Medical Image Analysis 88 (2023) 102841

Contents lists available at ScienceDirect

Medical Image Analysis

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

BolT: Fused window transformers for fMRI time series analysis

Hasan A. Bedel [a,b], Irmak Sivgin [a,b], Onat Dalmaz [a,b], Salman U.H. Dar [a,b], Tolga Çukur [a,b,c,*]

[a] *Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey*
[b] *National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara 06800, Turkey*
[c] *Neuroscience Program, Bilkent University, Ankara 06800, Turkey*

ARTICLE INFO

Keywords:
Functional MRI
Time series
Deep learning
Transformer
Classification
Connectivity
Explainability

ABSTRACT

Deep-learning models have enabled performance leaps in analysis of high-dimensional functional MRI (fMRI) data. Yet, many previous methods are suboptimally sensitive for contextual representations across diverse time scales. Here, we present BolT, a blood-oxygen-level-dependent transformer model, for analyzing multi-variate fMRI time series. BolT leverages a cascade of transformer encoders equipped with a novel fused window attention mechanism. Encoding is performed on temporally-overlapped windows within the time series to capture local representations. To integrate information temporally, cross-window attention is computed between base tokens in each window and fringe tokens from neighboring windows. To gradually transition from local to global representations, the extent of window overlap and thereby number of fringe tokens are progressively increased across the cascade. Finally, a novel cross-window regularization is employed to align high-level classification features across the time series. Comprehensive experiments on large-scale public datasets demonstrate the superior performance of BolT against state-of-the-art methods. Furthermore, explanatory analyses to identify landmark time points and regions that contribute most significantly to model decisions corroborate prominent neuroscientific findings in the literature.

1. Introduction

Functional MRI (fMRI) measures blood–oxygen-level-dependent (BOLD) responses that reflect changes in metabolic demand consequent to neural activity (Hillman, 2014; Rajapakse et al., 1998). Recording BOLD responses at a unique combination of spatio-temporal resolution and coverage, fMRI provides the means to study complex cognitive processes in the human brain Kubicki et al. (2003), Wang et al. (2005), Papma et al. (2017), Mensch et al. (2017). On the one hand, task-based fMRI enables researchers to associate stimulus or task variables with multi-variate responses across the brain Li et al. (2009), Venkataraman et al. (2009), Nishimoto et al. (2011). Regions that are co-activated in the presence of a particular variable are taken to be involved in the cortical representation of that variable (Simon et al., 2004), and they are considered to be functionally connected (Rogers et al., 2007). On the other hand, characteristic multi-variate responses are also eminent in the absence of external stimuli or task, when the subject is merely resting (Niu et al., 2021; Yeo et al., 2011; Van Dijk et al., 2010; Hu and Shi, 2006). In resting-state fMRI, co-activation patterns are typically used to define networks of brain regions, whose functional connectivity (FC) has been associated with various normal and disease states (Greicius, 2008; Lei et al., 2021; Iraji et al., 2015; Zhang et al., 2017). Many prior studies have linked behavioral traits and prominent neurological diseases with FC features of BOLD responses (Kong et al., 2019; Rajpoot et al., 2015; Müller-Oehring et al., 2018; Anderson et al., 2013).

Earlier fMRI studies adopted traditional machine learning (ML) to analyze multi-variate brain responses in order to decode task- or disease-related information. Since these ML methods use relatively compact models, feature extraction is typically employed to reduce dimensionality and factor out nuisance variability (McKeown and Sejnowski, 1998; Svensén et al., 2002). A prominent approach first expresses FC features as the temporal correlations of BOLD responses across separate brain regions, and then uses methods such as support vector machines or logistic regression to classify external variables (Pereira et al., 2009; De Martino et al., 2008; Zhang et al., 2015; Wang et al., 2019). Later studies have instead adopted deep learning (DL) given its ability to capture complex patterns in high-dimensional data (Heinsfeld et al., 2018; Li et al., 2020b; Duncan et al., 2019; Mlynarski et al., 2019; Kam et al., 2019). Various successful deep models have been proposed in the literature based on convolutional (Kawahara et al., 2017), graph (Parisot et al., 2018), or recurrent architectures (Fan et al., 2020; Wang et al., 2021) that process FC features. Yet, common FC features primarily reflect first-order inter-regional interactions, potentially disregarding higher-order

[*] Corresponding author at: Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey.
E-mail address: cukur@ee.bilkent.edu.tr (T. Çukur).

https://doi.org/10.1016/j.media.2023.102841
Received 19 September 2022; Received in revised form 7 February 2023; Accepted 10 May 2023
Available online 18 May 2023

interactions evident in recorded BOLD responses (Lahaye et al., 2003; Hu and Shi, 2007). To more directly assess information in fMRI data, several recent studies have instead built classifiers using recurrent models or vanilla transformer models (Dvornek et al., 2017; Nguyen et al., 2020; Malkiel et al., 2021) on BOLD responses. While powerful, these recent architectures can introduce high computational burden when processing long time series, and they do not embody explicit mechanisms to capture contextual representations of multi-variate data across diverse time scales (Ismail et al., 2019; Liégeois et al., 2019; Allen et al., 2014).

Here we propose a novel transformer architecture that directly operates on BOLD responses, BolT, for fMRI time-series classification. To capture local representations, BolT splits the time series into temporally-overlapping windows and employs a cascade of transformer blocks to encode window-specific representations of BOLD tokens (i.e., linear projections of responses measured across the brain at specific time points). To enhance expressiveness across broad time scales without elevating computational costs, BolT leverages a novel fused window attention mechanism that utilizes cross attention and token fusion among overlapping windows. While cross attention enables interactions between base BOLD tokens in a given window and fringe tokens in neighboring windows prior to encoding, token fusion enables integration of encoded representations across neighboring windows. To hierarchically transition from local to global representations, the extent of window overlap in transformer blocks is progressively increased across the cascade. BolT improves task performance by utilizing classification (CLS) tokens to capture task-oriented high-level features. Window-specific CLS tokens are introduced to maintain local sensitivity and compatibility with the hierarchical model structure. Meanwhile, task-relevant information exchange is promoted by a novel cross-window regularization that aligns these CLS tokens across windows. At the end of the cascade, the encoded CLS tokens are averaged across windows and a linear projection layer is used for classification.

Comprehensive demonstrations are reported for classification tasks on public datasets: gender detection from resting-state fMRI scans and cognitive task detection from task-based fMRI scans in the Human Connectome Project (HCP) dataset (Van Essen et al., 2013), and disease detection from resting-state fMRI scans in the Autism Brain Imaging Data Exchange (ABIDE) dataset (Di Martino et al., 2014). BolT achieves higher classification performance than prior traditional and deep-learning methods, including convolutional, graph, recurrent and transformer baselines. Ablation studies are presented that demonstrate the significant contribution of individual design elements to model performance, including learnable CLS tokens, split time windows, token fusion, cross attention, and cross-window regularization. To interpret the representational information captured by BolT, we devise an explanatory technique on the fused window attention operators. The proposed technique extracts gradient-weighted attention maps across the cascade to construct an importance map for BOLD tokens, and thereby identify landmark time points. A logistic regression analysis on the landmark points in then performed to identify brain regions that contribute most significantly to the model's decision. Explanatory analyses reveal task timings and relevant brain regions that corroborate established neuroscientific findings in the literature. Code for implementing BolT is publicly available at https://github.com/icon-lab/BolT.

Contributions

- We introduce a novel transformer architecture to efficiently and sensitively analyze fMRI BOLD responses.
- A novel fused window attention mechanism is proposed with progressively grown window size to hierarchically capture local-to-global representations.
- A novel cross-window regularization is proposed on global classification features to align high-level representations across the time series.
- An explanatory technique is introduced for BolT that evaluates the relevance of individual time points and brain regions to the classification decisions.

2. Related work

2.1. Traditional methods

Whole-brain fMRI data carry densely overlaid patterns of multi-variate responses, which can be difficult to isolate via uni-variate analysis (Penny et al., 2011; Woolrich et al., 2001). This has sparked interest in adoption of ML for multi-variate fMRI analysis (Norman et al., 2006; Haxby, 2012). Earlier studies in this domain used traditional classifiers such as support vectors machines (Song and Chen, 2014; Wang et al., 2007; Hojjati et al., 2017). Because high-dimensional data are paired with models of limited complexity, feature selection is key to improving sensitivity in traditional models (Bullmore et al., 1996; Xie et al., 2009; Poldrack, 2007). Accordingly, many traditional models are built on FC features derived from response correlations among brain regions-of-interest (ROIs), as these features are commonly considered to capture discriminative information about cognitive state (Zeng et al., 2012; Shen et al., 2010; Khazaee et al., 2016).

2.2. Deep learning methods on FC features

In recent years, DL models have been adopted to elevate sensitivity in fMRI analysis. Some studies have used multi-layer perceptron (MLP) or convolutional neural network (CNN) models to extract high-level features of fMRI data (Suk et al., 2016; Koyamada et al., 2015; Huang et al., 2017) and then to classify external variables (Sarraf and Tofighi, 2016a,b; Zhao et al., 2017). More commonly, classification models have been built based on FC features among brain ROIs for improved performance (Meszlényi et al., 2017; Kawahara et al., 2017; Xing et al., 2019). Given the brain's intrinsic structure, graph neural networks (GNN) have gained traction wherein individual ROIs denote nodes and FC features among ROIs determine edge weights (Li et al., 2021, 2019). To capture temporal variability in dynamic FC features, recurrent or transformer architectures have also been integrated to process the GNN outputs (Kim et al., 2021). However, GNN-based models might suffer from over smoothing (Chen et al., 2020) or squashing (Alon and Yahav, 2020) that can lower sensitivity to long-range dependencies. Furthermore, while methods that receive FC features as input can improve learning efficiency by mitigating nuisance variability, FC features typically reflect first-order interactions among ROIs, neglecting potential non-linear effects (Su et al., 2013).

2.3. Deep learning methods on BOLD responses

Recurrent networks: An alternative approach to building models on pre-extracted FC features is to directly analyze BOLD responses in fMRI time series. Given the high degree of temporal correlation in BOLD responses, recurrent neural networks (RNNs) have been proposed to sequentially process fMRI data across time given CNN-based or ROI-extracted spatial representations (Li et al., 2020a; Dvornek et al., 2017; Zhao et al., 2020). Previously reported recurrent architectures in the fMRI literature include vanilla long short-term memory (LSTM) models (Dvornek et al., 2017), and hybrid convolutional LSTM models (Li et al., 2020a; Zhao et al., 2020). While recurrent architectures are powerful in time series analysis, sequential processing introduces difficulties in model training on long time series due to vanishing gradients, and hence they may show suboptimal sensitivity to long-range interactions (Kerg et al., 2020).

Vanilla transformers: Transformer architectures based on self-attention mechanisms have recently been introduced to address limitations of recurrent networks (Vaswani et al., 2017). Given a sequence of tokens, self-attention operators filter their inputs based on inter-token similarity to integrate long-range contextual information. A feed-forward network block, typically selected as an MLP, then encodes latent representations of the contextualized tokens. Several recent studies have employed vanilla transformers that process the entire fMRI time series as a single sequence (Nguyen et al., 2020; Zhang et al., 2022). Vanilla transformers can manifest relatively limited sensitivity to local representations while emphasizing long-range temporal interactions. Moreover, they introduce quadratic computational complexity with respect to sequence length as self-attention requires similarity assessment between all pairs of tokens.

Efficient transformers: To sensitively analyze fMRI data across diverse time scales while mitigating computational burden, here we introduce a novel transformer architecture, BolT, based on a fused window self-attention mechanism (FW-MSA, Section 3.1). Unlike previous methods that receive as input pre-extracted FC features (Abraham et al., 2017; Parisot et al., 2017; Gadgil et al., 2020; Li et al., 2021; Kim et al., 2021), BolT performs learning on BOLD responses to improve sensitivity. Unlike vanilla transformers that process the time series as a single sequence to focus on global temporal representations (Nguyen et al., 2020; Zhang et al., 2022), BolT improves efficiency by splitting the time series into overlapping windows, and employs a cascaded transformer encoder that hierarchically extracts local-to-global representations.

Several recent studies have devised efficient transformer models with partially similar aims to our proposed approach. A computer vision study has introduced SwinT that restricts self-attention computations to non-overlapping local windows in a given sequence, and performs half-sequence-length shifts in the window position across alternating transformer blocks (Liu et al., 2021). SwinT implicitly captures cross-window interactions via the window shifts, and it does not utilize high-level CLS tokens. Instead, BolT explicitly captures cross-window interactions by using overlapping windows along with cross attention and token fusion between neighboring windows, and it utilizes dedicated CLS tokens to improve classification. A natural language processing study has proposed Longformer that restricts self-attention to a moving local window centered on each token for encoding the token in a single context, and it uses a global CLS token across the sequence that can degrade sensitivity to local representations (Beltagy et al., 2020). In contrast, BolT encodes each token appearing in multiple overlapping windows in multiple contexts and then fuses these encodings, and it uses window-specific CLS tokens that are aligned with cross-window regularization for enhanced sensitivity.

Efficient transformers have also been adopted in medical image analysis tasks. A hybrid CNN-transformer model, IFT-Net, has been proposed that reduces dimensionality of input data with a convolutional module prior to the transformer (Zhao et al., 2022). While this approach reduces the sequence length for self-attention computations, it can compromise temporal resolution and sensitivity to local representations in time series analysis. Instead, BolT maintains local sensitivity by preserving temporal dimensionality across the cascade. Another study has introduced HATNet as a hybrid CNN-transformer model where the transformer module sequentially computes intra- and inter-window attention on non-overlapping windows (Mehta et al., 2022). This sequential approach can be suboptimal since inter-window computations are performed on self-attention outputs that ignore cross-window interactions. In contrast, BolT simultaneously computes self- and cross-attention in overlapping windows. A recent fMRI study has introduced a cascaded transformer, TFF, that splits the time series into separate windows to focus on local temporal representations (Malkiel et al., 2021). TFF processes tokens in separate windows independently across the cascade and naively averages the encoded representations over windows, reducing sensitivity to long-range context. Instead, in each stage of the cascade, BolT uses FW-MSA modules to capture interactions that extend over broad time scales via learning-based fusion of information flow across windows.

3. Theory

For multi-variate analysis of four-dimensional (4D) fMRI data recorded in a subject, regional BOLD responses are first extracted using an external atlas parcellating the brain into R ROIs. The time series for a given ROI is taken as the average response across voxels within the ROI, z-scored to zero mean and unit variance. Our model learns to map these regional BOLD responses $x \in \mathbb{R}^{T\times R}$ (where T is the number of time samples in the fMRI scan) onto class labels y (e.g. subject gender, cognitive task) depending on the task. Note that transformers expect a sequence of tokens as input. Here, we refer to a learnable linear projection of BOLD responses measured at a particular time index as a BOLD token, i.e., $b^{(t)} = f_b(x^{(t)}) \in \mathbb{R}^N$, where t is the time index, f_b is the linear projection, N is the encoding dimensionality. The collection of BOLD tokens across the fMRI scan is then $b = (b^{(0)}, \dots, b^{(T-1)}) \in \mathbb{R}^{T\times N}$. Latent representations of BOLD tokens are computed by a cascade of transformer blocks in BolT (Fig. 1). The learned latent features h_f are then linearly projected onto individual class probabilities. To capture both local and global representations of BOLD tokens, the transformer blocks split the fMRI time series into F overlapping windows, and employ a novel fused window self-attention operator to assess interactions between base tokens in a given time window and fringe tokens in neighboring windows. In this section, we describe the architectural details of the proposed model and introduce an explanatory technique devised for BolT.

3.1. BolT

BolT embodies a cascade of transformer blocks and a final linear layer to perform classification. Unlike vanilla transformers (Vaswani et al., 2017), BolT comprises a novel FW-MSA module to enhance sensitivity to the diverse time scales of dynamic interactions in the brain, while maintaining linear scalability with the duration of fMRI scans (Fig. 2). A regular MSA layer uses global attention across tokens resulting in quadratic complexity. In contrast, FW-MSA computes local attention within compact time windows extracted from the fMRI scan. Temporal windowing restricts token-to-token interactions to a focal neighborhood surrounding each window. The resultant local precision serves to improve the capture of subtle changes in brain activation dynamics (Hutchison et al., 2013). To capture a window-level latent representation, a CLS token is also employed for use in downstream detection tasks (Dosovitskiy et al., 2020). Input CLS tokens are initialized as tied vectors across separate time windows. The CLS token for each window is concatenated to the query, value and key tokens in FW-MSA. The final layer uses output CLS tokens to linearly map their aggregate features onto class logits.

Fused window attention: FW-MSA enables cross-window interactions by attention calculation between base tokens in a given window and fringe tokens in neighboring windows. To do this, FW-MSA first splits the entire collection of BOLD tokens $b \in \mathbb{R}^{T\times N}$ into $F = (T-W)/s+1$ windows of size W and stride s. The receptive field of a given window contains W base tokens centrally, and L fringe tokens on either side of the base. While processing the ith window, FW-MSA receives as input a collection of CLS ($CLS_i \in \mathbb{R}^N$) and BOLD ($b_i \in \mathbb{R}^{(W+2L)\times N}$) tokens. Let $Q_i \in \mathbb{R}^{(1+W)\times N}$ denote queries for base tokens, and $K_i \in \mathbb{R}^{(1+W+2L)\times N}$ and $V_i \in \mathbb{R}^{(1+W+2L)\times N}$ denote keys and values for the union of base and fringe tokens. Assuming f_q, f_k and f_v are learnable linear projections, the query, key and value for the i-th window are:

$$\begin{aligned}
Q_i &= f_q(\{CLS_i, b^{(i\times s)}, \dots, b^{(i\times s+W-1)}\}),\\
K_i &= f_k(\{CLS_i, b^{(i\times s-L)}, \dots, b^{(i\times s+W+L-1)}\}),\\
V_i &= f_v(\{CLS_i, b^{(i\times s-L)}, \dots, b^{(i\times s+W+L-1)}\}). \qquad (1)
\end{aligned}$$

To leverage information in the temporal ordering of BOLD tokens, we incorporate a relative position bias in attention calculations (Liu et al., 2021; Yang et al., 2021):

$$\mathrm{Attention}(Q_i, K_i, V_i) = \mathrm{Softmax}(\frac{Q_i K_i^T}{\sqrt{d}} + B)V_i, \qquad (2)$$

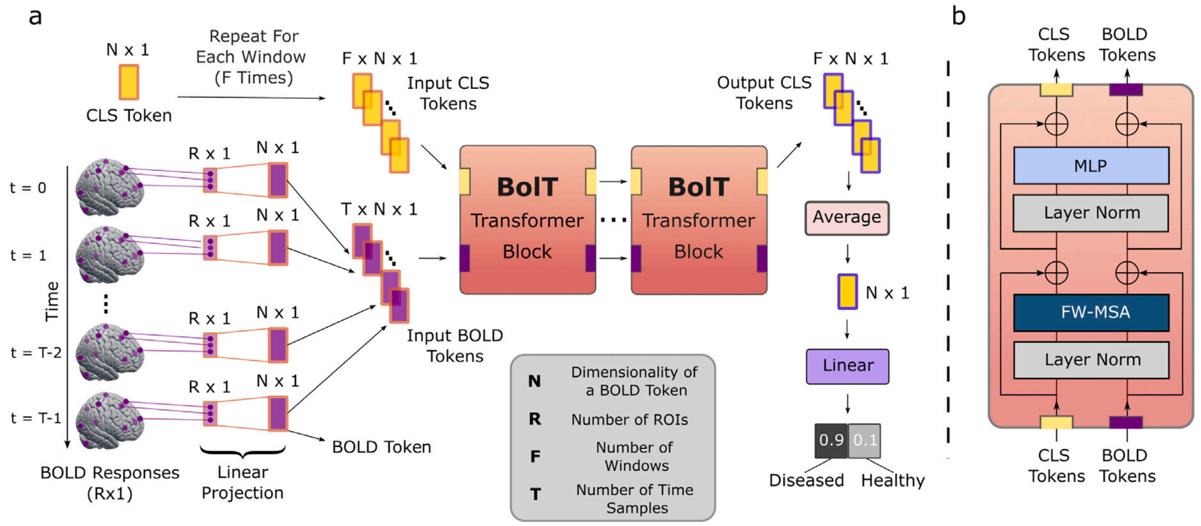

Fig. 1. (a) Overview of BolT. First, ROI-level BOLD responses are extracted from four-dimensional fMRI data. These responses are then projected by a learnable linear layer to obtain T BOLD tokens. Each BOLD token encodes ROI responses across the brain recorded at a specific time instant as an N-dimensional vector. A cascade of transformer blocks processes BOLD tokens across a collection of F temporally-overlapping windows within the time series. For each time window, a separate learnable classification (CLS) token is employed within the transformer blocks. The CLS tokens input to the first block are initialized as tied vectors across F windows, but they become window-specific following encoding through transformer blocks. The blocks compute latent representations of BOLD and CLS tokens; yet only the CLS tokens are used for the classification task at the output layer. (b) Inner architecture of the transformer block. Unlike vanilla transformers, BolT is equipped with a novel fused window multi-head self-attention (FW-MSA) layer to efficiently capture both local and global context within the fMRI time series.

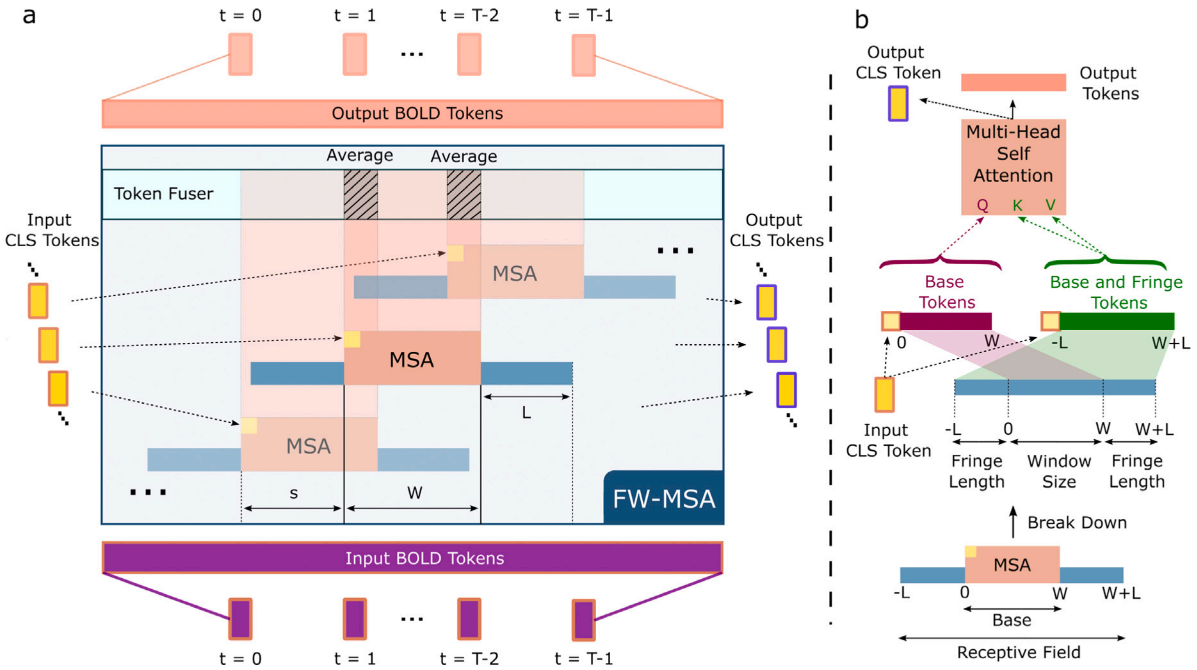

Fig. 2. (a) Schematic of the fused window multi-head self-attention (FW-MSA) module. Input BOLD tokens are separated into an overlapping set of time windows of size W and stride s. A CLS token is assigned to each window. A given window possesses a bilateral fringe region of size L to permit interactions with neighboring windows. Within-window interactions are captured via attention among base tokens, whereas cross-window interactions are captured via attention between base and fringe tokens. For each BOLD token, attention-derived latent representations are then fused across the separate time windows in which it appears. (b) Attention calculations within a window. Query to the FW-MSA is a BOLD token from the window base, whereas key and value are BOLD tokens from the broader receptive field including the fringe region.

where $B \in \mathbb{R}^{(1+W)\times(1+W+2L)}$ is a learnable positional bias matrix and d is the feature dimensionality of the attention head. FW-MSA includes multiple attention heads, albeit expressions are given for a single head for simplicity. B expresses the positioning of base and CLS tokens with respect to all tokens in the receptive field including base, fringe and CLS tokens. For BOLD tokens, B parametrizes the potential range $[-W-L+1, W+L-1]$ of relative distances among tokens in different positions of the receptive field. For the CLS token, it instead serves to distinguish the CLS token from BOLD tokens.

Token fusion: FW-MSA calculates latent representations of each BOLD token given surrounding local context. Because each token appears in multiple windows, a token fuser is used to aggregate the resultant representations:

$$b^{(i)}[m] = \frac{1}{P}\sum_{p=0}^{P-1} b_p^{(i)}[m-1], \tag{3}$$

where $m \in \{0,1,\ldots,M-1\}$ is the index of the transformer block, p is the index of the time window among P windows that contain a particular token, $b_p^{(i)}[m-1]$ is the ith input BOLD token, and $b^{(i)}[m]$ is the fused token. Token fusion facilitates exchange of information across windows, while maintaining a fixed number of BOLD tokens across transformer blocks. Following fusion, all tokens are forwarded to the MLP module.

Cross-window regularization: The first transformer block in BolT receives as input a single CLS token shared across the time windows. The transformer encoders then compute a unique CLS token for each window based on the BOLD tokens within its receptive field. If the latent space of window-level representations captured by the CLS tokens are largely incompatible, model performance in downstream classification tasks might be compromised. Thus, to encourage high-level representations that are consistent across time windows, we introduce a novel cross-window regularization as an additional loss term:

$$L_{CWR} = \frac{1}{NF} \sum_{i=0}^{F-1} \|CLS_i[M-1] - \frac{1}{F}(\sum_{j=0}^{F-1} CLS_j[M-1])\|_2^2, \tag{4}$$

where $CLS_i[M-1]$ is the encoded CLS token for the ith window at the output of the last transformer block (Mth). The regularization term in Eq. (4) penalizes the deviation of individual CLS tokens from their mean over windows.

3.2. Explanatory technique

We introduce an explanatory technique for BolT that generates importance weights for BOLD tokens to assess their contribution to a given decision. To do this, we first derive gradient-weighted attention maps as inspired by recent computer vision studies (Chefer et al., 2021). Yet, we propose an adapted procedure for map calculation in FW-MSA to cope with the overlap and fusion operations across time windows (Alg. 1). In the proposed procedure, gradient-weighted attention maps are calculated separately for each transformer block and each time window:

$$\bar{A}_{mi} = E_h((\nabla A_{mi} \odot A_{mi})^+). \tag{5}$$

In Eq. (5), $A_{mi} \in \mathbb{R}^{(1+W)\times(1+W+2L)}$ is the attention map produced by the FW-MSA layer at the mth block for the ith window, and the first row and column of A_{mi} are reserved for attention values related to the CLS token. The effect of map values onto the model output is characterized via ∇A_{mi}, i.e., the gradient of the loss function with respect to A_{mi}. Meanwhile, E_h denotes the averaging operator across attention heads for aggregation, $\odot$ is the Hadamard product to modulate the attention maps with the gradients, and $^+$ denotes rectification to prevent negative values. Within each transformer block, single-window attention maps are then aggregated to form a global attention map across the entire time series, $\bar{A}_G[m] \in \mathbb{R}^{(F+T)\times(F+T)}$ where F is the number of windows, T is the number of BOLD tokens. During aggregation, projections of fringe tokens that appear in multiple windows are averaged across windows. Assuming that $t' = F + i \times s$ is the starting index of base tokens in the ith window, $t'' = t' - L + p_l$ and $t''' = t' + W + L - p_r$ respectively denote the starting and ending indices of the fringe tokens, where $p_l = max(0, i\times s - L) - (i\times s - L)$ and $p_r = i\times s + W + L - min(T, i\times s + W + L)$ are correction factors to handle windows near the edges of the time series. Receiving as input $\bar{A}_{mi}$, the projection in the ith window is then expressed as:

$$\bar{A}_G[m](i,i) = \bar{A}_{mi}(0,0) \tag{6}$$

$$\bar{A}_G[m](i, t'' : t''') = \bar{A}_{mi}(0, 1+p_l : -p_r) \tag{7}$$

$$\bar{A}_G[m](t' : t'+W, i) = \bar{A}_{mi}(1 :, 0) \tag{8}$$

$$\bar{A}_G[m](t' : t'+W, t'' : t''') = \bar{A}_{mi}(1 :, 1+p_l : -p_r) \tag{9}$$

$$\bar{A}_G[m] = \bar{A}_G[m] \oslash A_{norm} \tag{10}$$

Eq. (6) captures self-attention for CLS_i, Eq. (7) captures attention between CLS_i and BOLD tokens within the receptive field of the ith window, Eq. (8) captures attention between base BOLD tokens and the CLS_i token, and Eq. (9) captures attention between based and fringe BOLD tokens. Note that $(a : -b+1)$ selects between the $(a+1)$-th element from the start and the bth element from the end. In Eq. (10), $A_{norm} \in \mathbb{R}^{(F+T)\times(F+T)}$ is an occurrence matrix that captures the number of times each token occurs across windows (i.e., all entries for a given token that appears in n windows are set to n), and $\oslash$ is Hadamard division used to normalize for repeated token occurrence.

Algorithm 1: Calculation of relevancy map

```
Input: $\{\{A_{0(0)}, ..., A_{0(F-1)}\}, ..., \{A_{M-1(0)}, ..., A_{M-1(F-1)}\}\}$: Set of F
       single-window attention maps from FW-MSA modules across
       M transformer blocks.
Output: $Rel[M]$: Relevancy map.
$Rel[0] = \mathbf{I}_{F+T}$  Initialize relevancy map
for $m = 0 : M-1$ do
    $\bar{A}_G[m] = \mathbf{0}_{F+T}$  Initialize global attention map
    for $i = 0 : F-1$ do
        $\bar{A}_{mi} \leftarrow E_h((\nabla A_{mi} \odot A_{mi})^+)$ Weighted attention map
        $t' \leftarrow F + i \times s$
        $p_l \leftarrow max(0, i \times s - L) - (i \times s - L)$
        $p_r \leftarrow i \times s + W + L - min(T, i \times s + W + L)$
        $t'' \leftarrow t' - L + p_l$
        $t''' \leftarrow t' + W + L - p_r$
        $\bar{A}_G[m](i,i) \leftarrow \bar{A}_{mi}(0,0)$ CLS to CLS attention
        $\bar{A}_G[m](i, t'' : t''') \leftarrow \bar{A}_{mi}(0, 1+p_l : -p_r)$ CLS to BOLD
        $\bar{A}_G[m](t' : t'+W, i) \leftarrow \bar{A}_{mi}(1 :, 0)$ BOLD to CLS
        $\bar{A}_G[m](t' : t'+W, t'' : t''') \leftarrow \bar{A}_{mi}(1 :, 1+p_l : -p_r)$
         BOLD to BOLD
    $\bar{A}_G[m] \leftarrow \bar{A}_G[m] \oslash A_{norm}$  Normalize for repeats
     $Rel[m+1] \leftarrow Rel[m] + \bar{A}_G[m]Rel[m]$  Update rel. map
return $Rel[M]$
```

Next, a token-relevance map $Rel[0]$ that represents the influence of each token onto other tokens in the time series is initialized as an identity matrix in $\mathbb{R}^{(F+T)\times(F+T)}$, implying that each token is initially self-relevant. The normalized attention maps are then used to progressively update the token-relevancy map across transformer blocks where $m \in [0, 1, \ldots, M-1]$:

$$Rel[m+1] = Rel[m] + \bar{A}_G[m]Rel[m]. \tag{11}$$

Following the calculation of the token-relevancy map at the final FW-MSA module, importance weights for input BOLD tokens $w_{imp} \in \mathbb{R}^T$ are finally derived as:

$$w_{imp} = \frac{1}{F} \sum_{i=0}^{F-1} Rel[M](i,\ F:). \tag{12}$$

Importance weight of a BOLD token for the classification task is taken as the across-window average of relevancy scores between the CLS tokens and the given BOLD token.

4. Methods

4.1. Datasets

Demonstrations were performed on fMRI data from the HCP S1200[1] (Van Essen et al., 2013) and ABIDE I releases[2] (Di Martino et al., 2014). In HCP S1200, resting-state fMRI data (HCP-Rest) were analyzed to predict gender, and task-based fMRI data (HCP-Task) were analyzed to predict cognitive task. In ABIDE I, resting-state fMRI data were analyzed to detect Autism Spectrum Disorder (ASD). Details about datasets are provided below.

HCP-Rest: Preprocessed fMRI data from 1200 subjects released by the WU-Minn HCP consortium were analyzed (Glasser et al., 2013).

[1] https://db.humanconnectome.org

[2] https://fcon_1000.projects.nitrc.org/indi/abide/

For each subject, the first session of resting-state scans was used, and incomplete scans with shorter than 1200 time samples were excluded. HCP-Rest comprised a total of 1093 scans from 594 female and 499 male subjects.

HCP-Task: Preprocessed fMRI data from 1200 subjects released by the WU-Minn HCP consortium were analyzed (Glasser et al., 2013). The first session of task-based scans was used, where each subject performed seven different tasks in separate runs: emotion, relational, gambling, language, social, motor, and working memory. Incomplete scans were excluded. HCP-Task comprised a total of 7450 scans from 594 female and 501 male subjects.

ABIDE-I: Preprocessed fMRI data released by the Preprocessed Connectomes Project were analyzed (Craddock et al., 2013; Di Martino et al., 2014). Low-quality scans that did not pass quality checks from all raters were excluded from analysis. ABIDE-I comprised a total of 871 scans from 403 patients with ASD and 468 healthy controls (Abraham et al., 2017).

4.2. Experimental procedures

Experiments were conducted in PyTorch on an NVIDIA RTX 3090 GPU. Modeling was performed via a nested cross-validation procedure with 10 outer and 1 inner folds. Accordingly, subjects were split into non-overlapping training (80%), validation (10%), and test sets (10%). For fair comparison, all competing methods used identical data splits. For each competing method, hyperparameter selection was performed based on performance in the first validation set and selected parameters that showed near optimal performance across all datasets and atlases were used thereafter. The selected parameters included learning rate $\in (10^{-6}, 10^{-1})$, number of epochs $\in (5, 100)$ and mini-batch size $\in (1, 100)$. Training was performed via the Adam optimizer. BolT was trained to minimize the following loss: $L = L_{CE} + \lambda \cdot L_{CWR}$ where L_{CE} is cross-entropy loss, and $\lambda = 0.1$ is the regularization coefficient for CWR loss set via cross-validation. Mean and standard deviation of model performance were reported across the test sets.

For each subject in the training set, fMRI time series were randomly cropped in the temporal dimension to 600 samples for HCP-Rest, 150 samples for HCP-Task, and 100 samples for ABIDE-I to improve stochasticity and learning efficiency (Kim et al., 2021). Functional data were registered to corresponding structural data for each subject, and aligned to the MNI template. ROI definitions were implemented using two public brain atlases: the Schaefer atlas (Schaefer et al., 2018) with 400 regions labeled across seven intrinsic connectivity networks, and the AAL atlas (Tzourio-Mazoyer et al., 2002) with 116 regions.

4.3. Implementation details

In this section, the architectural and hyperparameter details of BolT are summarized. BolT was trained for 20 epochs with a batch size of 32 and an initial learning rate of 10^{-4}. The learning rate was increased to 2x10^{-4} in the first 10 epochs and then gradually decreased to 10^{-5}. A linear projection layer matched the dimensionality of input BOLD responses to the hidden dimensionality of the transformer blocks. A cascade of four blocks was used, each composed of FW-MSA and MLP modules that used layer normalization and skip connections. A hidden dimensionality of 400 and 40 attention heads with 20 dimensions per head were prescribed. A dropout rate of 0.1 was used in both FW-MSA and MLP layers. For FW-MSA modules, given a desired window size W, stride s and fringe length L were set proportionately as follows:

$$s = W\alpha, \quad L = m(W - s)\beta = m(1 - \alpha)W\beta, \tag{13}$$

where $m \in \{0, 1, \dots, M - 1\}$ is the block index, $\alpha \in \mathbb{R}^+$ is the stride coefficient (i.e., proportionality constant), and $\beta \in Z^+$ is the fringe coefficient. Note that the fringe length was progressively grown over transformer blocks as the number of fused tokens increased. Here, cross-validated search for $W \in (10, 200)$, $\alpha \in (0.0, 1.0)$, $\beta \in (0, 3)$ was performed. Hyperparameters were selected as $W = 20$, $\alpha = 0.4$, $\beta = 2$.

4.4. Model complexity

Complexity of models that directly analyze BOLD responses as a temporal sequence depends on the extent of computations performed on the input sequence of BOLD tokens. Assume that the latent dimensionality of tokens is N in a sequence of length T. Recurrent architectures process the tokens in the input sequence serially. As such, a recurrent layer follows a well-known $O(N^2T)$ complexity that scales linearly with the sequence length (Dvornek et al., 2017; Zhao et al., 2020; Xing et al., 2019). A convolution layer also has a linear $O(kN^2T)$ complexity where k denotes the size of the convolution kernel. Meanwhile, vanilla transformers use regular MSA layers that exhaustively compute interactions between all time points in the input sequence, with T queries and T keys of dimensionality N. Thus, vanilla transformers such as BAnD and IFT-Net incur $O(NT^2)$ complexity that scales quadratically with the input sequence length (Nguyen et al., 2020; Zhao et al., 2022).

In contrast, the FW-MSA layer in BolT computes focal interactions between time points in overlapping time windows. Each window has $W + 1$ queries and $W + 2L + 1$ keys. The complexity within a single window is $O(NW^2 + NWL)$. Given a total of $(T - W)/s$ windows for the entire sequence, the overall complexity is $O(NTW^2/s + NTWL/s)$. Selecting s and L as outlined in Eq. (13), BolT incurs $O(NTW^{\frac{(1+\beta(m)(1-\alpha))}{\alpha}})$ complexity that linearly scales with sequence length. Other efficient transformers also show a similar linear trend (Liu et al., 2021; Malkiel et al., 2021; Beltagy et al., 2020; Mehta et al., 2022). For instance, $L = 0$ (no cross-window attention) and $s = \frac{W}{2}$ result in a linear $O(NTW)$ complexity in SwinT (Liu et al., 2021), and $L = 0$ and $s = W$ result in $O(NTW)$ complexity in TFF and HATNet (Malkiel et al., 2021).

4.5. Competing methods

BolT was demonstrated against several state-of-the-art methods for fMRI classification including recent transformer, graph, convolutional and recurrent network models, along with a traditional classifier. The architecture, loss function and learning rate scheduler for each competing method were adopted from the original proposing papers.

SVM: A traditional model operating on static FC features was considered (Abraham et al., 2017). An ℓ_2 regularized model with linear kernel was used. FC features were computed via Pearson's correlation between ROI-level responses. The cross-validated hyperparameter was a regularization weight of $C = 1$.

BrainNetCNN: A CNN model operating on static FC features of fMRI data was considered (Kawahara et al., 2017). ROI-level features were processed with a cascade of two edge-to-edge, one edge-to-node, and one node-to-graph convolutional layers followed by three linear layers. FC features were computed via Pearson's correlation. Cross-validated hyperparameters were 10^{-4} learning rate, 20 epochs, and 16 batch size.

BrainGNN: A GNN operating on static FC features of fMRI data was considered (Li et al., 2021). Taking ROIs as graph nodes, BrainGNN used a cascade of graph convolutional layers to assign nodes to clusters with learned embeddings, and used pooling layers to aggregate information with element-wise score normalization. FC features were computed via partial correlation (Li et al., 2021). Cross-validated hyperparameters were 10^{-2} learning rate, 50 epochs, and 100 batch size.

STAGIN: A GNN operating on dynamic FC features of fMRI data was considered (Kim et al., 2021). STAGIN processed shifted windows across the time series by a four-layer GNN taking ROIs as nodes, deriving edges based on FC features, and calculating node features via a recurrent unit. Features extracted from each graph by a squeeze-excitation readout module were consolidated onto a single latent feature by a transformer. The latent feature was linearly projected to class logits. A window size of 50 was prescribed, FC features were computed via Pearson's correlation (Kim et al., 2021). Cross-validated hyperparameters were 2 $\times 10^{-4}$ learning rate, 40 epochs, 8 batch size.

LSTM: An RNN model operating on BOLD responses was considered (Dvornek et al., 2017). A single-layer LSTM model averaged hidden states across time samples, and performed classification via a sigmoid activation layer. Cross-validated hyperparameters were 10^{-3} learning rate, 30 epochs, 64 batch size.

CNN-LSTM: A convolutional RNN model operating on BOLD responses was considered (Zhao et al., 2020). CNN-LSTM was adopted to use 1D convolutions. Hidden states were averaged across time samples to perform classification via a sigmoid activation layer. Cross-validated hyperparameters were 10^{-3} learning rate, 50 epochs, and 64 batch size.

GC-LSTM: A graph convolutional RNN model operating on FC features was considered (Xing et al., 2019). GC-LSTM was adopted to use the FC features to construct graph adjacency matrices on windowed time series, and it computed recurrent state updates via spectral graph convolution. The window size and stride were set to 120 and 2 s, respectively (Xing et al., 2019). Hidden states were averaged across time samples to perform classification via an MLP followed by a sigmoid activation layer (Xing et al., 2019). Cross-validated hyperparameters were a learning rate of 10^{-3}, 100 epochs, and a batch size of 16.

SwinT: A transformer model with a windowed attention mechanism proposed for computer vision tasks was considered (Liu et al., 2021). For fair comparison, the number and hidden dimensionality of transformer blocks, number of attention heads, and window size were matched with BolT. ROI-level responses were linearly projected to match the hidden dimensionality of the transformer blocks. Output tokens were averaged across windows and linearly projected onto class logits for classification (Beltagy et al., 2020). Cross-validated hyperparameters were 2×10^{-4} learning rate, 30 epochs, and 32 batch size.

Longformer: A transformer model with a windowed attention mechanism proposed for language tasks was considered (Beltagy et al., 2020). Longformer uses sliding window attention for BOLD tokens, and global attention for the CLS token. For fair comparison, the number and hidden dimensionality of transformer blocks, number of attention heads, and window size were matched with BolT. ROI-level responses were linearly projected to match the hidden dimensionality of the transformer blocks. The global CLS token output was linearly projected onto class logits for classification (Beltagy et al., 2020). Cross-validated hyperparameters were 2×10^{-4} learning rate, 40 epochs, and 32 batch size.

BAnD: A hybrid CNN-transformer model operating on BOLD responses was considered (Nguyen et al., 2020). The CNN module suffered from suboptimal learning, so for fair comparison BAnD was implemented with ROI-level inputs as in BolT. ROI-level responses were linearly projected to match the hidden dimensionality of the transformer. No windowing was performed on the time series. Cross-validated hyperparameters were 10^{-4} learning rate, 30 epochs, and 32 batch size.

TFF: A hybrid CNN-transformer model operating on BOLD responses was considered (Malkiel et al., 2021). The CNN module was observed to suffer from suboptimal learning, so for fair comparison TFF was implemented with ROI-level inputs as in BolT. ROI-level responses were linearly projected to match the hidden dimensionality of the transformer. The time series was split into windows of size 20 with shifts of 10, which were processed independently until the final layer where they were averaged (Malkiel et al., 2021). Cross-validated hyperparameters were 10^{-5} learning rate, 30 epochs, and 32 batch size.

IFT-Net: A hybrid CNN-transformer model proposed for medical image analysis was considered (Zhao et al., 2022). IFT-Net was adopted to use 1D convolutional projections in attention layers where ROI-level responses were taken as input channels. ROI-level responses were linearly projected to match the hidden dimensionality of the transformer blocks. The output feature vector was fed to a sigmoid activation layer for classification. Cross-validated hyperparameters were 10^{-4} learning rate, 30 epochs, and 8 batch size.

HATNet: A hybrid CNN-transformer model proposed for medical image analysis was considered (Mehta et al., 2022). HATNet was adopted to use 1D CNN modules with ROI-level responses taken as input channels. The output of the CNN module was split into windows of size 16 and processed via the transformer module for classification (Mehta et al., 2022). Cross-validated hyperparameters were 10^{-4} learning rate, 100 epochs, and 32 batch size.

5. Results

5.1. Ablation studies

We conducted a series of ablation studies to evaluate the contribution of individual design elements in BolT. The design elements included learnable and local (i.e., window-specific) CLS tokens, split time windows, token fusion, cross attention between base and fringe tokens, and cross-window regularization. Starting with a vanilla transformer variant, ablated variants were obtained by progressively introducing individual elements. The vanilla variant omitted all elements including CLS tokens, so classification was performed by linearly projecting the time average of encoded BOLD tokens at the output of the last transformer block. For all variants, the architecture and hyperparameters for the utilized components were matched with BolT. To assess the contribution of a learnable CLS token, a variant was formed by introducing a global CLS token into the vanilla variant. Note that local CLS tokens are not applicable in this case since windowing was omitted. To assess the contribution of windowing, two variants were formed that used split time windows and either global or local CLS tokens, albeit omitted token fusion, cross attention and cross-window regularization. Since cross attention was not used, these variants split the time series into non-overlapping windows by setting stride equal to window size and fringe coefficient to 0 (i.e., $s = W$, $L = 0$). For the global CLS variant, attention for the CLS token was computed with all BOLD tokens in the time series, whereas attention for a BOLD token was restricted to the window it resided in. For the local CLS variant, attention computations for all tokens were restricted to local windows. To assess the contribution of token fusion, a variant was formed that used local CLS tokens, splitting into overlapping windows (i.e., $s = W\alpha$ as in BolT) and token fusion, albeit omitted cross attention and cross-window regularization. Cross attention was omitted by setting $L = 0$. To assess the contribution of cross attention, a variant was formed that used local CLS tokens, splitting into overlapping windows, token fusion and cross-window attention (i.e., $L = m(1 - \alpha)W\beta$), albeit omitted cross-window regularization. Finally, to assess the contribution of cross-window regularization, a variant using all elements (i.e., BolT) was employed where regularization was achieved via adding the loss term in Eq. (4).

Table 1 lists performance metrics for all ablated variants. First, we find that introduction of a learnable, global CLS token into the vanilla variant consistently improves performance metrics, demonstrating the utility of high-level CLS over low-level BOLD tokens in classification tasks. Second, a performance loss is incurred when the global CLS token is used along with split time windows, suggesting that the global token does not adequately capture local representations. In contrast, local CLS tokens in combination with windowing yield a performance boost, indicating the importance of using window-specific CLS tokens to learn local representations. Third, fusion of tokens with repeated occurrence in overlapping windows yields a further improvement. Note that the BOLD token for a given time point appears in a single window and is encoded in a single context for a non-overlapping split, whereas it appears in multiple windows and is encoded in multiple contexts for an overlapping split. Thus, improvements due to token fusion demonstrate the benefit of encoding representations of time points in diverse contexts. Fourth, enabling cross attention between base tokens in each

Table 1
Performance of BolT variants ablated of essential design elements. Ablated elements were learnable CLS tokens (CLS), split time windows (Windowing), token fusion (Fusion), cross-attention between the base and fringe tokens (Cross Attn.), and cross-window regularization (CWR). When windowing is enabled, the annotation (G) denotes utilization of a global CLS token shared across windows, whereas (L) denotes utilization of local CLS tokens in each window. Results are shown based on both the Schaefer and AAL atlases. Accuracy, recall, precision and AUC metrics are reported as mean(std) across test folds. Bold-face indicates the top-performing model.

Atlas	CLS	Windowing	Fusion	Cross Attn.	CWR	Acc. (%)	Rec. (%)	Prec. (%)	AUC (%)
Schaefer	✗	✗	✗	✗	✗	86.35 ±3.56	85.76 ±4.94	84.74 ±4.68	94.59 ±1.57
	✓	✗	✗	✗	✗	89.65 ±3.36	88.38 ±5.63	89.15 ±4.78	96.03 ±1.23
	✓(G)	✓	✗	✗	✗	84.99 ±2.37	89.78 ±4.76	79.95 ±2.68	93.63 ±1.58
	✓(L)	✓	✗	✗	✗	89.65 ±1.85	88.77 ±3.27	88.71 ±3.06	96.79 ±0.90
	✓(L)	✓	✓	✗	✗	90.29 ±1.77	89.57 ±2.54	89.34 ±3.23	97.15 ±1.01
	✓(L)	✓	✓	✓	✗	91.03 ±2.12	89.97 ±2.70	90.42 ±3.30	97.09 ±1.10
	✓(L)	✓	✓	✓	✓	**91.85 ±3.05**	**90.58 ±4.97**	**91.51 ±3.07**	**97.35 ±1.06**
AAL	✗	✗	✗	✗	✗	80.01 ±3.06	78.20 ±5.61	78.22 ±3.88	88.74 ±3.14
	✓	✗	✗	✗	✗	83.12 ±3.58	80.80 ±5.38	82.25 ±5.05	90.88 ±1.91
	✓(G)	✓	✗	✗	✗	78.74 ±4.11	85.00 ±5.07	73.02 ±4.41	88.41 ±3.41
	✓(L)	✓	✗	✗	✗	86.85 ±3.06	87.20 ±4.30	84.78 ±4.65	93.57 ±2.25
	✓(L)	✓	✓	✗	✗	87.04 ±2.48	87.60 ±4.96	84.73 ±3.43	93.80 ±2.13
	✓(L)	✓	✓	✓	✗	87.13 ±3.03	85.80 ±5.39	86.17 ±4.22	94.12 ±1.98
	✓(L)	✓	✓	✓	✓	**87.31 ±2.69**	**86.99 ±4.49**	**85.65 ±4.01**	**94.29 ±2.05**

Table 2
Performance of BolT under varying window sizes W. Results are shown based on the Schaefer and AAL atlases. Accuracy, recall, precision and AUC metrics are reported as mean±std across test folds. Bold-face indicates the top-performing model.

Atlas	Window size	Acc. (%)	Rec. (%)	Prec. (%)	AUC (%)
Schaefer	W = 10	91.39 ± 2.67	89.98 ± 4.28	91.09 ± 2.81	97.49 ± 0.78
	W = 20	**91.85 ± 3.05**	**90.58 ± 4.97**	**91.51 ± 3.07**	97.35 ± 1.05
	W = 80	91.48 ± 2.17	90.18 ± 4.41	91.18 ± 2.86	**97.56 ± 0.95**
	W = 200	88.74 ± 2.73	84.77 ± 5.92	90.04 ± 2.34	96.32 ± 1.44
AAL	W = 10	86.58 ± 2.68	86.20 ± 4.77	84.88 ± 3.98	93.83 ± 2.10
	W = 20	**87.31 ± 2.69**	**86.99 ± 4.49**	**85.65 ± 4.01**	**94.29 ± 2.05**
	W = 80	86.39 ± 4.42	86.00 ± 4.56	84.71 ± 5.88	94.17 ± 2.30
	W = 200	85.57 ± 4.04	82.80 ± 5.60	85.40 ± 5.31	91.98 ± 3.72

window and neighboring fringe tokens increases performance, indicating the importance of this cross-attention mechanism for integration of contextual representations across neighboring windows. Lastly, we find that cross-window regularization that aligns window-specific CLS tokens contributes notably to model performance. This result indicates that the model benefits from coherence of representations in CLS tokens that are averaged across windows to implement classification. Overall, we find that the BolT model including all of its design elements yields the highest performance among all variants.

Compared to the non-windowed variant with CLS, the ablated variant with windowing and local CLS tokens yields a notable performance improvement on the AAL atlas, albeit it shows relatively stable performance on the Schaefer atlas (Table 1). Since the measured fMRI data and modeling procedures were identical for the two atlases, this difference is best attributed to an interaction between window size and the temporal characteristics of ROI responses. In theory, a larger W can be suited to analyze relatively slow varying responses, and a smaller W can be suited to analyze relatively fast varying responses. A power spectral density analysis shows that ROI responses based on the AAL atlas that groups voxels based on anatomical proximity carry higher energy at high temporal frequencies, whereas responses based on the Schaefer atlas that groups voxels based on functional similarity carry higher energy at low temporal frequencies (not reported). In turn, we observe that the performance benefits from split time windows are maximized at $W = 20$ for the AAL atlas, and at $W = 200$ for the Schaefer atlas (not reported). Note, however, that BolT does not only use basic windowing as in the ablated variant, but it also leverages additional window-related design elements including cross-attention between base and fringe tokens, token fusion and cross-window regularization. These elements promote information exchange across separate time windows, increasing the effective temporal receptive field for each window. As such, a small W can serve to sensitively analyze a broad range of responses. Table 2 lists performance metrics for BolT under varying window sizes. We find that optimal or near-optimal performance is attained at a compact window size of $W = 20$ commonly for both atlases. This result suggests that window-related design elements in BolT

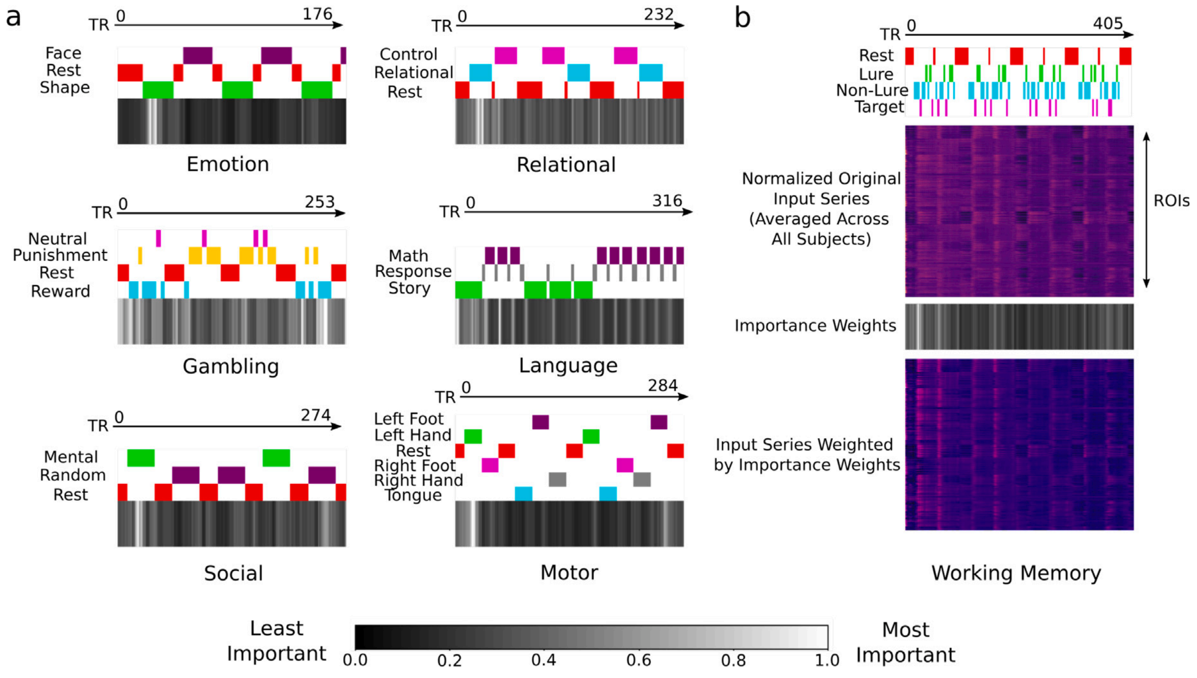

Fig. 3. Importance weights for individual cognitive tasks in HCP-Task averaged across subjects (see colorbar). (a) Results for emotion, relational, gambling, language, social and motor tasks. Subtask annotations are shown to outline task structure during fMRI scans. (b) Input fMRI time series and importance weights for the working memory task. Relevant ROIs and time points are emphasized by masking the input time series with importance weights.

Table 3
Performance of logistic-regression models given as input most important tokens determined via the explanatory technique. Results are also given for logistic-regression models based on random tokens. Boldface indicates the top-performing model in each classification task.

	LR (Important)	LR (Random)
HCP-Rest		
Acc. (%)	**77.74±6.82**	49.59 ± 2.79
Rec. (%)	**71.93±6.43**	26.86 ± 3.76
Prec. (%)	**77.98±8.47**	41.98 ± 4.37
AUC (%)	**83.70±5.64**	45.70 ± 1.25
HCP-Task		
Acc. (%)	**94.83±1.44**	15.14 ± 1.18
Rec. (%)	**94.81±1.44**	15.15 ± 1.18
Prec. (%)	**94.89±1.43**	15.51 ± 1.47
AUC (%)	**99.22±0.33**	51.19 ± 0.73
ABIDE-I		
Acc. (%)	**60.44±3.55**	51.23 ± 4.05
Rec. (%)	**46.81±4.80**	33.11 ± 5.36
Prec. (%)	**59.71±4.84**	47.68 ± 6.31
AUC (%)	**62.26±4.65**	49.33 ± 5.34

collectively introduce a degree of reliability against varying temporal frequency characteristics of BOLD responses.

Next, we examined the efficacy of the explanatory technique that computes an importance weight for each BOLD token (see Fig. 3). This importance weight is supposed to reflect the degree of discriminative information captured by the token for the respective detection task. We reasoned that if the explanatory technique computes reasonable weights, significant detection should be possible based on a subset of highly important tokens. To test this prediction, BOLD tokens in the time series were ordered according to their importance weights. ROI definitions based on the Schaefer atlas were used for this analysis, since they yielded better performance in BolT. Logistic-regression models were then built for the same detection task given as input a subset of five consecutive tokens (Tagliazucchi et al., 2012, 2011). Table 3 lists detection performance based on the most important subset of

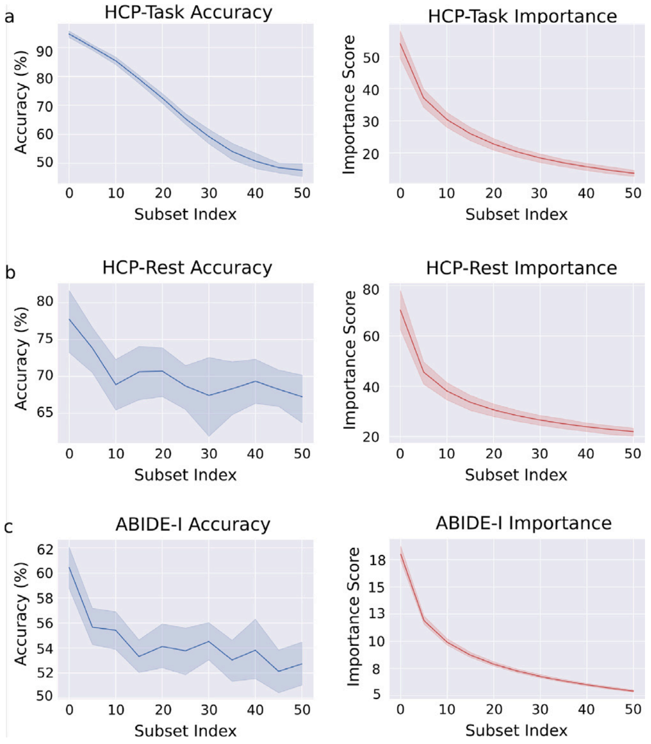

Fig. 4. Accuracy of logistic-regression models receiving as input a subset of five important tokens, while the total importance of the subset is systematically varied. Tokens were ordered according to their importance weights, different subsets of five consecutive tokens in the ordered list were selected. Results are shown for varying subset index for (a) HCP-Task, (b) HCP-Rest, (c) ABIDE-I. Subset index refers to the offset within the ordered list for the selected subset. Importance score was taken as the average importance weight of tokens in a given subset normalized by the minimum importance weight within the time series. As such, total importance of the subset and detection accuracy show a general decrease with increasing subset index.

Table 4
Performance of competing methods and BolT using Schaefer atlas for gender detection on the HCP-Rest dataset, task detection on the HCP-Task dataset and disease detection on the ABIDE-I dataset. Metrics are reported as mean±std across test folds. Boldface indicates the top-performing model in terms of each metric in individual classification tasks.

	HCP-Rest (Schaefer)				HCP-Task (Schaefer)				ABIDE-I (Schaefer)			
	Acc.(%)	Rec.(%)	Prec.(%)	AUC(%)	Acc.(%)	Rec.(%)	Prec.(%)	AUC(%)	Acc.(%)	Rec.(%)	Prec.(%)	AUC(%)
SVM	75.92 ±4.78	65.13 ±7.41	78.63 ±6.27	85.34 ±4.59	95.57 ±0.87	95.58 ±0.88	95.61 ±0.86	99.78 ±0.07	65.72 ±4.11	53.97 ±7.33	66.26 ±6.68	71.33 ±4.62
BrainNetCNN	84.16 ±3.73	82.97 ±6.63	82.39 ±3.04	90.95 ±3.85	96.42 ±0.82	96.42 ±0.87	96.52 ±0.82	99.81 ±0.06	66.78 ±3.90	63.82 ±8.73	65.50 ±4.63	73.15 ±5.62
BrainGNN	79.14 ±3.90	78.96 ±5.43	76.89 ±6.19	86.31 ±2.51	93.10 ±1.40	93.11 ±1.43	93.80 ±1.07	99.67 ±0.12	59.87 ±5.80	47.86 ±12.21	57.79 ±6.50	63.86 ±6.63
STAGIN	82.51 ±3.69	84.16 ±5.61	79.20 ±4.71	87.63 ±3.53	99.27 ±0.43	99.26 ±0.43	99.27 ±0.42	99.98 ±0.01	61.70 ±3.56	41.69 ±7.37	63.63 ±7.03	64.71 ±6.02
LSTM	81.59 ±4.03	82.16 ±3.00	78.99 ±6.37	90.46 ±2.37	98.35 ±0.65	98.34 ±0.65	98.41 ±0.58	99.94 ±0.04	64.55 ±5.41	57.14 ±16.20	62.85 ±6.98	68.88 ±5.87
CNN-LSTM	80.77 ±3.83	79.57 ±9.34	80.38 ±9.27	88.47 ±3.39	99.04 ±0.59	99.04 ±0.59	99.07 0.56	99.96 ±0.05	65.49 ±5.70	57.31 ±9.40	64.79 ±7.57	71.40 ±5.41
GC-LSTM	83.99 ±4.85	76.90 ±16.09	87.64 ±5.14	93.85 ±1.35	98.16 ±0.59	98.16 ±0.59	98.20 ±0.59	99.96 ±0.02	62.77 ±5.46	60.24 ±19.56	62.11 ±10.96	68.28 ±4.84
SwinT	79.41 ±2.49	77.35 ±4.54	77.67 ±3.63	87.49 ±1.71	99.54 ±0.27	99.54 ±0.27	99.55 ±0.27	**99.99** **±0.00**	68.56 ±4.74	60.75 ±7.60	68.27 ±6.56	74.13 ±4.15
Longformer	83.24 ±3.34	79.15 ±8.49	84.03 ±5.45	92.41 ±2.81	99.46 ±0.39	99.46 ±0.38	99.47 ±0.38	**99.99** **±0.00**	67.85 ±4.58	57.15 ±15.79	70.09 ±7.13	74.87 ±3.99
BaND	83.61 ±4.03	84.54 ±6.41	80.87 ±5.42	92.62 ±2.20	99.24 ±0.46	99.24 ±0.46	99.26 ±0.43	99.98 ±0.01	65.48 ±3.04	58.07 ±7.93	64.47 ±4.76	72.10 ±4.38
TFF	87.36 ±3.68	84.17 ±4.21	87.89 ±5.09	94.79 ±2.39	99.08 ±0.25	99.08 ±0.25	99.11 ±0.24	99.98 ±0.02	66.73 ±5.33	42.71 ±16.49	**78.93** **±10.19**	75.44 ±3.94
IFT-Net	82.97 ±2.84	79.76 ±10.28	83.72 ±7.27	93.12 ±2.05	97.58 ±3.62	97.58 ±3.62	98.08 ±2.59	99.91 ±0.16	61.88 ±5.52	51.59 ±23.88	63.82 ±9.42	68.50 ±6.44
HATNet	85.72 ±2.64	84.38 ±5.89	84.57 ±3.68	93.66 ±2.26	99.37 ±0.33	99.36 ±0.33	99.38 ±0.32	**99.99** **±0.00**	64.66 ±4.53	57.67 ±6.88	62.99 ±5.76	68.92 ±4.95
BolT	**91.85** **±3.05**	**90.58** **±4.97**	**91.51** **±3.07**	**97.35** **±1.06**	**99.66** **±0.35**	**99.66** **±0.35**	**99.67** **±0.34**	**99.99** **±0.00**	**71.28** **±4.62**	**64.85** **±7.94**	71.32 ±7.35	**77.56** **±3.44**

tokens against that based on a subset of randomly selected tokens. While random tokens perform near chance level, important tokens achieve substantially higher performance. We also reasoned that detection performance should scale with the overall importance of the selected token subset. To examine this issue, separate logistic-regression models were built while the overall importance of the token subset was systematically reduced (Fig. 4). Detection performance elevates near-monotonically with increasing levels of token importance. Taken together, these results indicate that the importance weights returned by the explanatory technique closely reflect the contribution of individual tokens to model decisions.

5.2. Comparative demonstration of BolT

We demonstrated BolT for three main tasks in fMRI analysis: gender detection on HCP-Rest, cognitive task detection on HCP-Task, and disease detection on ABIDE-I datasets. BolT was demonstrated against state-of-the-art traditional (SVM), CNN (BrainNetCNN), GNN (BrainGNN, STAGIN), RNN (LSTM, CNN-LSTM, GC-LSTM), and transformer (SwinT, Longformer, BaND, TFF, IFT-Net, HATNet) baselines. Demonstrations were performed using ROI definitions extracted via two different brain atlases. Performance metrics for competing methods for Schaefer atlas are listed in Table 4, and those for AAL atlas are listed in Table 5. For each detection task and based on each atlas, BolT outperforms all competing methods in each metric ($p < 0.05$, Wilcoxon signed-rank test), except for TFF that offers higher precision on ABIDE-I, SwinT that offers similar recall on ABIDE-I (AAL atlas), and SwinT, Longformer, and HATNet that offer similar AUC on HCP-Task. On average across atlases in gender detection, BolT improves (accuracy, recall, precision, AUC) by (8.39, 11.13, 7.41, 5.87)% over transformer baselines, (10.96, 13.78, 9.58, 7.95)% over RNN baselines, (12.67, 11.85, 14.49, 12.10)% over GNN baselines, (11.92, 13.59, 12.78, 10.84)% over the CNN baseline, and (15.76, 23.51, 13.99, 14.00)% over the traditional baseline. In cognitive task detection, BolT achieves improvements of (1.14, 1.14, 1.00, 0.03)% over transformer baselines, (2.35, 2.35, 2.21, 0.12)% over RNN baselines, (6.84, 6.82, 6.18, 0.82)% over GNN baselines, (6.09, 6.09, 5.98, 0.56)% over the CNN baseline, and (7.75, 7.73, 7.67, 0.82)% over the traditional baseline. Finally, in disease detection, BolT achieves improvements of (5.11, 10.56, -, 4.88)% over transformer baselines, (6.44, 9.63, 5.86, 7.22)% over RNN baselines, (8.58, 12.53, 9.43, 10.69)% over GNN baselines, (4.31, 0.64, 5.73, 4.46)% over the CNN baseline, and (4.05, 8.15, 4.23, 4.68)% over the traditional baseline. Taken together, these results indicate that BolT enables significant performance benefits in detection tasks over prior traditional and DL methods.

In general, we observe that DL models yield superior performance to the traditional SVM baseline on HCP-Rest and HCP-Task, whereas SVM outperforms CNN, GNN, RNN, and a subset of transformer baselines in disease detection on ABIDE-I. Note that HCP-Rest and HCP-Task were acquired using relatively standardized protocols and scanner hardware in a compact set of imaging sites. In contrast, ABIDE-I was curated by aggregating data from a larger number of sites with more substantial variations in imaging protocols and hardware. In turn, the resultant data heterogeneity can limit generalization performance for DL methods with relatively high complexity, while the simpler SVM method starts performing competitively. That said, we observe that windowed transformer models including BolT still outperform SVM in this case, implying a degree of reliability against data heterogeneity due to the generalization capabilities of self-attention operators combined with local sensitivity from split time windows. We also observe that all competing methods yield notably higher performance on HCP-Task, compared to HCP-Rest and ABIDE-I. This is expected as detecting divergent cognitive tasks from task-based fMRI scans that elicit responses in

Table 5
Performance of competing methods and BolT using AAL atlas for gender detection on the HCP-Rest dataset, task detection on the HCP-Task dataset and disease detection on the ABIDE-I dataset. Metrics are reported as mean±std across test folds. Boldface indicates the top-performing model in terms of each metric in individual classification tasks.

	HCP-Rest (AAL)				HCP-Task (AAL)				ABIDE-I (AAL)			
	Acc.(%)	Rec.(%)	Prec.(%)	AUC(%)	Acc.(%)	Rec.(%)	Prec.(%)	AUC (%)	Acc.(%)	Rec.(%)	Prec.(%)	AUC(%)
SVM	71.71 ±3.76	65.40 ±11.86	70.55 ±3.28	78.30 ±4.09	88.11 ±1.46	88.14 ±1.46	88.25 ±1.38	98.54 ±0.31	65.60 ±3.14	55.00 ±7.40	64.10 ±5.77	71.16 ±4.18
BrainNetCNN	71.16 ±4.91	67.40 ±9.21	69.20 ±6.47	78.99 ±5.12	90.56 ±1.06	90.58 ±1.07	90.73 ±1.04	99.04 ±0.27	64.01 ±6.04	60.24 ±9.33	61.86 ± 6.75	69.79 ±6.16
BrainGNN	69.79 ±4.59	72.40 ±9.70	65.25 ±4.36	77.85 ±4.59	79.73 ±5.92	79.85 ±5.88	81.72 ±4.77	97.05 ±1.29	61.40 ±4.66	58.30 ±10.86	58.40 ±6.74	65.68 ±5.61
STAGIN	76.18 ±3.10	72.20 ±9.22	75.00 ±3.75	83.07 ±3.15	98.87 ±0.60	98.86 ±0.61	98.91 ±0.57	99.95 ±0.03	61.52 ±3.49	52.69 ±7.38	60.12 ±4.94	66.68 ±4.36
LSTM	73.25 ±4.48	67.19 ±8.44	73.10 ±7.59	81.96 ±2.90	96.96 ±0.69	96.97 ±0.69	97.06 ±0.64	99.88 ±0.06	63.06 ±3.96	45.81 ±22.76	64.05 ±13.93	70.25 ±3.85
CNN-LSTM	74.81 ±3.15	66.40 ±9.20	76.25 ±5.28	82.88 ±3.30	97.77 ±0.60	97.76 ±0.61	97.83 ±0.57	99.91 ±0.05	63.65 ±5.42	49.47 ±17.04	66.49 ±11.86	68.78 ±4.13
GC-LSTM	77.27 ±5.29	77.79 ±14.57	77.61 ±11.50	89.59 ±1.93	93.14 ±1.39	93.14 ±1.39	93.74 ±1.09	99.57 ±0.14	60.05 ±5.36	48.27 ±19.80	61.04 ±9.04	64.67 ±6.90
SwinT	78.37 ±4.00	76.60 ±6.69	76.14 ±4.07	84.84 ±3.58	99.23 ±0.38	99.23 ±0.38	99.25 ±0.36	**99.99** **±0.01**	66.78 ±4.25	**60.48** **±6.38**	65.56 ±5.57	72.92 ±4.30
Longformer	76.28 ±4.30	63.00 ±13.94	83.53 ±9.68	87.80 ±3.10	99.11 ±0.42	99.11 ±0.43	99.13 ±0.41	**99.99** **±0.00**	64.77 ±4.71	58.69 ±18.18	64.99 ±9.42	71.42 ±4.70
BaND	78.55 ±4.12	70.40 ±8.38	80.46 ±5.18	87.39 ±2.84	98.16 ±0.45	98.16 ±0.45	98.20 ±0.41	99.93 ±0.03	63.12 ±3.60	45.18 ±12.53	65.25 ±6.33	68.65 ±3.61
TFF	82.57 ±4.05	81.80 ±8.59	81.34 ±7.41	91.14 ±2.98	97.43 ±1.00	97.43 ±1.00	97.55 ±0.90	99.90 ±0.07	65.84 ±3.87	46.30 ±15.83	**73.30** **±8.86**	74.14 ±4.16
IFT-Net	77.72 ±4.85	76.00 ±11.76	76.57 ±7.28	86.69 ±2.62	95.22 ±5.58	95.17 ±5.67	96.19 ±3.80	99.87 ±0.19	58.26 ±4.72	32.11 ±24.85	54.23 ±25.90	64.99 ±4.64
HATNet	78.37 ±2.15	74.60 ±5.58	77.23 ±1.54	87.43 ±2.96	97.97 ±0.35	97.96 ±0.35	98.02 ±0.32	99.96 ±0.02	61.16 ±5.64	54.59 ±7.15	58.79 ±6.55	66.52 ±5.10
BolT	**87.31** **±2.69**	**86.99** **±4.49**	**85.65** **±4.01**	**94.29** **±2.05**	**99.52** **±0.39**	**99.52** **±0.40**	**99.54** **±0.38**	**99.99** **±0.00**	**68.14** **±2.81**	**60.49** **±4.22**	67.52 ±4.07	**74.30** **±3.69**

largely non-overlapping brain networks is relatively easier compared to detection tasks on resting-state fMRI scans. Here, we preferred to report HCP-Task since it is a highly relevant, benchmark dataset that is frequently reported in methodology studies on task-based fMRI analysis. Yet, future studies are warranted to examine the utility of BolT in detecting BOLD-response differences among more similar cognitive tasks driving partly overlapping brain networks.

5.3. Explainability of BolT

To interpret the spatio-temporal patterns of brain activation that contribute to BolT's decisions, we employed the explanatory technique to calculate token importance weights. Importance weights for each cognitive task in HCP-Task are shown in Fig. 3. Landmark time points of high importance closely align with transitions in the temporal structure of task variables following an offset due to hemodynamic delay. For instance, periods of target maintenance following target appearance are attributed high importance in the working memory task, corresponding to abrupt changes in activation (Tagliazucchi et al., 2011).

We then leveraged the landmark time points to identify brain regions critical for the detection tasks. To do this, a logistic-regression model was trained on top-five most important BOLD tokens, and model weights were taken to reflect the importance of individual ROIs for task performance (Rahman et al., 2022). As shown in Fig. 5 for gender detection, we find important ROIs across the attention and somatosensory networks in male subjects, and ROIs in prefrontal/frontal cortices and default mode network (DMN) in female subjects. This is consistent with previous reports on stronger FC features across sensorimotor cortices in males and across DMN in females (Ritchie et al., 2018; Filippi et al., 2013). We further find important ROIs in visual networks for both genders. This result is aligned with a recent report suggesting that responses in visual regions might implicitly represent gender-discriminating information (Kim et al., 2021). As shown in Fig. 6 for task detection, brain regions implicated with the target task are attributed high importance (e.g., sensorimotor regions in the Motor task, temporal regions in the Language task). As shown in Fig. 7 for ASD detection, we find important ROIs in healthy controls across the frontal-parietal network (FPN), thought to mediate goal-oriented, cognitively demanding behavior (Uddin et al., 2019). In contrast, ASD patients manifest important ROIs across DMN, with commonly reported over-activation in ASD (Buckner et al., 2008; Abraham et al., 2017; Chen et al., 2021). Taken together, these results indicate that BolT effectively captures task-relevant patterns of brain activation in both normal and disease states.

6. Discussion

Here, we introduced a transformer architecture that efficiently captures local-to-global representations of time series to perform detection tasks based on fMRI scans. The proposed architecture learns latent representations of fMRI data via a novel fused window attention mechanism that incorporates long-range context with linear complexity in terms of scan length. Detection is then performed based on learned high-level classification tokens regularized across time windows. Demonstrations were performed on resting-state and task-based fMRI data with superior performance against state-of-the-art baselines including convolutional, graph and transformer models.

In this study, we primarily built classification models with categorical output variables for gender, cognitive task, and disease. To improve classification performance, learnable CLS tokens were included that provide a condensed high-level representation of corresponding time windows. Note that the human brain does not only represent categorical variables, but it is also assumed to carry information regarding

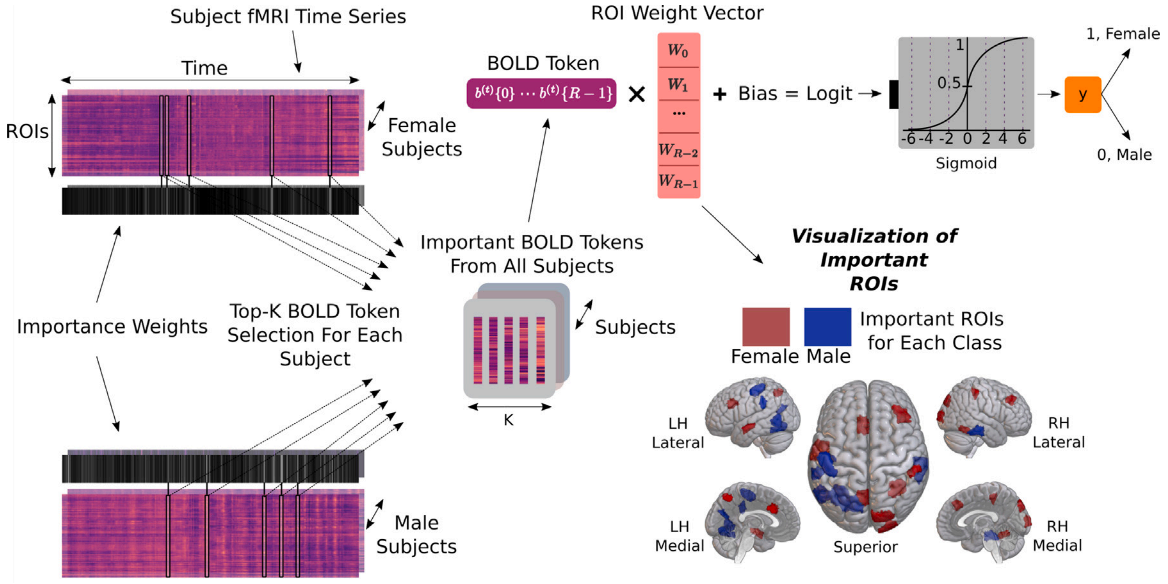

Fig. 5. Landmark time points (i.e. BOLD tokens) selected by BolT were used to identify brain regions critical for gender detection in HCP-Rest. A collection of K = 5 tokens were retrieved from each subject, characterizing responses across R ROIs. Next, a logistic regression model was trained to map the tokens in landmark time points onto the associated output class. Model weights reflect each ROI's contribution to the classification decision. For each class, the top 2 percent of most influential ROIs were visualized (i.e. female in red color, male in blue color).

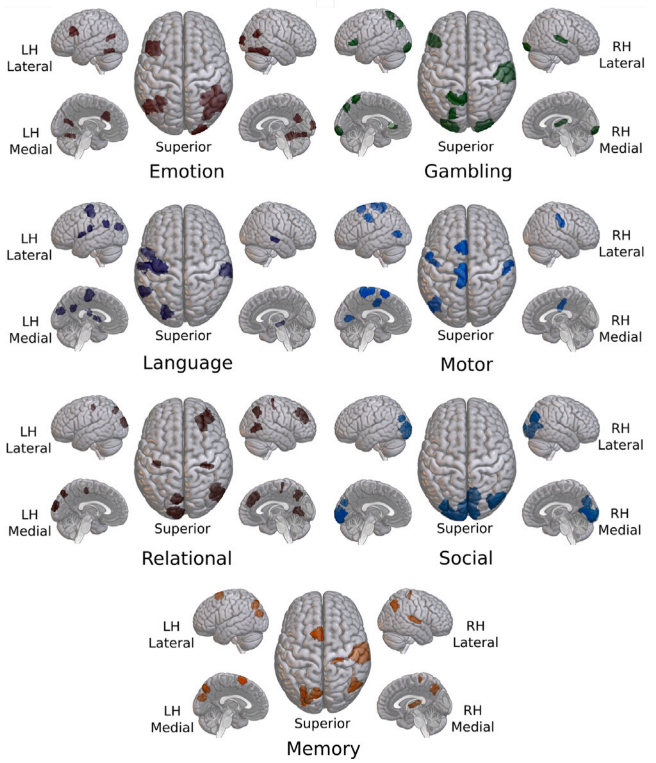

Fig. 6. Critical ROIs for cognitive task detection in HCP-Task. For each task, the top 2 percent of most influential ROIs were visualized. Elevated BOLD responses in highlighted ROIs imply the presence of the associated task.

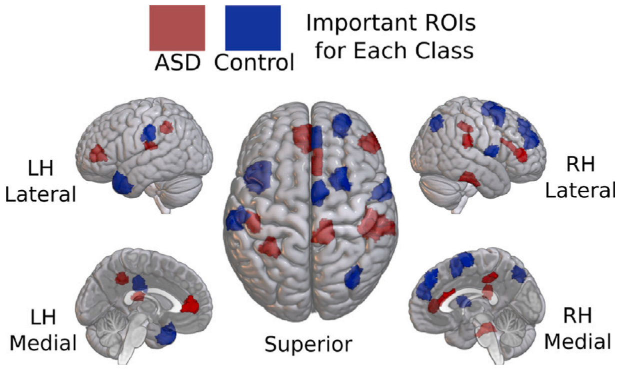

Fig. 7. Critical ROIs for ASD detection in ABIDE-I. The top 2 percent of most influential ROIs were visualized for ASD patients and healthy controls. Elevated BOLD responses in red ROIs imply the presence of the ASD condition, whereas elevated responses in blue ROIs imply a healthy condition.

continuous stimulus or task features (Çukur et al., 2013). To analyze cortical representations of such continuous features, BolT can be adapted to instead build regression models (Nishimoto et al., 2011; VanRullen and Reddy, 2019; Li et al., 2018). To do this, latent representations of BOLD tokens in downstream layers of BolT can be coupled with linear or ReLU activation functions. This can enable BolT to decode continuous stimulus or task variables from fMRI scans.

A common approach to interrogate brain function in computational neuroimaging studies is to build decoding models that predict stimulus or task features given as input measured brain activations (LaConte, 2011; Andersson et al., 2011). Following this framework, here we used BolT to build decoding models that detect external features based on measured BOLD responses. An alternative framework rests on analysis of brain function by building encoding models that predict brain activations given as input stimulus/task features (Nishimoto, 2021; Celik et al., 2021; Shahdloo et al., 2022; Anderson et al., 2016; Ngo et al., 2022). In cognitive neuroimaging studies, the experimental time course for the stimulus and/or cognitive task can be taken as input to BolT, and voxel-wise regression models can be built to estimate measured BOLD responses. It remains important future work to assess the efficacy of BolT in training encoding models.

Literature suggests that resting-state fMRI scans carry idiosyncratic information regarding disease progression in neurodevelopmental disorders (Uddin et al., 2010; Hohenfeld et al., 2018). Based on this literature, we considered ASD detection using solely information from resting-state fMRI scans. Recent studies suggest that auxiliary information on patient demographics or scan protocols might help facilitate disease detection (Dvornek et al., 2018). Moreover, some neurological diseases such as Alzheimer's or dementia have complementary imaging signatures in other modalities such as structural or diffusion-weighted MRI (Román and Pascual, 2012). Thus, it is reasonable to expect that disease detection performance with BolT can be further improved by incorporating auxiliary information as well as additional imaging modalities. Auxiliary information can be integrated via bypass channels near the output layers of BolT, whereas additional imaging modalities can be incorporated as added input channels alongside fMRI data.

As commonly practiced in many fMRI studies, here we first normalized each subject's brain volume onto an anatomical template, and then used an anatomical atlas to define brain ROIs. Average BOLD responses in individual ROIs were then provided as input to BolT. Note that this approach ensures relatively consistent and comprehensive ROI definitions across subjects, permitting analyses in brain regions that do not have well established functional-localization procedures (Flandin et al., 2002). Yet, spatial registration to a common template involves a potentially lossy transformation of fMRI data. Such losses can be mitigated by defining ROIs in the brain spaces of individual subjects as opposed to a template. To do this, the registration transform between the subject and template brain spaces can be estimated. ROI boundaries in the template brain space can then be backprojected onto the individual subject brain space by inverting the estimated transformation (Shahdloo et al., 2020, 2022). Alternatively, a CNN model can also incorporated in BolT to perform spatial encoding of volumetric MRI data prior to processing with the transformer blocks (Nguyen et al., 2020; Malkiel et al., 2021).

Here we trained all competing models from scratch on fMRI data from several hundred subjects for each detection task. Given their relatively higher complexity against convolutional models, transformers are generally considered to require substantial datasets for successful learning (Dosovitskiy et al., 2020; Güngör et al., 2022). In applications where only compact datasets are available, pre-training and transfer learning procedures can be adopted to initialize the network weights in transformer architectures (Devlin et al., 2018; Dalmaz et al., 2022b). Reliable augmentation via image synthesis based on advanced procedures such as diffusion modeling can also help alleviate data scarcity (Dar et al., 2022; Özbey et al., 2022). Alternatively, complexity of self-attention modules can be mitigated by replacing regular dot-product attention operators with efficient kernelized operators (Zhang et al., 2022). Federated learning across multiple institutions might facilitate learning from large, diverse datasets without introducing privacy risks (Elmas et al., 2022; Dalmaz et al., 2022a). Lastly, unsupervised learning strategies can also be adopted to permit training on partially labeled fMRI datasets from a larger subject cohort (Malkiel et al., 2021; Korkmaz et al., 2022). A systematic exploration of the data efficiency of BolT against competing models remains an important topic for future research.

7. Conclusion

In this study, we introduced a novel transformer model to improve classification performance on fMRI time series. BolT leverages fused window attention to capture local interactions among temporally-overlapped time windows, and hierarchically grows window overlap to capture global representations. Token fusion and cross-window regularization are used to effectively integrate latent representations across the time series. Here, demonstrations were performed for gender and disease detection from resting-state fMRI and task detection from task-based fMRI. Furthermore, an explanatory technique was devised to interpret model decisions in terms of landmark time points and brain regions. Collectively, the proposed approach holds great promise for sensitive and explainable analysis of multi-variate fMRI data. BolT may help detect other neurological disorders with characteristic influences on fMRI activation patterns, and classification of more intricate task variables during cognitive processing.

CRediT authorship contribution statement

Hasan A. Bedel: Conceptualization, Methodology, Data curation, Software, Formal analysis, Visualization, Investigation, Validation, Writing – original draft, Writing – review & editing. **Irmak Sivgin:** Methodology, Data curation, Software, Formal analysis, Visualization, Investigation, Validation, Writing – original draft. **Onat Dalmaz:** Formal analysis, Visualization, Investigation, Validation, Writing – original draft. **Salman U.H. Dar:** Formal analysis, Visualization, Investigation, Validation, Writing – original draft. **Tolga Çukur:** Conceptualization, Methodology, Supervision, Funding acquisition, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The url for the code for the study is provided in the manuscript. The public repositories containing data have been listed in the manuscript.

Acknowledgments

This study was supported in part by a TUBITAK BIDEB scholarships, Turkey awarded to H.A. Bedel and O. Dalmaz, and by a TUBA GEBIP 2015 fellowship, Turkey, a BAGEP 2017 fellowship, Turkey, and a TUBITAK 121N029, Turkey grant awarded to T. Çukur. Funds have been provided by the European Joint Programme Neurodegenerative Disease Research (JPND) 2020 call "Novel imaging and brain stimulation methods and technologies related to Neurodegenerative Diseases" for the Neuripides project 'Neurofeedback for self-stImulation of the brain as therapy for ParkInson Disease'. The Neuripides project has received funding from the following funding organizations under the aegis of JPND: The Netherlands, The Netherlands Organization for Health Research and Development (ZonMw); Germany, Federal Ministry of Education and Research (BMBF); Czech Republic, Ministry of Education, Youth and Sports (MEYS); France, French National Research Agency (ANR); Canada, Canadian Institutes of Health Research (CIHR); Turkey, Scientific and Technological Research Council of Turkey (TUBITAK).

References

Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G., 2017. Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. NeuroImage 147, 736–745.

Allen, E.A., Damaraju, E., Plis, S.M., Erhardt, E.B., Eichele, T., Calhoun, V.D., 2014. Tracking whole-brain connectivity dynamics in the resting state. Cerebral Cortex 24 (3), 663–676.

Alon, U., Yahav, E., 2020. On the bottleneck of graph neural networks and its practical implications. arXiv:2006.05205.

Anderson, J.S., Nielsen, J.A., Ferguson, M.A., Burback, M.C., Cox, E.T., Dai, L., Gerig, G., Edgin, J.O., Korenberg, J.R., 2013. Abnormal brain synchrony in down syndrome. NeuroImage: Clinical 2, 703–715.

Anderson, A.J., Zinszer, B.D., Raizada, R.D., 2016. Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. NeuroImage 128, 44–53.

Andersson, P., Pluim, J.P., Siero, J.C., Klein, S., Viergever, M.A., Ramsey, N.F., 2011. Real-time decoding of brain responses to visuospatial attention using 7T fMRI. PLoS One 6 (11), e27638.

Beltagy, I., Peters, M.E., Cohan, A., 2020. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

Buckner, R.L., Andrews-Hanna, J.R., Schacter, D.L., 2008. The brain's default network: anatomy, function, and relevance to disease. Ann. New York Acad. Sci. 1124 (1), 1–38.

Bullmore, E., Rabe-Hesketh, S., Morris, R., Williams, S., Gregory, L., Gray, J., Brammer, M., 1996. Functional magnetic resonance image analysis of a large-scale neurocognitive network. NeuroImage 4 (1), 16–33.

Celik, E., Keles, U., Kiremitçi, İ., Gallant, J.L., Çukur, T., 2021. Cortical networks of dynamic scene category representation in the human brain. Cortex 143, 127–147.

Chefer, H., Gur, S., Wolf, L., 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: IEEE/CVF International Conference on Computer Vision. pp. 397–406.

Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., Sun, X., 2020. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In: AAAI Conference on Artificial Intelligence, vol.34(04). pp. 3438–3445.

Chen, Y.-Y., Uljarevic, M., Neal, J., Greening, S., Yim, H., Lee, T.-H., 2021. Excessive functional coupling with less variability between salience and default-mode networks in Autism Spectrum Disorder. Biol. Psychiatry.

Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., Khundrakpam, B.S., Lewis, J.D., Li, Q., Milham, M., et al., 2013. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. Front. Neuroinform. 7.

Çukur, T., Nishimoto, S., Huth, A.G., Gallant, J.L., 2013. Attention during natural vision warps semantic representation across the human brain. Nature Neurosci. 16 (6), 763–770.

Dalmaz, O., Mirza, U., Elmas, G., Özbey, M., Dar, S.U., Ceyani, E., Avestimehr, S., Çukur, T., 2022a. One model to unite them all: personalized federated learning of multi-contrast MRI synthesis. arXiv preprint arXiv:2207.06509.

Dalmaz, O., Yurt, M., Çukur, T., 2022b. ResViT: Residual vision transformers for multi-modal medical image synthesis. IEEE Trans. Med. Imaging 41 (10), 2598–2614.

Dar, S.U., Öztürk, Ş., Korkmaz, Y., Elmas, G., Özbey, M., Güngör, A., Çukur, T., 2022. Adaptive diffusion priors for accelerated MRI reconstruction. arXiv preprint arXiv:2207.05876.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage 43 (1), 44–58.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry 19 (6), 659–667.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929.

Duncan, J.S., Insana, M.F., Ayache, N., 2019. Biomedical imaging and analysis in the age of big data and deep learning [scanning the issue]. Proc. IEEE 108 (1), 3–10.

Dvornek, N.C., Ventola, P., Duncan, J.S., 2018. Combining phenotypic and resting-state fMRI data for autism classification with recurrent neural networks. In: IEEE International Symposium on Biomedical Imaging. IEEE, pp. 725–728.

Dvornek, N.C., Ventola, P., Pelphrey, K.A., Duncan, J.S., 2017. Identifying autism from resting-state fMRI using long short-term memory networks. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 362–370.

Elmas, G., Dar, S.U., Korkmaz, Y., Ceyani, E., Susam, B., Özbey, M., Avestimehr, S., Çukur, T., 2022. Federated learning of generative image priors for MRI reconstruction. IEEE Trans. Med. Imaging http://dx.doi.org/10.1109/TMI.2022.3220757.

Fan, L., Su, J., Qin, J., Hu, D., Shen, H., 2020. A deep network model on dynamic functional connectivity with applications to gender classification and intelligence prediction. Front. Neurosci. 14, 881.

Filippi, M., Valsasina, P., Misci, P., Falini, A., Comi, G., Rocca, M.A., 2013. The organization of intrinsic brain activity differs between genders: A resting-state fMRI study in a large cohort of young healthy subjects. Hum. Brain Mapp. 34 (6), 1330–1343.

Flandin, G., Kherif, F., Pennec, X., Malandain, G., Ayache, N., Poline, J.-B., 2002. Improved detection sensitivity in functional MRI data using a brain parcelling technique. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 467–474.

Gadgil, S., Zhao, Q., Pfefferbaum, A., Sullivan, E.V., Adeli, E., Pohl, K.M., 2020. Spatio-temporal graph convolution for resting-state fMRI analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 528–538.

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

Greicius, M., 2008. Resting-state functional connectivity in neuropsychiatric disorders. Curr. Opin. Neurol. 21 (4), 424–430.

Güngör, A., Askin, B., Soydan, D.A., Saritas, E.U., Top, C.B., Çukur, T., 2022. TranSMS: Transformers for super-resolution calibration in magnetic particle imaging. IEEE Trans. Med. Imaging 41 (12), 3562–3574.

Haxby, J.V., 2012. Multivariate pattern analysis of fMRI: the early beginnings. NeuroImage 62 (2), 852–855.

Heinsfeld, A.S., Franco, A.R., Craddock, R.C., Buchweitz, A., Meneguzzi, F., 2018. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. NeuroImage 17, 16–23.

Hillman, E.M., 2014. Coupling mechanism and significance of the BOLD signal: a status report. Annu. Rev. Neurosci. 37, 161–181.

Hohenfeld, C., Werner, C.J., Reetz, K., 2018. Resting-state connectivity in neurodegenerative disorders: Is there potential for an imaging biomarker? NeuroImage 18, 849–870.

Hojjati, S.H., Ebrahimzadeh, A., Khazaee, A., Babajani-Feremi, A., Initiative, A.D.N., et al., 2017. Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM. J. Neurosci. Methods 282, 69–80.

Hu, Z., Shi, P., 2006. Interregional functional connectivity via pattern synchrony. In: International Conference on Control, Automation, Robotics and Vision. IEEE, pp. 1–6.

Hu, Z., Shi, P., 2007. Nonlinear analysis of BOLD signal: biophysical modeling, physiological states, and functional activation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 734–741.

Huang, H., Hu, X., Zhao, Y., Makkie, M., Dong, Q., Zhao, S., Guo, L., Liu, T., 2017. Modeling task fMRI data via deep convolutional autoencoder. IEEE Trans. Med. Imaging 37 (7), 1551–1561.

Hutchison, R.M., Womelsdorf, T., Allen, E.A., Bandettini, P.A., Calhoun, V.D., Corbetta, M., Della Penna, S., Duyn, J.H., Glover, G.H., Gonzalez-Castillo, J., et al., 2013. Dynamic functional connectivity: promise, issues, and interpretations. NeuroImage 80, 360–378.

Iraji, A., Benson, R.R., Welch, R.D., O'Neil, B.J., Woodard, J.L., Ayaz, S.I., Kulek, A., Mika, V., Medado, P., Soltanian-Zadeh, H., et al., 2015. Resting state functional connectivity in mild traumatic brain injury at the acute stage: independent component and seed-based analyses. J. Neurotrauma 32 (14), 1031–1045.

Ismail, A.A., Gunady, M., Pessoa, L., Corrada Bravo, H., Feizi, S., 2019. Input-cell attention reduces vanishing saliency of recurrent neural networks. Adv. Neural Inf. Process. Syst. 32.

Kam, T.-E., Zhang, H., Jiao, Z., Shen, D., 2019. Deep learning of static and dynamic brain functional networks for early MCI detection. IEEE Trans. Med. Imaging 39 (2), 478–487.

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

Kerg, G., Kanuparthi, B., Alias Parth Goyal, A.G., Goyette, K., Bengio, Y., Lajoie, G., 2020. Untangling tradeoffs between recurrence and self-attention in artificial neural networks. Adv. Neural Inf. Process. Syst. 33, 19443–19454.

Khazaee, A., Ebrahimzadeh, A., Babajani-Feremi, A., 2016. Application of advanced machine learning methods on resting-state fMRI network for identification of mild cognitive impairment and Alzheimer's disease. Brain Imag. Behav. 10 (3), 799–817.

Kim, B.-H., Ye, J.C., Kim, J.-J., 2021. Learning dynamic graph representation of brain connectome with spatio-temporal attention. Adv. Neural Inf. Process. Syst. 34.

Kong, R., Li, J., Orban, C., Sabuncu, M.R., Liu, H., Schaefer, A., Sun, N., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., et al., 2019. Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. Cerebral Cortex 29 (6), 2533–2551.

Korkmaz, Y., Dar, S.U., Yurt, M., Özbey, M., Cukur, T., 2022. Unsupervised MRI reconstruction via zero-shot learned adversarial transformers. IEEE Trans. Med. Imaging 41 (7), 1747–1763.

Koyamada, S., Shikauchi, Y., Nakae, K., Koyama, M., Ishii, S., 2015. Deep learning of fMRI big data: a novel approach to subject-transfer decoding. arXiv:1502.00093.

Kubicki, M., McCarley, R.W., Nestor, P.G., Huh, T., Kikinis, R., Shenton, M.E., Wible, C.G., 2003. An fMRI study of semantic processing in men with schizophrenia. NeuroImage 20 (4), 1923–1933.

LaConte, S.M., 2011. Decoding fMRI brain states in real-time. NeuroImage 56 (2), 440–454.

Lahaye, P.-J., Poline, J.-B., Flandin, G., Dodel, S., Garnero, L., 2003. Functional connectivity: studying nonlinear, delayed interactions between BOLD signals. NeuroImage 20 (2), 962–974.

Lei, B., Yu, S., Zhao, X., Frangi, A.F., Tan, E.-L., Elazab, A., Wang, T., Wang, S., 2021. Diagnosis of early Alzheimer's disease based on dynamic high order networks. Brain Imag Behav. 15 (1), 276–287.

Li, X., Dvornek, N.C., Zhou, Y., Zhuang, J., Ventola, P., Duncan, J.S., 2019. Graph neural network for interpreting task-fMRI biomarkers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 485–493.

Li, K., Guo, L., Nie, J., Li, G., Liu, T., 2009. Review of methods for functional brain connectivity detection using fMRI. Comput. Med. Imaging Graph. 33 (2), 131–139.

Li, W., Lin, X., Chen, X., 2020a. Detecting Alzheimer's disease Based on 4D fMRI: An exploration under deep learning framework. Neurocomputing 388, 280–287.

Li, Y., Liu, J., Tang, Z., Lei, B., 2020b. Deep spatial-temporal feature fusion from adaptive dynamic functional connectivity for MCI identification. IEEE Trans. Med. Imaging 39 (9), 2818–2830.

Li, H., Satterthwaite, T.D., Fan, Y., 2018. Brain age prediction based on resting-state functional connectivity patterns using convolutional neural networks. In: IEEE International Symposium on Biomedical Imaging. pp. 101–104.

Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L.H., Ventola, P., Duncan, J.S., 2021. Braingnn: Interpretable brain graph neural network for fMRI analysis. Med. Image Anal. 74, 102233.

Liégeois, R., Li, J., Kong, R., Orban, C., Van De Ville, D., Ge, T., Sabuncu, M.R., Yeo, B., 2019. Resting brain dynamics at different timescales capture distinct aspects of human behavior. Nature Commun. 10 (1), 1–9.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Malkiel, I., Rosenman, G., Wolf, L., Hendler, T., 2021. Pre-training and Fine-tuning Transformers for fMRI Prediction Tasks. arXiv:2112.05761.

McKeown, M.J., Sejnowski, T.J., 1998. Independent component analysis of fMRI data: examining the assumptions. Hum. Brain Mapp. 6 (5–6), 368–372.

Mehta, S., Lu, X., Wu, W., Weaver, D., Hajishirzi, H., Elmore, J.G., Shapiro, L.G., 2022. End-to-end diagnosis of breast biopsy images with transformers. Med. Image Anal. 79, 102466.

Mensch, A., Mairal, J., Bzdok, D., Thirion, B., Varoquaux, G., 2017. Learning neural representations of human cognition across many fMRI studies. Adv. Neural Inf. Process. Syst. 30.

Meszlényi, R.J., Buza, K., Vidnyánszky, Z., 2017. Resting state fMRI functional connectivity-based classification using a convolutional neural network architecture. Front. Neuroinform. 11, 61.

Mlynarski, P., Delingette, H., Criminisi, A., Ayache, N., 2019. Deep learning with mixed supervision for brain tumor segmentation. J. Med. Imaging 6 (3), 034002.

Müller-Oehring, E.M., Kwon, D., Nagel, B.J., Sullivan, E.V., Chu, W., Rohlfing, T., Prouty, D., Nichols, B.N., Poline, J.-B., Tapert, S.F., et al., 2018. Influences of age, sex, and moderate alcohol drinking on the intrinsic functional architecture of adolescent brains. Cerebral Cortex 28 (3), 1049–1063.

Ngo, G.H., Nguyen, M., Chen, N.F., Sabuncu, M.R., 2022. A transformer-Based neural language model that synthesizes brain activation maps from free-form text queries. Med. Image Anal. 81, 102540.

Nguyen, S., Ng, B., Kaplan, A.D., Ray, P., 2020. Attend and decode: 4d fMRI task state decoding using attention models. In: Machine Learning for Health. PMLR, pp. 267–279.

Nishimoto, S., 2021. Modeling movie-evoked human brain activity using motion-energy and space-time vision transformer features. BioRxiv.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. 21 (19), 1641–1646.

Niu, C., Cohen, A.D., Wen, X., Chen, Z., Lin, P., Liu, X., Menze, B.H., Wiestler, B., Wang, Y., Zhang, M., 2021. Modeling motor task activation from resting-state fMRI using machine learning in individual subjects. Brain Imag. Behav. 15 (1), 122–132.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences 10 (9), 424–430.

Özbey, M., Dar, S.U., Bedel, H.A., Dalmaz, O., Özturk, Ş., Güngör, A., Çukur, T., 2022. Unsupervised medical image translation with adversarial diffusion models. arXiv preprint arXiv:2207.08208.

Papma, J.M., Smits, M., De Groot, M., Mattace Raso, F.U., van der Lugt, A., Vrooman, H.A., Niessen, W.J., Koudstaal, P.J., van Swieten, J.C., van der Veen, F.M., et al., 2017. The effect of hippocampal function, volume and connectivity on posterior cingulate cortex functioning during episodic memory fMRI in mild cognitive impairment. Eur. Radiol. 27 (9), 3716–3724.

Parisot, S., Ktena, S.I., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., Rueckert, D., 2018. Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. Med. Image Anal. 48, 117–130.

Parisot, S., Ktena, S.I., Ferrante, E., Lee, M., Moreno, R.G., Glocker, B., Rueckert, D., 2017. Spectral graph convolutions for population-based disease prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 177–185.

Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., 2011. Statistical parametric mapping: the analysis of functional brain images. Elsevier.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45 (1), S199–S209.

Poldrack, R.A., 2007. Region of interest analysis for fMRI. Soc. Cogn. Affect. Neurosci. 2 (1), 67–70.

Rahman, M., Mahmood, U., Lewis, N., Gazula, H., Fedorov, A., Fu, Z., Calhoun, V.D., Plis, S.M., et al., 2022. Interpreting models interpreting brain dynamics. Sci. Rep. 12 (1), 1–15.

Rajapakse, J.C., Kruggel, F., Maisog, J.M., Yves von Cramon, D., 1998. Modeling hemodynamic response for analysis of functional MRI time-series. Hum. Brain Mapp. 6 (4), 283–300.

Rajpoot, K., Riaz, A., Majeed, W., Rajpoot, N., 2015. Functional connectivity alterations in epilepsy from resting-state functional MRI. PLoS One 10 (8), e0134944.

Ritchie, S.J., Cox, S.R., Shen, X., Lombardo, M.V., Reus, L.M., Alloza, C., Harris, M.A., Alderson, H.L., Hunter, S., Neilson, E., et al., 2018. Sex differences in the adult human brain: evidence from 5216 UK biobank participants. Cerebral Cortex 28 (8), 2959–2975.

Rogers, B.P., Morgan, V.L., Newton, A.T., Gore, J.C., 2007. Assessing functional connectivity in the human brain by fMRI. Magn. Reson. Imag. 25 (10), 1347–1357.

Román, G., Pascual, B., 2012. Contribution of neuroimaging to the diagnosis of Alzheimer's disease and vascular dementia. Arch. Med. Res. 43 (8), 671–676.

Sarraf, S., Tofighi, G., 2016a. Classification of Alzheimer's disease using fMRI data and deep learning convolutional neural networks. arXiv:1603.08631.

Sarraf, S., Tofighi, G., 2016b. Deep learning-based pipeline to recognize Alzheimer's disease using fMRI data. In: Future Technologies Conference. IEEE, pp. 816–820.

Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T., 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cerebral Cortex 28 (9), 3095–3114.

Shahdloo, M., Çelik, E., Çukur, T., 2020. Biased competition in semantic representation during natural visual search. NeuroImage 216, 116383.

Shahdloo, M., Çelik, E., Ürgen, B.A., Gallant, J.L., Çukur, T., 2022. Task-dependent warping of semantic representations during search for visual action categories. J. Neurosci. 42 (35), 6782–6799.

Shen, H., Wang, L., Liu, Y., Hu, D., 2010. Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. NeuroImage 49 (4), 3110–3121.

Simon, O., Kherif, F., Flandin, G., Poline, J.-B., Riviere, D., Mangin, J.-F., Le Bihan, D., Dehaene, S., 2004. Automatized clustering and functional geometry of human parietofrontal networks for language, space, and number. NeuroImage 23 (3), 1192–1202.

Song, X., Chen, N.-k., 2014. A SVM-based quantitative fMRI method for resting-state functional network detection. Magn. Resonan. Imag. 32 (7), 819–831.

Su, L., Wang, L., Shen, H., Feng, G., Hu, D., 2013. Discriminative analysis of non-linear brain connectivity in schizophrenia: an fMRI Study. Front. Hum. Neurosci. 7, 702.

Suk, H.-I., Wee, C.-Y., Lee, S.-W., Shen, D., 2016. State-space model with deep learning for functional dynamics estimation in resting-state fMRI. NeuroImage 129, 292–307.

Svensén, M., Kruggel, F., Benali, H., 2002. ICA of fMRI group study data. NeuroImage 16 (3), 551–563.

Tagliazucchi, E., Balenzuela, P., Fraiman, D., Chialvo, D.R., 2012. Criticality in large-scale brain fMRI dynamics unveiled by a novel point process analysis. Front. Physiol. 3, 15.

Tagliazucchi, E., Balenzuela, P., Fraiman, D., Montoya, P., Chialvo, D.R., 2011. Spontaneous BOLD event triggered averages for estimating functional connectivity at resting state. Neurosci. Lett. 488 (2), 158–163.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15 (1), 273–289.

Uddin, L.Q., Supekar, K., Menon, V., 2010. Typical and atypical development of functional human brain networks: insights from resting-state fMRI. Front. Syst. Neurosci. 21.

Uddin, L.Q., Yeo, B., Spreng, R.N., 2019. Towards a universal taxonomy of macro-scale functional human brain networks. Brain Topograph. 32 (6), 926–942.

Van Dijk, K.R., Hedden, T., Venkataraman, A., Evans, K.C., Lazar, S.W., Buckner, R.L., 2010. Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. J. Neurophysiol. 103 (1), 297–321.

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

VanRullen, R., Reddy, L., 2019. Reconstructing faces from fMRI patterns using deep generative neural networks. Commun. Biol. 2 (1), 1–10.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Venkataraman, A., Van Dijk, K.R., Buckner, R.L., Golland, P., 2009. Exploring functional connectivity in fMRI via clustering. In: IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 441–444.

Wang, Z., Childress, A.R., Wang, J., Detre, J.A., 2007. Support vector machine learning-based fMRI data group analysis. NeuroImage 36 (4), 1139–1151.

Wang, L., Li, K., Hu, X.P., 2021. Graph convolutional network for fMRI analysis based on connectivity neighborhood. Netw. Neurosci. 5 (1), 83–95.

Wang, D., Shi, L., Yeung, D.S., Heng, P.-A., Wong, T.-T., Tsang, E.C., 2005. Support vector clustering for brain activation detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 572–579.

Wang, C., Xiao, Z., Wu, J., 2019. Functional connectivity-based classification of autism and control using SVM-RFECV on rs-fMRI data. Phys. Med. 65, 99–105.

Woolrich, M.W., Ripley, B.D., Brady, M., Smith, S.M., 2001. Temporal autocorrelation in univariate linear modeling of fMRI data. NeuroImage 14 (6), 1370–1386.

Xie, S.-y., Guo, R., Li, N.-f., Wang, G., Zhao, H.-t., 2009. Brain fMRI processing and classification based on combination of PCA and SVM. In: International Joint Conference on Neural Networks. pp. 3384–3389.

Xing, X., Li, Q., Wei, H., Zhang, M., Zhan, Y., Zhou, X.S., Xue, Z., Shi, F., 2019. Dynamic spectral graph convolution networks with assistant task training for early mci diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 639–646.

Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., Gao, J., 2021. Focal self-attention for local-global interactions in vision transformers. arXiv:2107.00641.

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

Zeng, L.-L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., Zhou, Z., Li, Y., Hu, D., 2012. Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. Brain 135 (5), 1498–1507.

Zhang, X., Hu, B., Ma, X., Xu, L., 2015. Resting-state whole-brain functional connectivity networks for MCI classification using L2-regularized logistic regression. IEEE Trans. Nanobiosci. 14 (2), 237–247.

Zhang, Y., Zhang, H., Chen, X., Lee, S.-W., Shen, D., 2017. Hybrid high-order functional connectivity networks using resting-state functional MRI for mild cognitive impairment diagnosis. Sci. Rep. 7 (1), 1–15.

Zhang, J., Zhou, L., Wang, L., Liu, M., Shen, D., 2022. Diffusion kernel attention network for brain disorder classification. IEEE Trans. Med. Imaging.

Zhao, C., Chen, W., Qin, J., Yang, P., Xiang, Z., Frangi, A.F., Chen, M., Fan, S., Yu, W., Chen, X., et al., 2022. IFT-net: Interactive fusion transformer network for quantitative analysis of pediatric echocardiography. Med. Image Anal. 82, 102648.

Zhao, Y., Dong, Q., Zhang, S., Zhang, W., Chen, H., Jiang, X., Guo, L., Hu, X., Han, J., Liu, T., 2017. Automatic recognition of fMRI-derived functional networks using 3-D convolutional neural networks. IEEE Trans. Biomed. Eng. 65 (9), 1975–1984.

Zhao, C., Li, H., Jiao, Z., Du, T., Fan, Y., 2020. A 3d convolutional encapsulated long short-term memory (3dconv-lstm) model for denoising fmri data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 479–488.