ensuring cross-participant correspondence of the associated fMRI signals.

This study focuses on cortical regions and thus excludes subcortical areas from analysis. We used the Destrieux atlas (Destrieux *et al.*, 2010) to parcellate the cortical surface into 75 areas per hemisphere. The mean fMRI signal was calculated by averaging over vertices within each cortical area, and a 150-by-150 functional connectivity matrix was constructed using the Pearson correlation between these average signals (green panel in Fig. 1). Specifically, we began by transforming the *z* to *r* correlation values with a hyperbolic tangent function, which scaled them between −1 and 1. For each row in the matrix, the values of the top 10% (one of the adjustable hyperparameters) of connections were retained, whereas all others were zeroed. The remaining connections were almost all positive, except for ∼5000 (<10% of all connections) with negative values from 23 voxels. The voxels with negative connections were located in ventral subcortical regions. We are not interested in these areas, and so we also zeroed the connection. Negative correlations were zeroed out and 90% of the lowest positive correlations were removed. The upper triangular matrix was then converted into a vector and used as the functional feature.

For the eye-tracking data, we used the time stamps to extract effective data points and synchronized eye behavior features across participants. Blink sessions were not considered.

To account for potential correlations between phenotypic measures within the “cognition” domain, we performed principal components analysis on these measures (Finn & Bandettini, 2021). The same principal components analysis strategy was applied to both the training and testing sets, using the means and standard deviations of the training set. We then used the first principal component to classify participants into four groups based on their scores. It is important to note that participant number balance is taken into consideration when selecting thresholds for grouping. Each group was assigned a label $l \in \mathscr{L}$.

Construction of graphs

Supposing we have a dataset of *M* participants, our objective is to assign each participant a cognitive group label *l*. We construct a graph $\mathscr{G} = \{\mathscr{V}, \mathscr{E}, A\}$ to represent the entire cohort as shown in Fig. 1, where $v \in \mathscr{V}$ is a node of the graph, the participants in this work. Edges *Es* as well as the adjacent matrix *A* encode the similarity between participants. On this graph structure, only a subset of nodes is labeled (e.g. *M* labeled nodes out of *N* nodes in total), leaving the rest of the nodes unlabeled (*N-M* nodes). Our goal is to assign each unlabeled node a label, a cognitive level in this work, in a semi-supervised fashion through the use of a GCN trained on the subset of labelled graph vertices. Intuitively, label information will be propagated over the graph under the assumption that nodes connected with high-edge weights are more comparable and similar (these edges provide a non-grid neighborhood for convolution), such that a similar label is more likely to be propagated to it.

In this work, mfMRI derived functional connectome is used as the node feature. Two sets of edges are defined on eye tracking data. Edge (i) is similarity between gaze trajectories of two participants, in which the gaze trajectory of participant *v* is denoted by g^v, which is a 2 × *t* vector. Its two rows record the coordinates of *x*- and *y*-dimensions on the screen. We use 2D Pearson correlation to measure the similarity of eye movement trajectories from two participants (Gao *et al.*, 2022). For edge (ii), similarity between temporal variation of pupil size, the pupil size along the time line of participant *v*, p^v, is a 1D vector. Likewise, the Pearson correlation coefficient between two vectors is defined as the edge weight.

Classification of population via A-CensNet

Basics of CensNet. As only a subset of nodes in the graph are labeled, CensNet is trained to propagate these labels throughout the entire graph. The implementation of CensNet has been previously described (Jiang *et al.*, 2019), and we provide a summary next:

For spectral graph convolution, normalized graph Laplacian of a graph $\mathscr{G} = \{\mathscr{V}, \mathscr{E}, A\}$ is computed: $\mathscr{L} = I_N - D^{-1/2} A D^{-1/2}$ where I_N is the identity matrix and D is the diagonal degree matrix. One of the important steps is the layer-wise propagation rule based on an approximated graph spectral kernel as follows:

$$H^{l+1} = \sigma\left(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} H^l W^l\right) \tag{1}$$

where $\tilde{A} = A + I_N$ and $\tilde{D}$ is the degree matrix, H^l and W^l are the hidden feature matrix and learnable weight of the *l*th layer.

Building on this foundation, the proposed CensNet model incorporates both node and edge convolution layers. As shown in Fig. 1, the graph can be represented by both a node-center version (yellow box) and an edge-center version (green box) for convenience. In the node convolution layer, the embedding of nodes in the white box is updated while the edge adjacency matrix and edge features in the green box remain unchanged. A similar

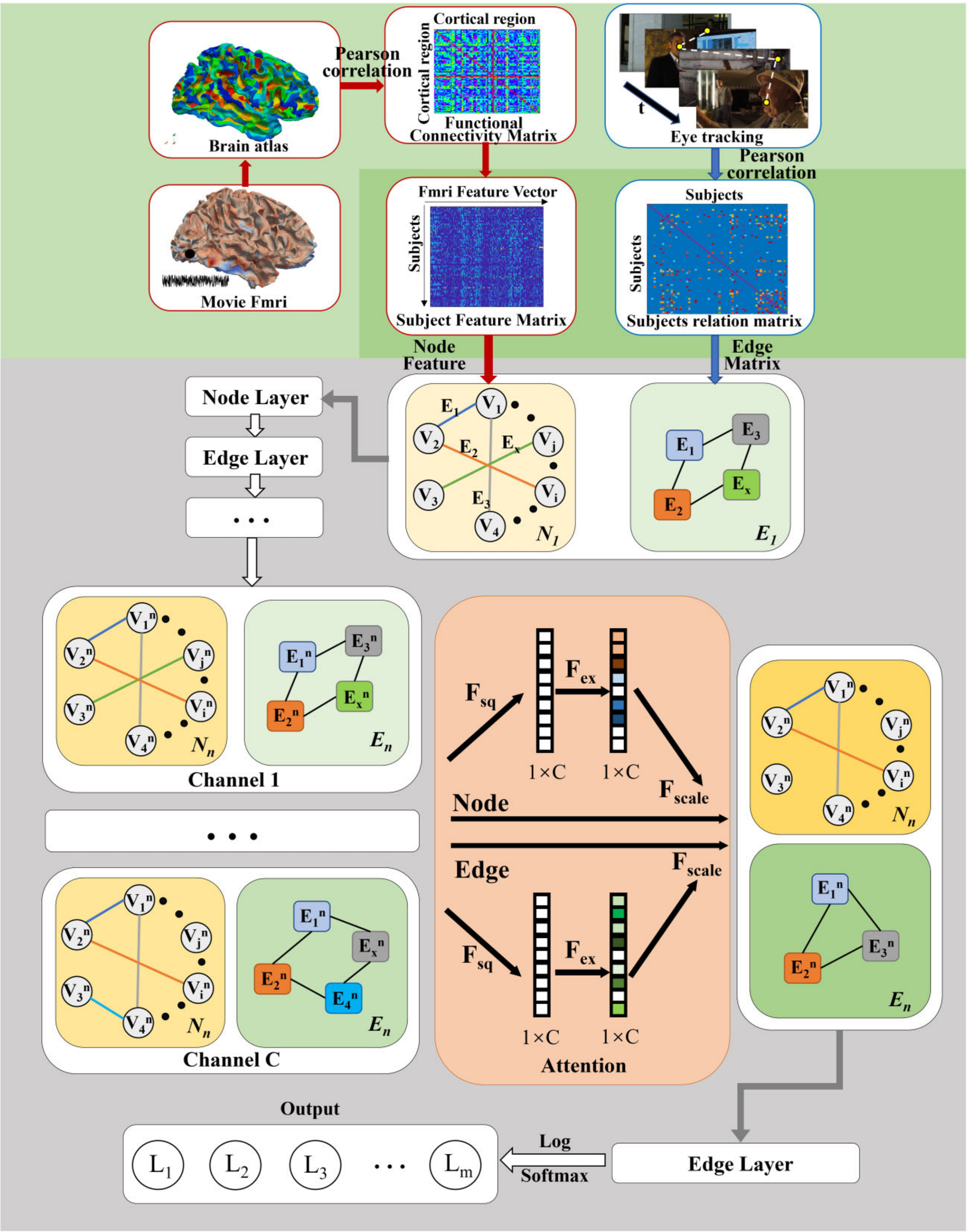
Brain atlas
Pearson correlation
Cortical region
Cortical region
Functional Connectivity Matrix
t
Eye tracking
Pearson correlation
Movie Fmri
Fmri Feature Vector
Subjects
Subject Feature Matrix
Subjects
Subjects
Subjects relation matrix
Node Feature
Edge Matrix
Node Layer
Edge Layer
• • •
E_1 V_1 V_2 E_2 E_x V_j V_3 E_3 V_i V_4 N_1
E_1 E_3 E_x E_2 E_1
V_1^n V_2^n V_j^n V_3^n V_i^n V_4^n N_n
E_1^n E_3^n E_x^n E_2^n E_n
Channel 1
• • •
V_1^n V_2^n V_j^n V_3^n V_i^n V_4^n N_n
E_1^n E_x^n E_2^n E_4^n E_n
Channel C
F_{sq}
F_{ex}
1×C
1×C
F_{scale}
Node
Edge
F_{scale}
F_{sq}
F_{ex}
1×C
1×C
Attention
V_1^n V_2^n V_j^n V_3^n V_i^n V_4^n N_n
E_1^n E_3^n E_2^n E_n
Output
L_1 L_2 L_3 • • • L_m
Log
Softmax
Edge Layer

Figure 1: The preprocessings and flowchart of A-CensNet. Data preprocessing stages are shown in the lighter-green panel. Construction of the graph is in the darker-green panel. Node feature generation is highlighted with red arrows and edge generation with blue arrows. The graph is presented by its node-center version (yellow box) and edge-center version (green box) and is fed to CensNet through different channels. The gray frame highlights the proposed A-CensNet flow. The Attention module (SENet) is highlighted by a orange panel. Note that SENet is inserted in the middle of a CensNet procedure, before the last round of node convolution.

update is then performed in the edge layer. This node-edge switching is accomplished using the following equations:

1. *Propagation rule for node layer.*

$$H_v^{l+1} = \sigma\left(T\Phi\left(H_e^l P_e\right)T^T \odot \tilde{A}_v H_v^l W_v^l\right) \tag{2}$$

where $\tilde{A}_v = \tilde{D}_v^{-\frac{1}{2}} \ (A_v + I_{Nv})\tilde{D}_v^{-\frac{1}{2}}$, $T \in \mathbb{R}^{Nv \times Ne}$ is a binary matrix that indicates whether an edge connects a node. P_e is a learnable weight vector, Φ denotes the diagonalization operation. $\odot$ denotes the element-wise product.

2. *Propagation rule for edge layer.*

$$H_e^{l+1} = \sigma\left(T^T\Phi\left(H_v^l P_v\right)T \odot \tilde{A}_e H_e^l W_e^l\right) \tag{3}$$

where $\tilde{A}_e = \tilde{D}_e^{-\frac{1}{2}} \ (A_e + I_{Ne})\tilde{D}_e^{-\frac{1}{2}}$.

Usually, CensNet ends up with a last round of node layer updating. The loss function is defined as

$$\mathscr{L}\ (\Theta) = -\sum_{l \in Y_L} \sum_{f\ =\ 1}^{F} Y_{lf} log M_{lf} \tag{4}$$

where Y_L is the subset of nodes with labels, M is the softmax results of the last node layer where node feature map has F dimensions.

For the squeeze-and-excitation attention block, as we used different eye movement features to construct heterogeneous graphs, the main purpose of the squeeze-and-excitation module is to integrate these graphs into one framework, by taking each graph as a channel and weighing them. The channel attention block SENet (Hu *et al.*, 2018) is introduced to integrate multiple graphs that share the same nodes but have different node features and edges. Note that a typical CensNet includes (i) the node-and-edge switching embedding plus (ii) an additional round of node embedding (for node classification). The SENet is inserted between (i) and (ii) (carneose panel in Fig. 1). That is, multiple graphs are sent to different channels, where node-and-edge switching embedding is performed. Then, SENet integrates multiple channels into one hybrid graph to complete the last round of node embedding for node classification. Specifically, as illustrated in carneose panel in Fig. 1, the updated node features first pass through a squeeze operation, which aggregates the feature maps across spatial dimension $M \times H_n$ (M: participant number, H_n: node characteristics) to produce a channel descriptor. This is followed by an excitation operation, which is learned for each channel by a self-gating mechanism based on channel dependence. The input channels are then reweighted and fed to the SENet, where F_{sq} is global average pooling function that compresses the characteristics of each channel into a real number, where c (= 4 in this work) denotes the channel number. Then, a 1 × c vector is sent to $F_{ex}(\cdot, W)$ which is implemented by a fully connected layer, where W is a learnable weight for every input channel. Then, F_{scale} scales the 1 × c vector and node-and-edge features multiplied by W to yield the new features. The process is expressed by:

$$z_{node} = \boldsymbol{F}_{sq}\ \left(\mathscr{F}_{node}\right)\ = \frac{1}{M \times H_n} \sum_{i\ =\ 1}^{M} \sum_{j\ =\ 1}^{H_n} \mathscr{F}_{node}\left(i,\ j\right) \tag{5}$$

$$s_{node} = \boldsymbol{F}_{ex}\ \left(z_{node},\ W\right)\ =\ \sigma\left(g\left(z_{node}, W\right)\right)\ =\ \sigma\left(W_2\delta\left(W_1 z_{node}\right)\right) \tag{6}$$

Edge features undergo the same process.

Statistics

Previous studies (Destrieux *et al.*, 2010; Finn & Bandettini, 2021; Gao *et al.*, 2022; He *et al.*, 2020) used regression to predict cognitive scores, and Pearson or Spearman correlation was used to assess prediction accuracy (He *et al.*, 2020). However, small dataset size significantly reduces accuracy, particularly when the sample size is below 100 (He *et al.*, 2020). As our dataset only includes 81 participants, we adopt a classification approach, dividing participants evenly into groups based on cognitive scores and evaluating accuracy using AUC, following the recommendation in (Gao *et al.*, 2022). For the purpose of comparison, we modify the state-of-the-art regression-based methods (Finn & Bandettini, 2021; Gao *et al.*, 2022; He *et al.*, 2020; Kawahara *et al.*, 2017) to classification by adjusting the output layer dimensions and activation function. We use the same loss function and AUC calculation method as A-CensNet, while keeping other layers in their default configuration.

Results

Implementation details

In our application, participants are divided to four cognitive groups (~20 participants in each one, a total of 81 participants). The specific selection criteria of the 81 participants are summarized as follows. Eye movement information is one of the major foci of our selection of participants. The eye movement information in HCP data provides the 2D coordinates of fixation point and pupil area according to the time stamp. When participants are in the state of blinking, the eye tracker will not be able to completely record the coordinates of fixation point and pupil area at this moment. The eye-movement information at this moment is invalid. In this situation, we reorganize the eye movement data, retain only the valid eye movement information, set a threshold for the size of the valid eye movement data file, and then exclude the participants with a lot of invalid information (the file size does not reach the threshold, where effective data reach 15 000 KB and all data is 40 000 KB). In this paper, we use the data in Movies 2 and 3. We identify the individuals who meet the requirements on both of the two movie datasets, yielding 81 participants, an intersection between the two datasets.

We evaluate our method with 50% labeled data in training set, while equally splitting the remaining data sets as validation and test sets (25 to 25%). That is, we randomly select 10 participants from each cognitive group for training (40 participants for training), 20 participants for validation, and 21 participants for testing. The selection of participants is random.

We experiment on preserving {10%, 15%, 20%} top graph edges by their weights, and find that preserving 10% nodes and 10% edges yields the best prediction performance (further details are found in section Ablation study). We try different settings of learning rate from {0.05, 0.01, 0.005, 0.001}, dropout {0.2, 0.3, 0.4, 0.5}, and hidden {16, 32, 64, 128, 512, 1024}, and find that the best performance is yielded by a learning rate setting of 0.005, dropout of 0.2, and hidden of 1024. The following results presented are also based on this parameter setting.

Basic model studies

Since the A-CensNet structure is proposed by adding the Attention mechanism to CensNet, which serves as the basic model of this method, we first investigate whether the performance of CensNet is superior to a more basic model GCN. The results are reported in Table 1. Note that 65 participants (a subset of

Table 1: Basic model studies. The prediction accuracy is measured by AUC. Linear model uses only fMRI, other models use graph structure with eye trajectory similarity as the edge definition. Digits in parentheses indicate the number of participants used.

Models	Linear (65)	GCN (65)	CensNet (65)	CensNet (81)
AUC	41.94 ± 0.81	48.51 ± 0.94	49.75 ± 0.65	50.36 ± 0.71

81 participants) are used in our previous works (Gao *et al.*, 2022) to demonstrate the superiority of GCN over linear method, and the results are reported in Table 1 for reference.

We first apply CensNet to the same 65 participants in (Gao *et al.*, 2022). The AUC is 49.75 ± 0.65, higher than that of GCN (65) (48.51 ± 0.94). Note that the linear method (RidgeClassifier (Pedregosa *et al.*, 2011) yields the worst performance (41.94 ± 0.81). This comparison supports the choice of CensNet as the basic model, and demonstrates that prediction performance of cognition can be improved by a consecutive convolution on both graph nodes and edges, compared with single node convolution step on GCN. When all 81 participants are used in CensNet, prediction performance is further improved (50.36 ± 0.71), supporting the conclusion in (He *et al.*, 2020) that the abundance of training samples largely improves the prediction performance. Note that the effect of dataset size is not the major interest of this work. We only use the 81 participants as a possible lower bound of dataset size, to investigate how far the upper bound of prediction performance can be pushed by means of algorithm refinement.

Ablation study

Effects of attention module (SENet)

Inserting attention module (SENet) into CensNet at different positions (as shown in Figs 1 and 2) could yield different results. Here, we conduct ablation experiments on the original CensNet with no attention module ("AttentionNo" for short) and two different A-CensNets with the attention module inserted into the Middle of the CensNet ("Attention-Middle", Fig. 1) and in front of CensNet ("Attention-Before", Fig. 2). The prediction accuracy by AUC of 100 repeated experiments are reported in Table 2.

Since the original CensNet (the "Attention-No" row in Table 2) does not integrate different graphs. It is constructed as four different graph structures: the fMRI connectivities of Movies 2 and 3 are used as node features, respectively, and the corresponding eye movement trajectories and pupil area during viewing are used to reconstruct edges, respectively. In general, we find that using pupil area to construct edge yields better results than eye movement (Pupil > Track). The "Attention-Middle" model feeds the four graph structures into CensNet, and fuses them through the attention module after updating node and edge features. Instead, the "Attention-Before" model updates weights and merges the graphs by the attention module before the update of node and edge features. The "Attention-Middle" model yields the best result (54.63 ± 0.65), while the "Attention-Before" model yields the worst results (50.24 ± 0.67); even worse than a single graph in the "Attention-No" row. This comparison suggests that a node-edge embedding could yield latent features more sensitive to individual variations, such that a channel attention works better at this deep feature space than being applied immediately after the original shallow features ("Attention-Before").

In addition, we find that the weights in the attention module associated with multiple channels do not significantly change from the initial ones when the algorithm converge, suggesting that all channels equally contribute to the final classification decision-making and that the node and edge feature do compensate each other.

Effects of channels and feature fusion strategy

To investigate the influence of the number of channels on performance, we conduct a series of experiments using two channels. The AUC prediction accuracy from 100 repeated experiments is presented in Table 3.

The results in Table 3 indicate that using two channels does not lead to a significant improvement in accuracy. Specifically, within a single movie dataset (indicated by gray rows in Table 3), when fixing the node feature and using eye trajectory and pupil size variation as the two channels, the performance is not superior to that of the single-channel model in the "Attention-No" section of Table 2.

The results in Tables 2 and 3 show that integrating different eye features is crucial for improving the prediction accuracy. The experiments in the "orange" rows of Table 3 show that using only one eye feature (either trajectory or pupil size) does not significantly improve the performance compared to the single-channel models. Concatenating the features from two datasets (green row) yields the lowest AUC among all experiments, suggesting that feature concatenation alone is not sufficient for improving the performance. Feature concatenation does not yield satisfied results by means of the original CensNet model (yellow rows). These results indicate that multimodal features can only compensate for each other in a "deep" embedding space, but not via a "shallow" fusion. Therefore, the integration strategy across datasets is extremely important.

Finally, it is worth noting that the sparsity of the graph is a very important parameter, and results are shown in Table 4. We use 10% nodes and 10% edges as they yield the best prediction performance.

Comparison with the state-of-the-art

We compare our results with those of state-of-the-art methods listed in Table 5. The results obtained via linear model and GCN were previously reported in Gao *et al.* (2022) for Movie 2 and are included here for reference. It should be noted that linear (Finn & Bandettini, 2021), FNN (He *et al.*, 2020), and BrainNetCNN (He *et al.*, 2020) are not GNN-related methods, as they do not involve edges but only rely on mfMRI features.

Our results demonstrate that all nonlinear deep neural networks provide a significant improvement in contrast to the linear method (41.94 ± 0.81). Within each algorithm, the concatenation of mfMRI features from two datasets (mfMRI2 + 3 rows) does not improve the prediction accuracy in contrast to that of a single dataset. For instance, BrainNetCNN achieves the second-best performance on Movie 2 (53.42 ± 0.63), indicating the importance of the strategy selection for concatenating features from multiple datasets. Within a single movie dataset (unshaded rows), models that integrate multiple modalities of features, such as mfMRI and eye behavior, outperform models (FNN and BrainNetCNN) that use a single modality (mfMRI). After integrating mfMRI and eye behavior from multiple datasets, our results surpass all state-of-

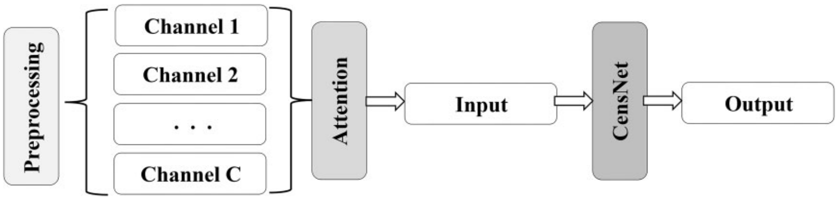

Figure 2: As a comparison, attention module is moved before multiple channels of CensNet.

Table 2: Ablation studies. The prediction accuracy is measured by AUC. Graph structure is represented by {Node, Edge}. C: channel. Trj: eye movement trajectory. Ppl: pupil size variation. The index after mfMRI, Trj and Ppl indicates which movie dataset it comes from. Red: the highest AUC. Blue: the second-highest one.

Models	Graph structure and AUC			
Attention-No	{mfMRI2, Trj2}	{mfMRI2, Ppl2}	{mfMRI3, Trj3}	{mfMRI3, Ppl3}
	50.36 ± 0.71	52.11 ± 0.62	51.49 ± 0.75	51.83 ± 0.85

Attention-Middle	C1: {mfMRI2, Trj2} C2: {mfMRI2, Ppl2}
	C3: {mfMRI3, Trj3} C4: {mfMRI3, Ppl3}			

	54.63 ± 0.65			

Attention-Before	C1: {mfMRI2, Trj2} C2: {mfMRI2, Ppl2}
	C3: {mfMRI3, Trj3} C4: {mfMRI3, Ppl3}			

	50.24 ± 0.67			

Table 3: Two-channel attention ablation study. "+" denotes a temporal concatenation of two features.

Model	Attention-Middle

C1: {mfMRI2, Trj2}
C2: {mfMRI2; Ppl2}
	50.66 ± 0.67

C1: {mfMRI3, Trj3}
C2: {mfMRI3, Ppl3}
	51.79 ± 0.67

C1: {mfMRI2, Trj2}
C2: {mfMRI3, Trj3}
	49.35 ± 0.69

C1: {mfMRI2, Ppl2}
C2: {mfMRI3, Ppl3}
	50.38 ± 0.62

C1: {mfMRI2 + mfMRI3, Trj2 + Trj3}
C2: {mfMRI2 + mfMRI3, Ppl2 + Ppl3}
	46.21 ± 0.55

Single channel {mfMRI2 + mfMRI3, Trj2 + Trj3}	51.94 ± 0.81
Single channel {mfMRI2 + mfMRI3, Ppl2 + Ppl3}	45.45 ± 0.43
Our model (four channels)	**54.63 ± 0.65**

Table 4: The influence of graph node and edge sparsity on prediction. The prediction accuracy is measured by AUC.

Node (reserved)	Edge (reserved)	AUC
10%	10%	54.63 ± 0.65
10%	15%	51.58 ± 0.91
15%	15%	50.93 ± 0.91

the-art methods. These findings highlight the effectiveness of integrating brain activity and eye behavior into a single framework for cognition prediction. Moreover, with a limited number of participants, the integration of multiple loads of stimuli via attention modules could significantly improve the prediction performance.

Table 5: Comparison with the state-of-the-art. Bold font highlights the highest AUC and italic font highlights the second-highest one. "+" denotes a temporal concatenation of two features. NA denotes that no edge is defined in the model. Shaded rows highlight the results on the joint use of Movies 2 and 3.

Models	Graph structure	AUC
Linear (Finn & Bandettini, 2021)	{mfMRI2, NA}	41.94 ± 0.81
GCN (Gao *et al.*, 2022)	{mfMRI2, Trj2}	48.51 ± 0.94
CensNet (Jiang *et al.*, 2019)	{mfMRI2, Trj2}	49.75 ± 0.65
	{ mfMRI2,Trj2}	50.36 ± 0.71
	{ mfMRI2,Ppl2}	52.91 ± 0.73
FNN (He *et al.*, 2020)	{mfMRI2, NA}	50.81 ± 0.66
	{mfMRI3, NA}	49.45 ± 0.85
	{mfMRI2 + 3, NA}	50.54 ± 1.01
BrainNet CNN (He *et al.*, 2020)	{mfMRI2, NA}	53.42 ± 0.63
	{mfMRI3, NA}	49.81 ± 0.78
	{mfMRI2 + 3, NA}	50.66 ± 0.82
A-CensNet	Four channels	**54.63 ± 0.65**

Discussion and Conclusion

We propose A-CensNet to predict participants' cognitive scores, with participants taken as nodes, mfMRI derived functional connectivity as node features, and different eye-tracking feature similarities between participants as heterogeneous graph edges. These graphs from different dataset are all taken as different channels. The proposed model integrates graph embeddings from multiple channels into one. This model outperforms the one using single modality, single channel, and state-of-the-art methods. Our results indicate that the brain functional activity patterns and the behavior patterns might complement each other in interpreting trait-like phenotypes, and might provide new clues to studies of diseases with cognitive abnormality.

Currently, we use classification paradigm to evaluate the prediction performance. Since the distribution of score is continuous, regression is theoretically a more suitable model and has been adopted in many previous studies (Bzdok & Ioannidis, 2019; Finn

Table 6: Multiple: classification prediction results. The prediction accuracy is measured by AUC.

Number of classes	4	6	8	10	12
AUC	54.63 ± 0.65	56.34 ± 0.65	56.95 ± 0.48	51.53 ± 0.60	45.52 ± 0.38

& Bandettini, 2021; Finn *et al.*, 2017; Gal *et al.*, 2022; He *et al.*, 2020; Li *et al.*, 2019; Pereira *et al.*, 2020; Sui *et al.*, 2020). However, the mapping between scores and brain connectivity fingerprint could be far from linear within a cohort and the huge inter-individual variability would lead to a large deviation from the group-wise trend. In short, the predictability of regression could hugely degrade due to the noisy scores and brain activity with huge inter-individual variabilities. For example, in Finn & Bandettini (2021), the regression prediction accuracy for cognition measured by Pearson's *r* is around 0.20 (the best performance on Movie 2 is close to 0.40). Therefore, we use classification scheme instead, to improve the tolerance of the model to noise. In fact, if we increase the class numbers (to the limit, each participant is classified into a unique group), the classification problem ultimately becomes a regression one. In fact, in addition to the four-group classification, we evenly divide the participants into six and eight groups, respectively, and AUCs via our model are 56.34 ± 0.65 and 56.95 ± 0.48, respectively (AUC is 54.63 ± 0.65 for four groups), demonstrating the robustness of the algorithm to classification scheme.

Indeed, to make the results more like a regression, we divide the 81 participants into more groups and report the AUC and accuracy in Table 6. Note that if the class number is 81, to the limit, the classification will be equivalent to regression.

From the results, we can see that with the increase of the number of classes, the classification accuracy increases first but dramatically drops below the chance line (AUC = 0.50) when the class number is 12. Although regression is the ultimate goal of prediction, prediction accuracy measured by a correlation coefficient is very low due to the limitation of dataset size, and thus cannot be a trustworthy metric to compare the performance with the state-of-the-art. Even though the AUCs via the classification scheme are still low in this work, in line with the low Pearson's *r* (around 0.20) via a regression scheme (Finn & Bandettini, 2021), the key point of this work is to demonstrate the improvement brought by the introduction of node-edge joint embedding and attention module for multiple source integration, which has been well validated by a variety of comparative studies. Given that the dataset size might be a critical factor to the prediction performance, as mentioned in the Introduction section, we expect that our algorithm applied on a large-size dataset could achieve a significantly improved precision that still outperforms the state-of-the-art.

The BrainNetCNN model in Finn & Bandettini, 2021 has yielded good results using only fMRI. GCN was adopted simply because it is relatively easier to be implemented on a cohort. By contrast, BrainNetCNN (Kawahara *et al.*, 2017) was specifically developed for brain connectome data and, at the current stage, it is not straightforward for us to used it to fuse eye-movement information with fMRI data. Nevertheless, we may take the advantages of this model and realize the feature fusion before a graph model.

In this work, we only predict cognitive scores. In fact, we also apply our algorithm to emotion score prediction. But the prediction performance is not as successful as that on cognition (AUC = 0.45 via GCN for emotion, contrast by AUC = 0.49 via GCN for cognition), in line with previous studies (Finn & Bandettini, 2021). A possible reason is that the scores in the cognition domain in HCP dataset are performance-based, while those in the emotion domain are self-reported, which can suffer from bias and may be less biologically valid. Another possible reason is that emotion measures simply do not robustly correlate with static functional brain connectivity. Unlike measures including cognition and fluid intelligence, emotion could be dynamically evoked by the prosperous content of a movie clip. The instant arousal of emotion could hardly be promptly reflected by connectivities estimated from the signal correlation, a static measure that covers the whole span of the signal duration. Adopting or adapting an algorithm that focuses more on temporal resolution and relation, such as LSTM (Graves & Graves, 2012) and BERT (Sun *et al.*, 2019) could improve the performance of predicting instant in-task behavior.

Even though the work was applied to healthy participants, the modalities used here, including mfMRI and eye tracking, are essential to advancing the discovery of brain imaging-based markers of psychiatric illness (Eickhoff *et al.*, 2020). Eye tracking has become one of the most popular tools in the research field of psychiatric illness, especially the ones with attention abnormality (Armstrong & Olatunji, 2012; Guillon *et al.*, 2014). For example, context-dependent different visual attention patterns to social stimuli have been found in autism spectrum disorders in contrast to controls (Guillon *et al.*, 2014). Note that our framework can be easily extend to incorporate more eye movement features by simply adding more channels to the SE module, in which multiple eye-movement features are used to represent various definitions of the similarity between individuals. Likewise, it has been demonstrated that using movie fMRI for psychiatric imaging can improve data quality and quantity. Movie watching can decrease repetitive behavioral demands and increase scanner tolerability. It may thus be particularly useful in populations that have difficulties during scanning, including those with psychiatric illness. Moreover, since a movie-watching scheme can magnify the inter-individual variability (Eickhoff *et al.*, 2020; Finn & Bandettini, 2021), a carefully chosen or designed purpose-built movie could potentially magnify the inter-group difference, which, in turn, could greatly facilitate the identification of imaging-based brain markers (Eickhoff *et al.*, 2020). From the perspective of techniques, we have demonstrated a possibility of integrating multiple sources of information to improve the cognition prediction performance. Since cognitive decline is usually an obvious manifestation of progression of many psychiatric diseases, we postulate that the proposed method could be applied to disease study. For example, it might be used to quantify the extent to which a patient's cognitive ability deviate from the trajectory of controls and even predict such deviation ahead of time.

It is worth noting that although the proposed model is not specifically designed for a movie-watching paradigm but the application is only to movie-watching, eye movement data are used, which can only be collected under the paradigm of natural stimulus but this is not possible during a resting state. However, features of other modalities can be jointly used with resting-state fMRI data. For example, we can use resting-state fMRI features as nodal features while using demography metrics, such as age, race, and sex, to measure the similarity between and define edges between participants, and use this graph structure to predict any other personal traits, including the cognition score in this study.

Finally, our method show promise in improving cognition prediction by cross-modality fusion. This avenue deserves further efforts. A possible solution could be cross-modality attention, where we may carry out feature extraction at common frequency domains as suggested in Huddar *et al.* (2020). After that, we can align the features of the two modalities in the time dimension and feed them to the cross-modality models (such as Wei *et al.*, 2020). Finally, the aligned feature can be used feed to the cross-attention module for inter-modal correlation (e.g. Wei *et al.*, 2020), where transformer modules, followed by a 1D-CNN and pooling operation, are used to fuse features.

Supplementary Data

Supplementary data are available at *Psychoradiology Journal* online.

Author contributions

J.G. led the formal analysis, methodology development, and writing (equal) of the original draft; L.Z. was in charge of data curation (equal); T.Z. was charge of validation (equal) and reviewing the writing (equal); C.L. was in charge of data curation (equal) and reviewing the writing (equal); Z.H. was in charge of data curation (equal); Y.W. supported the methodology development and manuscript editing; S.Z. and T.L. supported the conceptualization and funding acquisition; L.G. supported the supervision and funding acquisition; J.H. supported the project administration and supervision; X.J. supported the conceptualization and reviewed the writing; T.Z. led the funding acquisition, supervision, and writing.

Conflict of interests

Prof. Tianming Liu holds the position of Editorial Board Member for *Psychoradiology* and is blinded from reviewing or making decisions for the manuscript.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (31971288, U1801265, 61936007, 62276050, 61976045, U20B2065, U1801265 and 61936007); the National Key R&D Program of China under Grant 2020AAA0105701; High-level researcher start-up projects (Grant No. 06100–22GH0202178); Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University CX2022052.

References

Al-Aidroos N, Said CP, Turk-Browne NB (2012) Top-down attention switches coupling between low-level and high-level areas of human visual cortex. *Proc Natl Acad Sci USA* **109**:14675–80.

Armstrong T, Olatunji BO (2012) Eye tracking of attention in the affective disorders: a meta-analytic review and synthesis. *Clin Psychol Rev* **32**:704–23.

Axer M, Amunts K (2022) Scale matters: the nested human connectome. *Science* **378**:500–4.

Baars BJ, Gage NM (2010) *Cognition, Brain, and Consciousness: Introduction to Cognitive Neuroscience*. The Occupational Therapy Association of South Africa, Pretoria: Academic Press.

Barack DL, Krakauer JW (2021) Two views on the cognitive brain. *Nat Rev Neurosci* **22**:359–71.

Barch DM, Burgess GC, Harms MP, *et al.* (2013) Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* **80**:169–89.

Beatty J, Lucero-Wagoner B (2000) The pupillary system. In: JT Cacioppo, LG Tassinary, GG Berntson (eds). *Handbook of psychophysiology*. Cambraidge, United Kingdom: Cambridge University Press, 142–62.

Bressler SL, Menon V (2010) Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn Sci* **14**:277–90.

Bullmore E, Sporns O (2012) The economy of brain network organization. *Nat Rev Neurosci* **13**:336–49.

Bzdok D, Ioannidis JP (2019) Exploration, inference, and prediction in neuroscience and biomedicine. *Trends Neurosci* **42**:251–62.

Carter BT, Luke SG (2020) Best practices in eye tracking research. *Int J Psychophysiol* **155**:49–62.

Deco G, Jirsa VK, McIntosh AR (2011) Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nat Rev Neurosci* **12**:43–56.

Defferrard M, Bresson X, Vandergheynst P (2016) Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in Neural Information Processing Systems*. United States: The MIT Press, 29.

Destrieux C, Fischl B, Dale A, *et al.* (2010) Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* **53**:1–15.

Diamond A (2013) Executive functions. *Annu Rev Psychol* **64**:135.

Eickhoff SB, Milham M, Vanderwal T (2020) Towards clinical applications of movie fMRI. *Neuroimage* **217**:116860.

Einhäuser W (2017) The pupil as marker of cognitive processes. In *Computational and Cognitive Neuroscience of Vision*. Singapore: Springer, 141–69.

Finn ES, Bandettini PA (2021) Movie-watching outperforms rest for functional connectivity-based prediction of behavior. *Neuroimage* **235**:117963.

Finn ES, Scheinost D, Finn DM, *et al.* (2017) Can brain state be manipulated to emphasize individual differences in functional connectivity? *Neuroimage* **160**:140–51.

Fox MD, Raichle ME (2007) Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat Rev Neurosci* **8**:700–11.

Gal S, Coldham Y, Tik N, *et al.* (2022) Act natural: functional connectivity from naturalistic stimuli fMRI outperforms resting-state in predicting brain activity. *Neuroimage* **258**:119359.

Gallistel CR, King AP (2011) *Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience*. 9600 Garsington Road, Oxford: John Wiley & Sons9781444359763.

Gao J, Li C, He Z, *et al.* (2022) Prediction of cognitive scores by movie-watching fmri connectivity and eye movement via spectral graph convolutions. *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*.

Glasser MF, Sotiropoulos SN, Wilson JA, *et al.* (2013) The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**:105–24.

Grady CL, Rieck JR, Nichol D, *et al.* (2021) Influence of sample size and analytic approach on stability and interpretation of brain-behavior correlations in task-related fMRI data. *Hum Brain Mapp* **42**:204–19.

Graves A, Graves A (2012) Long short-term memory. *Supervised Sequence Labelling with Recurrent Neural Networks Studies in Computational Intelligence*, Berlin, Heidelberg: Springer, **385**, 37–45.

Griffanti L, Salimi-Khorshidi G, Beckmann CF, *et al.* (2014) ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *Neuroimage* **95**:232–47.

Guillon Q, Hadjikhani N, Baduel S, *et al.* (2014) Visual social attention in autism spectrum disorder: insights from eye tracking studies. *Neurosci Biobehav Rev* **42**:279–97.

Harvey PD (2019) Domains of cognition and their assessment. *Dialogues Clin Neurosci* **21**:227–37

He T, An L, Chen P, *et al.* (2022) Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nat Neurosci* **25**:795–804.

He T, Kong R, Holmes AJ, *et al.* (2020) Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *Neuroimage* **206**:116276.

Hess EH, Polt JM (1960) Pupil size as related to interest value of visual stimuli. *Science* **132**:349–50.

Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–41

Huddar MG, Sannakki SS, Rajpurohit VS (2020) Attention-based word-level contextual feature extraction and cross-modality fusion for sentiment analysis and emotion classification. *Int J Intell Eng Informatics* **8**:1.

Huijbers W, Van Dijk KR, Boenniger MM, *et al.* (2017) Less head motion during MRI under task than resting-state conditions. *Neuroimage* **147**:111–20.

Jiang X, Ji P, Li S (2019) *CensNet: Convolution with Edge-Node Switching in Graph Neural Networks*. Macao,China: IJCAI, 2656–62.

Kannurpatti SS, Rypma B, Biswal BB (2012) Prediction of task-related BOLD fMRI with amplitude signatures of resting-State fMRI. *Front Syst Neurosci* **6**:7.

Kawahara J, Brown CJ, Miller SP, *et al.* (2017) BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. *Neuroimage* **146**:1038–49.

Laeng B, Alnaes D (2019) Pupillometry. In *Eye Movement Research*. Cham: Springer, 449–502

Laeng B, Sirois S, Gredebäck G (2012) Pupillometry: a window to the preconscious? *Persp Psychol Sci* **7**:18–27.

LeMoult J, Gotlib IH (2019) Depression: a cognitive perspective. *Clin Psychol Rev* **69**:51–66.

Lezak MD, Howieson DB, Loring DW, *et al.* (2004) *Neuropsychological Assessment*. Oxford University Press, USA.

Li J, Kong R, Liégeois R, *et al.* (2019) Global signal regression strengthens association between resting-state functional connectivity and behavior. *Neuroimage* **196**:126–41.

Lim JZ, Mountstephens J, Teo J (2020) Emotion recognition using eye-tracking: taxonomy, review and current challenges. *Sensors* **20**:2384.

Liu G, Locascio JJ, Corvol J-C, *et al.* (2017) Prediction of cognition in Parkinson's disease with a clinical–genetic score: a longitudinal analysis of nine cohorts. *Lancet Neurol* **16**:620–9.

Lohse GL, Johnson EJ (1996) A comparison of two process tracing methods for choice tasks. *Organ Behav Hum Decis Process* **68**:28–43.

Malinen S, Hlushchuk Y, Hari R (2007) Towards natural stimulation in fMRI—issues of data analysis. *Neuroimage* **35**:131–9.

Mathôt S (2018) Pupillometry: psychology, physiology, and function. *J Cogn* **1**:16.

Pedregosa F, Varoquaux G, Gramfort A, *et al.* (2011) Scikit-learn: machine learning in Python. *J Machine Learn Res* **12**:2825–30.

Pereira TD, Shaevitz JW, Murthy M (2020) Quantifying behavior to understand the brain. *Nat Neurosci* **23**:1537–49.

Pick S, Goldstein LH, Perez DL, *et al.* (2019) Emotional processing in functional neurological disorder: a review, biopsychosocial model and research agenda. *J Neurol Neurosurg Psychiatry* **90**:704–11.

Son J, Ai L, Lim R, *et al.* (2020) Evaluating fMRI-based estimation of eye gaze during naturalistic viewing. *Cereb Cortex* **30**:1171–84.

Sonkusare S, Breakspear M, Guo C. (2019) Naturalistic stimuli in neuroscience: critically acclaimed. *Trends Cogn Sci* **23**:699–714.

Stern Y (2012) Cognitive reserve in ageing and Alzheimer's disease. *Lancet Neurol* **11**:1006–12.

Sui J, Jiang R, Bustillo J, *et al.* (2020) Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises. *Biol Psychiatry* **88**:818–28.

Sun F, Liu J, Wu J, *et al.* (2019) *BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, Beijing, China. https://doi.org/10.1145/3357384.3357895

Sun J, Wang B, Niu Y, *et al.* (2020) Complexity analysis of EEG, MEG, and fMRI in mild cognitive impairment and Alzheimer's disease: a review. *Entropy* **22**:239.https://doi.org/10.3390/e22020239

Thiebaut de Schotten M, Forkel SJ. (2022) The emergent properties of the connected brain. *Science* **378**:505–10.

Van Den Heuvel MP, Sporns O. (2011) Rich-club organization of the human connectome. *J Neurosci* **31**:15775–86.

Wei X, Zhang T, Li Y, *et al.* (2020) Multi-modality cross attention network for image and sentence matching. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10941–50.

Wolters AF, van de Weijer SCF, Leentjens AFG, *et al.* (2019) Resting-state fMRI in Parkinson's disease patients with cognitive impairment: a meta-analysis. *Parkinsonism Relat Disord* **62**:16–27.

Yang J, Gohel S, Vachha B. (2020) Current methods and new directions in resting state fMRI. *Clin Imaging* **65**:47–53.

Received: 5 April 2023; **Revised:** 14 June 2023; **Accepted:** 6 July 2023