BoiT: Fused Window Transformers for fMRI Time Series Analysis

Hasan A. Bedel, Irmak Sivgin, Onat Dalmaz, Salman UH. Dar and Tolga Çukur

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 BoiT, a blood-oxygen-level-dependent transformer model, for analyzing multi-variate fMRI time series. BoiT 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. Here, we present BolT, a blood-oxygen-level-dependent transformer model, for analyzing multi-variate fMRI time series. BoiT 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. 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 BoiT 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.

Index Terms—functional MRI, time series, deep learning, transformer, classification, connectivity, explainability.

I. INTRODUCTION

Functional MRI (fMRI) measures blood-oxygen-level-dependent (BOLD) responses that reflect changes in metabolic demand consequent to neural activity [1], [2]. 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 [3]–[6]. On the one hand, task-based fMRI enables researchers to associate stimulus or task variables with multi-variate responses across the brain [7]–[9]. Regions that are co-activated in the presence of a particular variable are taken to be involved in the cortical representation of that variable [10], and they are considered to be functionally connected [11]. 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 [12]–[15]. 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 [7], [16]–[18]. Many prior studies have linked behavioral traits and prominent neurological diseases with FC features of BOLD responses [19]–[22].

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 [23], [24]. 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 [25]–[28]. Later studies have instead adopted deep learning (DL) given its ability to capture complex patterns in high-dimensional data [29]–[33]. Various successful deep models have been proposed in the literature based on convolutional [34], graph [35], or recurrent architectures [36], [37] that process FC features. Yet, common FC features primarily reflect first-order inter-regional interactions, potentially disregarding higher-order interactions evident in recorded BOLD responses [38], [39]. To more directly access information in fMRI data, several recent studies have instead built classifiers using recurrent models or vanilla transformer models [40]–[42] 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 [43]–[45].

Here we propose a novel transformer architecture that directly operates on BOLD responses, BoiT, for fMRI time-series classification. To capture local representations, BoiT 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).
A logistic regression analysis on the landmark points in the Autism Brain Imaging Data Exchange (ABIDE) dataset [47].

At the end of the cascade, the encoded 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 [46], and disease detection from resting-state fMRI scans in the Autism Brain Imaging Data Exchange (ABIDE) dataset [47].

BoiT 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 BoiT, 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 time series 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 BoiT is publicly available at https://github.com/icon-lab/BoiT.

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 BoiT that evaluates the relevance of individual time points and brain regions to the classification decisions.

II. RELATED WORK

A. Traditional methods

Whole-brain fMRI data carry densely overlaid patterns of multi-variate responses, which can be difficult to isolate via univariate analysis [48], [49]. This has sparked interest in adoption of ML for multi-variate fMRI analysis [50], [51]. Earlier studies in this domain used traditional classifiers such as support vectors machines [52]–[54]. Because high-dimensional data are paired with models of limited complexity, feature selection is key to improving sensitivity in traditional models [55]–[57]. 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 [58]–[60].

B. 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 [61]–[63] and then to classify external variables [64]–[66]. More commonly, classification models have been built based on FC features among brain ROIs for improved performance [34], [67], [68]. 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 [69], [70]. To capture temporal variability in dynamic FC features, recurrent or transformer architectures have also been integrated to process the GNN outputs [71]. However, GNN-based models might suffer from over smoothing [72] or squashing [73] 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 [74].

C. 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 [40], [75], [76]. Previously reported recurrent architectures in the fMRI literature include vanilla long short-term memory (LSTM) models [40], and hybrid convolutional LSTM models [75], [76]. 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 [77].

Vanilla transformers: Transformer architectures based on self-attention mechanisms have recently been introduced to address limitations of recurrent networks [78]. 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 [41], [79]. 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 III-A). Unlike previous methods that receive as input pre-extracted FC features [69], [71], [80]–[82], 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 [41], [79], 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 [83]. 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 [84]. 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, IFTTNet, has been proposed that reduces dimensionality of input data with a convolutional module prior to the transformer [85]. 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 [86]. 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 [42]. 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.

III. 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 \( b = (b(0), ..., b(T-1)) \in \mathbb{R}^{T \times N} \). Latent representations of BOLD tokens are computed by a cascade of transformer blocks in BolT (Figure 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.

A. BolT

BolT embodies a cascade of transformer blocks and a final linear layer to perform classification. Unlike vanilla transformers [78], 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 (Figure 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 [87]. To capture a window-level
latent representation, a $CLS$ token is also employed for use in downstream detection tasks [88]. 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 $i$-th 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:

$$Q_i = f_q(\{CLS_i, b(i\times s), ..., b(i\times s+W-1)\}),$$
$$K_i = f_k(\{CLS_i, b(i\times s-L), ..., b(i\times s+W+L-1)\}),$$
$$V_i = f_v(\{CLS_i, b(i\times s-L), ..., b(i\times s+W+L-1)\}).$$

To leverage information in the temporal ordering of BOLD tokens, we incorporate a relative position bias in attention calculations [83], [89]:

$$Attention(Q_i, K_i, V_i) = \text{Softmax}(\frac{Q_iK_i^T}{\sqrt{d}} + B)V_i,$$  \hspace{1cm} (2)

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:

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

where $m \in \{0, 1, ..., M-1\}$ is the index of the time window among $P$ windows that contain a particular token, $b^{(i)}_p[m-1]$ is the $i$-th input BOLD
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.

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:

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

(4)

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

**B. 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 [90]. 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:

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

(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 $m$-th block for the $i$-th 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 $\nabla A_{mi}$, 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, $\tilde{A}_G[m] \in \mathbb{R}^{(F+T) \times (F+T)}$ 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 $i$-th 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 $i$-th window is then expressed as:

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

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

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

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

$$\bar{A}_G[m] = \bar{A}_G[m] \odot A_{norm}$$

Eq. (6) captures self-attention for $CLS_i$. Eq. (7) captures attention between $CLS_i$ and BOLD tokens within the receptive field of the $i$-th window. Eq. (8) captures attention between base BOLD tokens and the $CLS_i$ token, and Eq. (9) captures attention between base and fringe BOLD tokens. Note that $(a : -b + 1)$ selects between the $(a+1)$-th element from the start and the $b$-th 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 $\odot$ is Hadamard division used to normalize for repeated token occurrence.

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 $m \in [0, 1, ..., M - 1]$:

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

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:

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

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.

### IV. METHODS

#### A. Datasets

Demonstrations were performed on fMRI data from the HCP S1200 [46] and ABIDE I releases [47]. 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 [91]. 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 [91]. 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 [47], [92]. 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 [80].

#### B. 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 (10%) sets.
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 $\epsilon \in (10^{-6}, 10^{-1})$, number of epochs $\ell \in (5, 100)$ and mini-batch size $b \in (1, 100)$. Training was performed via the Adam optimizer. BolT was trained to minimize the following loss: $L = L_{CE} + \lambda \cdot L_{CW}$ 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 [71]. 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 [93] with 400 regions labeled across seven intrinsic connectivity networks, and the AAL atlas [94] with 116 regions.

C. 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 $2 \times 10^{-4}$ 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 kernels was used. Each block used layer normalization and skip connections. A dropout rate of 0.1 was used in both FW-MSA and MLP layers. FW-MSA modules, given a desired window size $W$, stride $s$ and fringe length $L$ were set proportionately as follows:

$$s = W \alpha, \quad L = m (W - s) \beta = m (1 - \alpha) W \beta,$$

where $m \in \{0, 1, ..., M - 1\}$ is the block index, $\alpha \in \mathbb{R}^+$ is the stride coefficient (i.e., proportionality constant), and $\beta \in \mathbb{Z}^+$ 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 $W \in (10, 200)$, $\alpha \in (0, 0.1, 1.0)$, $\beta \in (0, 3)$ was performed. Hyperparameters were selected as $W = 20$, $\alpha = 0.4$, $\beta = 2$.

D. 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 [40], [68], [76]. 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 BAN and IFT-Net incur $O(N^2T)$ complexity that scales quadratically with the input sequence length [41], [85].

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 $O(NTW(1+\beta(\alpha)(1-\alpha)))$ complexity that linearly scales with sequence length. Other efficient transformers also show a similar linear trend [42], [83], [84], [86]. For instance, $L = 0$ (no cross-window attention) and $s = \frac{W}{2}$ result in a linear $O(NTW)$ complexity in SwinT [83], and $L = 0$ and $s = W$ result in $O(NTW)$ complexity in TFF and HATNet [42].

E. 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 [80]. An $\ell_2$ 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 [34]. 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 [69]. 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 [69]. 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 [71]. 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
TABLE I: Performance of BolT variants ablated of essential design elements. Ablated elements were learnable \textit{CLS} tokens (\textit{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 \textit{CLS} token shared across windows, whereas (L) denotes utilization of local \textit{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  | Windowing | Fusion | Cross Attn. | CWR | Acc. (%) | Rec. (%) | Prec. (%) | AUC (%) |
|--------|-----------|--------|-------------|-----|----------|----------|-----------|---------|
| Schaefer | x         | x      | x           | x   | 86.35    | 85.76    | 84.74     | 94.59   |
|         |           |        |             |     | ±3.56    | ±4.94    | ±4.68     | ±1.57   |
|         | ✓         | x      | x           | x   | 89.65    | 88.38    | 89.15     | 96.03   |
|         |           |        |             |     | ±3.36    | ±5.63    | ±4.78     | ±1.23   |
| ✓(G)   | ✓         | x      | x           | x   | 84.99    | 89.78    | 79.95     | 93.63   |
|         |           |        |             |     | ±2.37    | ±4.76    | ±2.68     | ±1.58   |
| ✓(L)   | ✓         | x      | x           | x   | 89.65    | 88.77    | 88.71     | 96.79   |
|         |           |        |             |     | ±1.85    | ±3.27    | ±3.06     | ±0.90   |
| ✓(L)   | ✓         | ✓      | x           | x   | 90.29    | 89.57    | 89.34     | 97.15   |
|         |           |        |             |     | ±1.77    | ±2.54    | ±3.23     | ±1.01   |
| ✓(L)   | ✓         | ✓      | ✓           | x   | 91.03    | 89.97    | 90.42     | 97.09   |
|         |           |        |             |     | ±2.12    | ±2.70    | ±3.30     | ±1.10   |
| ✓(L)   | ✓         | ✓      | ✓           | ✓   | 91.85    | 90.58    | 91.51     | 97.35   |
|         |           |        |             |     | ±3.05    | ±4.97    | ±3.07     | ±1.06   |

| AAL    | x         | x      | x           | x   | 80.01    | 78.20    | 78.22     | 88.74   |
|        |           |        |             |     | ±3.06    | ±5.61    | ±3.88     | ±3.14   |
| ✓      | x         | x      | x           | x   | 83.12    | 80.80    | 82.25     | 90.88   |
|        |           |        |             |     | ±3.58    | ±5.38    | ±5.05     | ±1.91   |
| ✓(G)   | ✓         | x      | x           | x   | 78.74    | 85.00    | 73.02     | 88.41   |
|        |           |        |             |     | ±4.11    | ±5.07    | ±4.41     | ±3.41   |
| ✓(L)   | ✓         | x      | x           | x   | 86.85    | 87.20    | 84.78     | 93.57   |
|        |           |        |             |     | ±3.06    | ±4.30    | ±4.65     | ±2.25   |
| ✓(L)   | ✓         | ✓      | x           | x   | 87.04    | 87.60    | 84.73     | 93.80   |
|        |           |        |             |     | ±2.48    | ±4.96    | ±3.43     | ±2.13   |
| ✓(L)   | ✓         | ✓      | ✓           | x   | 87.13    | 85.80    | 86.17     | 94.12   |
|        |           |        |             |     | ±3.03    | ±5.39    | ±4.22     | ±1.98   |
| ✓(L)   | ✓         | ✓      | ✓           | ✓   | 87.31    | 86.99    | 85.65     | 94.29   |
|        |           |        |             |     | ±2.69    | ±4.49    | ±4.01     | ±2.05   |

TABLE II: 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        | \textbf{91.85±3.05} | \textbf{90.58±4.97} | \textbf{91.51±3.07} | 97.35±1.05 |
|         | W=80        | 91.48±2.17 | 90.18±4.41 | 91.18±2.86 | \textbf{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        | \textbf{87.31±2.69} | \textbf{86.99±4.49} | \textbf{85.65±4.01} | \textbf{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 |

latent feature was linearly projected to class logits. A window size of 50 was prescribed, FC features were computed via
Pearson’s correlation [71]. 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 [40]. 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 [76]. 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 [68]. 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 sec, respectively [68]. Hidden states were averaged across time samples to perform classification via an MLP followed by a sigmoid activation layer [68]. 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 [83]. For fair comparison, the number and hidden dimensionality of transformer blocks, number of attention heads, and window size were matched with BoiT. 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 [84]. Cross-validated hyperparameters were $2 \times 10^{-4}$ learning rate, 30 epochs, and 32 batch size.

**Longformer**: A transformer model with a windowed attention mechanism proposed for language tasks was considered [84]. 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 BoiT. 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 [84]. Cross-validated hyperparameters were $2 \times 10^{-4}$ learning rate, 40 epochs, and 32 batch size.

**BAnD**: A hybrid CNN-transformer model operating on BOLD responses was considered [41]. The CNN module suffered from suboptimal learning, so for fair comparison BAnD was implemented with ROI-level inputs as in BoiT. 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 [42]. The CNN module was observed to suffer from suboptimal learning, so for fair comparison TFF was implemented with ROI-level inputs as in BoiT. 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 [42]. 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 [85]. 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 [86]. 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 [86]. Cross-validated hyperparameters were $10^{-4}$ learning rate, 100 epochs, and 32 batch size.

### Table III: 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. (%)       | Rec. (%)    | Prec. (%)  | AUC (%)       |
|                     | 77.74±6.82    | 71.93±6.43 | 77.98±8.47 | 83.70±5.64   |
| **HCP-Task**        | Acc. (%)       | Rec. (%)    | Prec. (%)  | AUC (%)       |
|                     | 94.83±1.44    | 94.81±1.44 | 94.89±1.43 | 99.22±0.33   |
| **ABIDE-I**         | Acc. (%)       | Rec. (%)    | Prec. (%)  | AUC (%)       |
|                     | 60.44±3.55    | 46.81±4.80 | 59.71±4.84 | 62.26±4.65   |

### V. Results

#### A. Ablation studies

We conducted a series of ablation studies to evaluate the contribution of individual design elements in BoiT. 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 BoIT. 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 BoIT) 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., BoIT) was employed where regularization was achieved via adding the loss term in Eq. 4.

Table I 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 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 BoiT 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 BoiT 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 II lists performance metrics for BoiT 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 BoiT 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, 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 BoiT. Logistic-regression models were then built for the same detection task given as input a subset of five consecutive tokens [95], [96]. Table III lists detection performance based on the most important subset of 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.

B. Comparative demonstration of BoiT

We demonstrated BoiT 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. BoiT 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 IV, and those for AAL atlas are listed in Table V. For each detection task and based on each atlas, BoiT 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, BoiT 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. Finally, in disease detection, BoiT achieves improvements of (5.11, 10.56, -1, 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 BoiT enables significant performance benefits in detection tasks over prior traditional and DL methods.

In general, we observe that DL models yield superior performance than 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...
BEDEL et al.: BOLT: FUSED WINDOW TRANSFORMERS FOR FMRI TIME SERIES ANALYSIS

TABLE V: 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.

| Method     | HCP-Rest (AAL) | HCP-Task (AAL) | ABIDE-I (AAL) |
|------------|----------------|----------------|---------------|
|            | Acc.(%) Rec.(%) Prec.(%) | Acc.(%) Rec.(%) Prec.(%) | Acc.(%) Rec.(%) Prec.(%) |
| SVM        | 71.71 ± 3.76 65.40 ± 11.86 | 88.11 ± 1.46 88.14 ± 1.46 | 65.60 ± 3.14 55.00 ± 7.40 |
| BrainNetCNN | 71.16 ± 4.91 67.40 ± 9.21 | 90.56 ± 1.06 90.58 ± 1.07 | 64.01 ± 6.04 60.24 ± 9.33 |
| BrainGNN   | 69.79 ± 4.59 72.40 ± 9.70 | 79.73 ± 5.92 79.85 ± 5.88 | 61.40 ± 6.40 58.30 ± 9.33 |
| STAGIN     | 76.18 ± 3.10 72.20 ± 9.22 | 98.87 ± 0.60 98.86 ± 0.61 | 61.52 ± 3.49 52.69 ± 7.38 |
| LSTM       | 73.25 ± 4.48 73.10 ± 8.44 | 96.96 ± 0.69 96.97 ± 0.69 | 63.06 ± 3.96 45.81 ± 22.76 |
| CNN-LSTM   | 74.81 ± 3.15 66.40 ± 9.20 | 97.77 ± 0.60 97.76 ± 0.61 | 63.65 ± 5.42 49.47 ± 17.04 |
| GC-LSTM    | 77.27 ± 5.29 77.79 ± 14.57 | 93.14 ± 1.39 93.14 ± 1.39 | 60.05 ± 5.36 48.27 ± 19.80 |
| SwinT      | 78.37 ± 4.00 76.60 ± 6.69 | 99.23 ± 0.38 99.23 ± 0.38 | 66.78 ± 4.25 60.48 ± 6.38 |
| Longformer | 76.28 ± 4.30 63.00 ± 13.94 | 99.11 ± 0.42 99.11 ± 0.43 | 64.77 ± 4.71 58.69 ± 18.18 |
| BaND       | 78.55 ± 4.12 70.40 ± 8.38 | 98.16 ± 0.45 98.16 ± 0.45 | 63.12 ± 3.60 45.18 ± 12.53 |
| TFF        | 82.57 ± 4.05 81.80 ± 8.59 | 97.43 ± 1.00 97.43 ± 1.00 | 65.84 ± 3.87 46.30 ± 15.83 |
| IFF-Net    | 77.72 ± 4.85 76.00 ± 11.76 | 95.22 ± 5.58 95.17 ± 5.67 | 58.26 ± 4.72 32.11 ± 24.85 |
| HATNet     | 78.37 ± 2.15 74.60 ± 5.58 | 97.97 ± 0.35 97.96 ± 0.35 | 61.61 ± 5.64 54.59 ± 7.15 |
| BOLT       | 87.31 ± 2.69 86.99 ± 4.49 | 99.52 ± 0.39 99.52 ± 0.40 | 68.14 ± 2.81 60.49 ± 4.22 |

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 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.

C. 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 [96].

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 [97]. 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 [98], [99]. 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 [71]. 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 [100]. In contrast, ASD patients manifest important ROIs across DMN, with commonly reported over-activation in ASD [80], [101], [102]. Taken together, these results indicate that BoiT effectively captures task-relevant patterns of brain activation in both normal and disease states.

VI. 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 continuous stimulus or task features [103]. To analyze cortical representations of such continuous features, BoiT can be adapted to instead build regression models [9], [104], [105]. To do this, latent representations of BOLD tokens in downstream layers of BoiT can be coupled with linear or ReLU activation functions. This can enable BoiT 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 [106], [107]. Following this framework, here we used BoiT to build decoding models that detect external features based on measured BOLD responses. An alternative
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.

framework rests on analysis of brain function by building encoding models that predict brain activations given as input stimulus/task features [108]–[112]. 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 [113], [114]. 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 [115]. Moreover, some neurological diseases such as Alzheimer’s or dementia have complementary imaging signatures in other modalities such as structural or diffusion-weighted MRI [116]. 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 [117]. 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 [110], [118]. Alternatively, a CNN model can also incorporated in BolT to perform spatial encoding of volumetric MRI data prior to processing with the transformer blocks [41], [42].

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 [88], [119]. 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 [120], [121]. Reliable augmentation via image synthesis based on advanced procedures such as diffusion modeling can also help alleviate data scarcity [122], [123]. Alternatively, complexity of self-attention modules can be mitigated by replacing regular dot-product attention operators with efficient kernelized operators [79]. Federated learning across multiple institutions might facilitate learning from large, diverse datasets without introducing privacy risks [124], [125]. Lastly, unsupervised learning strategies can also be adopted to permit training on partially labeled fMRI datasets from a larger subject cohort [42], [126]. A systematic exploration of the data efficiency of BolT against competing models remains an important topic for future research.

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.
In this study, we introduced a novel transformer model to improve classification performance on fMRI time series. BoiT 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. BoiT may help detect other neurological disorders with characteristic influences on fMRI activation patterns, and classification of more intricate task variables during cognitive processing.

**REFERENCES**

[1] E. M. Hillman, “Coupling mechanism and significance of the BOLD signal: a status report,” Annual Review of Neuroscience, vol. 37, pp. 161–181, 2014.

[2] J. C. Rajajakse, F. Kruggel, J. M. Masiog, and D. Yves von Cranon, “Modeling hemodynamic response for analysis of functional MRI time-series,” Human Brain Mapping, vol. 6, no. 4, pp. 283–300, 1998.

[3] M. Kubicki, R. W. McCarley, P. G. Nestor, T. Huh, R. Kikinis, M. E. Sontgen, and C. G. Wible, “An fMRI study of semantic processing in men with schizophrenia,” NeuroImage, vol. 20, no. 4, pp. 1923–1933, 2003.

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

[5] J. M. Papma, M. Smits, M. De Groot, F. U. Mattace Raso, A. van der Lught, H. A. Vrooman, W. J. Niessen, P. J. Koudstaal, J. C. van Swieten, F. M. van der Veen et al., “The effect of hippocampal function, volume and connectivity on posterior cingulate cortex functioning during episodic memory fMRI in mild cognitive impairment,” European Radiology, vol. 27, no. 9, pp. 3716–3724, 2017.

[6] A. Mensch, J. Mairal, D. Bzdok, B. Thirion, and G. Varoquaux, “Learning neural representations of human cognition across many fMRI studies,” Advances in Neural Information Processing Systems, vol. 30, 2017.

[7] K. Li, L. Guo, J. Nie, G. Li, and T. Liu, “Review of methods for functional brain connectivity detection using fMRI,” Computerized Medical Imaging and Graphics, vol. 33, no. 2, pp. 131–139, 2009.

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

[9] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, “Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies,” Current Biology, vol. 21, no. 19, pp. 1641–1646, 2011.

[10] O. Simon, F. Kerher, G. Flandin, J.-B. Poline, D. Riviere, J-F. Mangin, D. Le Bihan, and S. Dehaene, “Automated clustering and functional geometry of human parietofrontal networks for language, space, and number,” NeuroImage, vol. 23, no. 3, pp. 1192–1202, 2004.

[11] B. P. Rogers, V. L. Morgan, A. T. Newton, and J. C. Gore, “Assessing functional connectivity in the human brain by fMRI,” Magnetic Resonance Imaging, vol. 25, no. 10, pp. 1347–1357, 2007.

[12] C. Niu, A. D. Cohen, X. Wen, Z. Chen, P. Lin, X. Liu, B. H. Menze, B. Bistler, Y. Wang, and M. Zhang, “Modeling motor task activation from resting-state fMRI using machine learning in individual subjects,” Brain Imaging and Behavior, vol. 15, no. 1, pp. 122–132, 2021.

[13] B. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sahuenco, D. Lashkari, M. Holmanhead, J. L. Rollman, J. W. Smoller, L. Zöllei, J. R. Polimeni et al., “The organization of the human cerebral cortex estimated by intrinsic functional connectivity,” Journal of Neurophysiology, 2011.
application to autism spectrum disorder and Alzheimer’s disease,” *Medical Image Analysis*, vol. 48, pp. 117–130, 2018.

[36] L. Fan, J. Su, J. Qin, D. Hu, and H. Shen, “A deep network model on dynamic functional connectivity with applications to gender classification and intelligence prediction,” *Frontiers in Neuroscience*, vol. 14, p. 881, 2020.

[37] L. Wang, K. Li, and X. P. Hu, “Graph convolutional network for fMRI analysis based on connectivity neighborhood,” *Network Neuroscience*, vol. 5, no. 1, pp. 83–95, 2021.

[38] P.-J. Lahaye, J.-B. Poline, G. Flandin, S. Dodel, and L. Garnero, “Functional connectivity: studying nonlinear, delayed interactions between BOLD signals,” *NeuroImage*, vol. 20, no. 2, pp. 962–974, 2003.

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

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

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

[42] I. Malikel, G. Rosenman, L. Wolf, and T. Hendler, “Pre-training and Fine-tuning Transformers for fMRI Prediction Tasks,” *arXiv:2112.05761*, 2021.

[43] A. A. Ismail, M. Gunady, L. Pessoa, H. Corrada Bravo, and S. Feizi, “Input-cell attention reduces vanishing saliency of recurrent neural networks,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[44] R. Liégeois, J. Li, R. Kong, C. Orban, D. Van De Ville, T. Ge, M. R. Sabuncu, and B. Yeo, “Resting brain dynamics at different timescales capture distinct aspects of human behavior,” *Nature Communications*, vol. 10, no. 1, pp. 1–9, 2019.

[45] E. A. Allen, E. Damaraju, S. M. Pliis, E. B. Erhardt, T. Eichele, and V. D. Calhoun, “Tracking whole-brain connectivity dynamics in the resting state,” *Cerebral cortex*, vol. 24, no. 3, pp. 663–676, 2014.

[46] D. V. Van Essen, D. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium et al., “The WU-Minn human connectome project: an overview,” *NeuroImage*, vol. 80, pp. 62–79, 2013.

[47] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alcaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dupretto et al., “The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism,” *Molecular psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

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

[49] M. W. Woolrich, B. D. Ripley, M. Brady, and S. M. Smith, “Temporal autoregression in univariate linear modeling of fMRI data,” *NeuroImage*, vol. 14, no. 6, pp. 1370–1386, 2001.

[50] K. A. Norman, S. M. Polyn, G. J. D. J. D. Greys, and J. V. Haxby, “Beyond mind-reading: multi-voxel pattern analysis of fMRI data,” *Trends in Cognitive Sciences*, vol. 10, no. 9, pp. 424–430, 2006.

[51] J. V. Haxby, “Multivariate pattern analysis of fMRI: the early beginnings,” *NeuroImage*, vol. 62, no. 2, pp. 852–855, 2012.

[52] X. Song and N.-k. Chen, “A SVM-based quantitative fMRI method for resting-state functional network detection,” *Magnetic Resonance Imaging*, vol. 32, no. 7, pp. 819–831, 2014.

[53] Z. Wang, A. R. Childress, J. Wang, and J. A. Detre, “Support vector machine learning-based fMRI data group analysis,” *NeuroImage*, vol. 36, no. 4, pp. 1139–1151, 2007.

[54] S. H. Hojjati, A. Ebrahimizadeh, A. Khazaee, A. Babajani-Feremi, A. D. N. Initiative et al., “Predicting convergence from MCI to AD using resting-state fMRI: An exploration under deep learning framework,” *arXiv:2006.05205*, 2020.

[55] L. Su, L. Wang, H. Shen, G. Feng, and D. Hu, “Discriminative analysis of non-linear brain connectivity in schizophrenia: an fMRI Study,” *Frontiers in Human Neuroscience*, vol. 7, p. 702, 2013.

[56] W. Li, X. Lin, and X. Chen, “Detecting Alzheimer’s disease Based on 4D fMRI: An exploration under deep learning framework,” *Neurocomputing*, vol. 388, pp. 280–287, 2020.

[57] C. Zhao, H. Li, Z. Jiao, T. Du, and Y. Fan, “A 3d convolutional unsupervised long short-term memory (3dconv-lstm) model for denoising fmi data,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 479–488.

[58] G. Kerg, B. Kanaparthi, A. G. ALPHA PARTH GOYAL, K. Goyette, Y. Bengio, and G. Lajoie, “Untangling tradeoffs between recurrence and self-attention in artificial neural networks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 19443–19454, 2020.

[59] A. Voswani, N. Shazeer, L. Vaswani, J. Parmar, M. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[79] J. Zhang, L. Zhou, L. Wang, M. Liu, and D. Shen, “Diffusion Kernel Attention Network for Brain Disorder Classification,” IEEE Transactions on Medical Imaging, 2022.

[80] A. Abraham, M. P. Milham, A. Di Martino, R. C. Craddock, D. Maras, B. Thirion, and G. Varoquaux, “Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example,” NeuroImage, vol. 147, pp. 745–754, 2017.

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

[82] G. Gouili, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, E. Adeli, and K. M. Pohl, “Spatio-temporal graph convolution for resting-state fMRI analysis,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020, pp. 528–538.

[83] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10,012–10,022.

[84] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” arXiv preprint arXiv:2004.05150, 2020.

[85] C. Zhao, W. Chen, J. Qin, P. Yang, Z. Xiang, A. F. Frangi, M. Chen, S. J. Ritchie, S. R. Cox, X. Shen, M. V. Lombardo, L. M. Reus, C. Alloza, M. A. Harris, H. L. Alderson, S. Hunter, E. Neilson et al., “Sex differences in the adult human brain: evidence from 5216 UK biobank participants,” Cerebral Cortex, vol. 30, no. 8, pp. 2959–2975, 2020.

[86] M. Filippi, P. Valsasina, P. Misci, A. Falini, G. Comi, and A. M. Rocca, “The organization of intrinsic brain activity differs between genders: A resting-state fMRI study in a large cohort of young healthy subjects,” Human Brain Mapping, vol. 34, no. 6, pp. 1330–1343, 2013.

[87] L. Q. Uddin, B. Yeo, and R. N. Spreng, “Towards a universal taxonomy of macro-scale functional human brain networks,” Brain Topography, vol. 32, no. 6, pp. 926–942, 2019.

[88] R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter, “The brain’s default network: anatomy, function, and relevance to disease,” Annals of the New York Academy of Sciences, vol. 1124, no. 1, pp. 1–38, 2008.

[89] Y.-Y. Chen, M. Ujarevic, J. Neal, S. Greening, H. Yim, and T.-H. Lee, “Excessive functional coupling with less variability between salience and default-mode networks in Autism Spectrum Disorder,” Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 2021.

[90] T. Çukur, S. Nishimoto, A. G. Huth, and J. L. Gallant, “Attention during natural vision wars semantic representation across the human brain,” Nature Neuroscience, vol. 16, no. 6, pp. 763–770, 2013.

[91] R. VanRullen and L. Reddy, “Reconstructing faces from fMRI patterns using deep generative neural networks,” Communications Biology, vol. 2, no. 1, pp. 1–10, 2019.

[92] H. Li, T. D. Satterthwaite, and Y. Fan, “Brain age prediction based on resting-state functional connectivity patterns using convolutional neural networks,” in International Symposium on Biomedical Imaging, 2018, pp. 101–104.

[93] S. M. LaCone, “Decoding fMRI brain states in real-time,” NeuroImage, vol. 82, no. 2, pp. 440–454, 2011.

[94] P. Andersson, J. P. Pluim, J. C. Siero, S. Klein, M. A. Viergever, and N. F. Ramsey, “Real-time decoding of brain responses to visuospatial attention using 7T fMRI,” PloS One, vol. 6, no. 11, pp. e27638, 2011.

[95] S. Nishimoto, “Modeling movie-evoked human brain activity using motion-energy and space-time vision transformer features,” bioRxiv, 2021.

[96] E. Celik, U. Keles, I. Kiremitci, J. L. Gallant, and T. Çukur, “Cortical networks of dynamic scene category representation in the human brain,” Cortex, vol. 143, pp. 127–147, 2021.

[97] M. Shahdloo, E. Celik, B. A. Urjen, J. L. Gallant, and T. Çukur, “Task-Dependent Warping of Semantic Representations During Search for a Name,” in Visual Action Categories,” Journal of Neuroscience, 2022.

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

[99] G. H. Ngo, M. Nguyen, N. F. Chen, and M. R. Sabuncu, “A transformer-based neural network model that synthesizes brain activation maps from free-form text queries,” Medical Image Analysis, vol. 81, p. 102540, 2022.

[100] L. Q. Uddin, K. Supekar, and V. Menon, “Typical and atypical development of functional human brain networks: insights from resting-state fMRI,” Frontiers in Systems Neuroscience, p. 21, 2010.

[101] C. Hohenfeld, C. J. Werner, and K. Reetz, “Resting-state connectivity in neurodegenerative diseases: Is there potential for an imaging biomarker?” NeuroImage: Clinical, vol. 18, pp. 849–870, 2018.

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

[103] G. Ronpian and B. Pascual, “Contribution of neuroimaging to the diagnosis of Alzheimer’s disease and vascular dementia,” Archives of Medical Research, vol. 43, no. 8, pp. 671–676, 2012.

[104] L. R. Buckner, F. Kherif, X. Pennec, G. Malandain, N. Ayache, and J.-B. Poline, “Improved detection sensitivity in functional MRI data using a brain parcelling technique,” in International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, 2002, pp. 467–474.

[105] M. Shahdloo, E. Celik, T. Çukur, “Biased competition in semantic category representation in the human brain,” in Proceedings of the New York Academy of Sciences, vol. 2, no. 1, pp. 13683, 2020.

[106] A. Gungor, B. Askin, D. A. Soydan, E. U. Saritas, C. B. Top, and B. U. Kircilar, “ResViT: Residual vision transformer for multi-modal medical image synthesis,” IEEE Transactions on Medical Imaging, vol. 41, no. 10, pp. 2598–2614, 2022.
[122] S. U. Dar, Ş. Öztürk, Y. Korkmaz, G. Elmas, M. Özbey, A. Güngör, and T. Çukur, “Adaptive diffusion priors for accelerated MRI reconstruction,” arXiv preprint arXiv:2207.05876, 2022.

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

[124] G. Elmas, S. U. Dar, Y. Korkmaz, E. Ceyani, B. Susam, M. Özbey, S. Avestimehr, and T. Çukur, “Federated learning of generative image priors for MRI reconstruction,” IEEE Transactions on Medical Imaging, 2022.

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

[126] Y. Korkmaz, S. U. Dar, M. Yurt, M. Özbey, and T. Çukur, “Unsupervised MRI reconstruction via zero-shot learned adversarial transformers,” IEEE Transactions on Medical Imaging, vol. 41, no. 7, pp. 1747–1763, 2022.