University of Essex

Research Repository

EFS-NET:EEG-fNIRS Multi-Scale Fusion Network Based On Spatial Calibration

Accepted for publication in Cognitive Neurodynamics

Research Repository link: https://repository.essex.ac.uk/43687/

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the published version if you wish to cite this paper.
https://doi.org/10.1007/s11571-026-10529-w

www.essex.ac.uk

University of Essex

EFS-NET:EEG-fNIRS Multi-Scale Fusion Network Based On Spatial Calibration

Wenhao Gu[1†], Ian Daly[2], Xinjie He[1], Shurui Li[1], Andrew Ty Lau[3], Xingyu Wang[1], Andrzej Cichocki[4], Jing Jin[1,5*], Yong Chen[6*]

[1] The Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China

[2] Brain-Computer Interfacing and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom

[3] Shanghai Lansheng Brain Hospital Investment Co., Ltd, Shanghai, China

[4] The Systems Research Institute of Polish Academy of Sciences, 01-447bWarsaw, and Nicolaus Copernicus University(UMK), 87-100 Torun, Poland

[5] School of Mathematics, East China University of Science and Technology, Shanghai 200237, China

[6] the Yangtze Delta Guoshu (Shanghai) Digital Medical Technology Co., Ltd., Shanghai, China

*Corresponding author(s). E-mail(s): jinjingat@gmail.com; cheny@nice-dm.com.cn;

Abstract. Hybrid brain-computer interfaces (hBCIs) integrate multiple neuroimaging modalities and utilize their complementary information to address the inherent limitations of single-modality neural signal decoding. For electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) hybrid BCIs, advanced fusion algorithms are crucial to fully exploit the superior spatial localization capability of fNIRS and the millisecond-level temporal resolution of EEG. This work proposes an end-to-end spatial calibration-based multi-scale EEG-fNIRS fusion network named EFS-Net, which organically integrates EEG and fNIRS signals through a multi-scale spatio-temporal fusion architecture. The network consists of three complementary functional branches: a multi-scale temporal convolution branch for capturing rapidly changing cortical electrophysiological features of EEG, an EEG spatial branch for constructing latency-compensated cortical topographies to adapt to the delayed hemodynamic response of fNIRS, and a spatially calibrated fNIRS spatial branch for dynamically fusing spatial feature maps with EEG counterparts to generate temporally aligned and spatially enhanced neural representations. This three-branch fusion structure constructs abundant spatio-temporal feature embeddings and improves the discriminability of neural features. Evaluated on two public datasets including Word Generation (WG) and Mental Arithmetic (MA) with a rigorous subject-specific leave-one-session-out cross-validation protocol, EFS-Net achieves classification accuracies of 77.71±8.23% and 81.69±9.49% on the WG and MA datasets respectively, which surpasses state-of-the-art unimodal algorithms and traditional fusion models. Visualization results demonstrate that the designed alignment strategy can restore realistic cortical spatial distribution characteristics, providing a feasible solution for personalized neural signal decoding in hybrid BCIs.

Keywords: Electroencephalography (EEG),Functional near-infrared spectroscopy (fNIRS), Hybrid brain-computer interface, Spatio-temporal fusion, Multi-scale fusion network

1 Introduction

In neuro-rehabilitation, brain-computer interfaces (BCIs) have opened new avenues for neural recovery and motor restoration. Electroencephalography (EEG) offers millisecond-level temporal resolution, enabling fine-grained capture of cortical electrophysiological dynamics [1], [2]. By extracting salient features,such as event-related potentials (ERPs) or rhythmic modulations — it decodes motor intention and cognitive states [3], [4], supporting personalized feedback and guidance during rehabilitative training.

Functional near-infrared spectroscopy (fNIRS) maps cortical hemodynamics with high spatial fidelity by exploiting the differential absorption of near-infrared light by oxy- and deoxy-hemoglobin [5]. Its non-invasiveness, robustness to motion artifacts, and suitability for prolonged wear make it ideal for longitudinal monitoring and closed-loop therapeutic paradigms. By quantifying neurovascular coupling and neuroplastic reorganization, fNIRS provides critical biomarkers for rehabilitation protocols [6],[7].

However, both modalities have inherent limitations. EEG suffers from low signal-to-noise ratio and significant non-stationarity [8],while fNIRS—despite its motion artifact robustness—is constrained by the slow hemodynamic response of neurovascular coupling, which typically entails a several-second delay. Thus, a hybrid EEG-fNIRS architecture is a promising strategy to synergistically harness EEG's temporal precision and fNIRS's spatial specificity.

Recent advancements in deep learning have significantly refined the art of feature extraction within single modalities. In the realm of deep learning for EEG decoding, recent architectures have made significant strides in capturing complex spatiotemporal dynamics. Notably, TCANet [9] adopted a hierarchical approach by integrating multi-scale convolutions with temporal convolutions and self-attention, effectively balancing local feature extraction with global dependency modeling. Simultaneously, MSCFormer [10] further pushed the boundaries by synergizing multi-branch CNNs with Transformer encoders, demonstrating that capturing long-range global dependencies is crucial for handling inter-subject variability,showcasing the power of hybrid architectures in single-modality scenarios.

Despite these advancements in unimodal feature refinement, the challenge of multimodal integration remains profound.Multimodal integration in hybrid BCIs is both promising and profoundly challenging [11], [12]. Contemporary EEG-fNIRS fusion paradigms fall into two dominant categories: feature-level fusion and decision-level fusion [13].

At the feature level, integration occurs either in the input layer or within intermediate latent representations. One approach leverages three-dimensional spatial remapping. For instance, due to the non-identical cortical coverage of EEG and fNIRS channels, FGA-Net [14] projects heterogeneous EEG and fNIRS sensor arrays onto a common two-dimensional manifold. Linear interpolation then harmonizes positional correspondences, converting raw channel activations into spatially coherent feature maps. Attention-augmented convolutions are subsequently used to model cross-modal interactions, allowing cortical loci to exchange contextual information across modalities.

An alternative strategy employs feature disentanglement. For example, Ming et al. adopt a dual-decoder topology to explicitly separate EEG- and fNIRS-derived representations. A gradient re-balancing

scheme dynamically equalizes the learning rates of the two pathways, preventing modality collapse [15]. Building on this approach, Xu et al. use a Mamba-based architecture to segregate and classify latent features. Both groups incorporate contrastive loss functions that utilize task-relevant shared representations while preserving modality-specific private encodings, thereby fusing complementary neural signatures [16]. Joint multimodal representations (JMRs) further leverage cross-modal attention to enable bidirectional spatio-temporal interaction, improving decoding accuracy [17].

In decision-level fusion, the optimal single-modality feature extractor is first selected; thereafter, with a loss function and late-feature-interaction strategy, the optimal performance of each modality is preserved for the final classification decision [18]. This approach delays feature interaction until the model's final layer. Shi et al. [19] propose a lightweight fusion framework where a multi-head attention mechanism fuses independent embeddings extracted from each modality, followed by a softmax layer to output classification results; in the same work, they benchmark fNIRS HbO against HbR. Adopting a similar approach, Yuan et al. [20] stack convolutional and concatenation layers in their model to pair-wise mix EEG with each of the three fNIRS contrasts, learning features via multiple loss functions and branches.

In EEG-fNIRS fusion frameworks, a key challenge is the intrinsic latency of the fNIRS hemodynamic response. Neurovascular coupling delays the HbO/HbR peak by 4-8 s after the neural event, with this offset persisting for 6-10 s post-task [21], [22]. This naturally misaligns the fNIRS time axis with EEG, causing temporal mismatch of cross-modal features and limiting early-decoding performance.

Notably, this delay exhibits pronounced participant- and region-specific heterogeneity: Kwak et al.'s segmented analysis shows higher accuracy in the 6-10 s window than in the 1-5 s window [14], and Hosni demonstrates substantial inter-participant variability in fNIRS latency—some participants only reach peak classification during the post-task resting period [23]. A fixed-shift strategy inevitably introduces systematic bias in cross-session validation, one reason multimodal fusion can underperform single-modality approaches.

The proposed EFS-Net fundamentally differs in architectural design from existing EEG-fNIRS fusion paradigms. Addressing the lack of physical interpretability caused by spatial misalignment between EEG and fNIRS in non-projective methods, and the neglect of hemodynamic delays in traditional projective approaches (e.g., FGA-Net), EFS-Net constructs a three-branch specialized extraction architecture tailored to signal characteristics. In contrast to strategies like STA-Net that rely on generic attention mechanisms and lack targeted signal extraction, our model does not merely stack modules but implements differentiated feature mining: utilizing multi-scale convolutions to specifically enhance the high-frequency temporal dynamics of EEG, while performing refined spatial topological feature extraction for fNIRS. Furthermore, leveraging EEG-derived physiological spatial priors guides fNIRS features back to authentic cortical topology, and dynamically computes delay-aware weights via cosine similarity across time lags to adaptively capture subject-specific hemodynamic latency variations within sessions. This innovation, combining dynamic temporal adaptation, physiological spatial constraints, and modality-specific extraction, enables EFS-Net to achieve superior decoding performance and biological interpretability.

Our contributions are three-fold:

1.We propose EFS-Net, a novel dynamic fusion network that explicitly leverages the complementary strengths of EEG (temporal precision) and fNIRS (spatial fidelity) for effective multi-modal fusion.

2.We design a spatio-temporal alignment module to dynamically accommodate the inherent hemodynamic latency of fNIRS and its session-wise fluctuations, thereby mitigating temporal

asynchrony between modalities.

3.We conduct comprehensive ablation studies and visual analyses, providing interpretable insights to validate the design rationale and necessity of each core component.

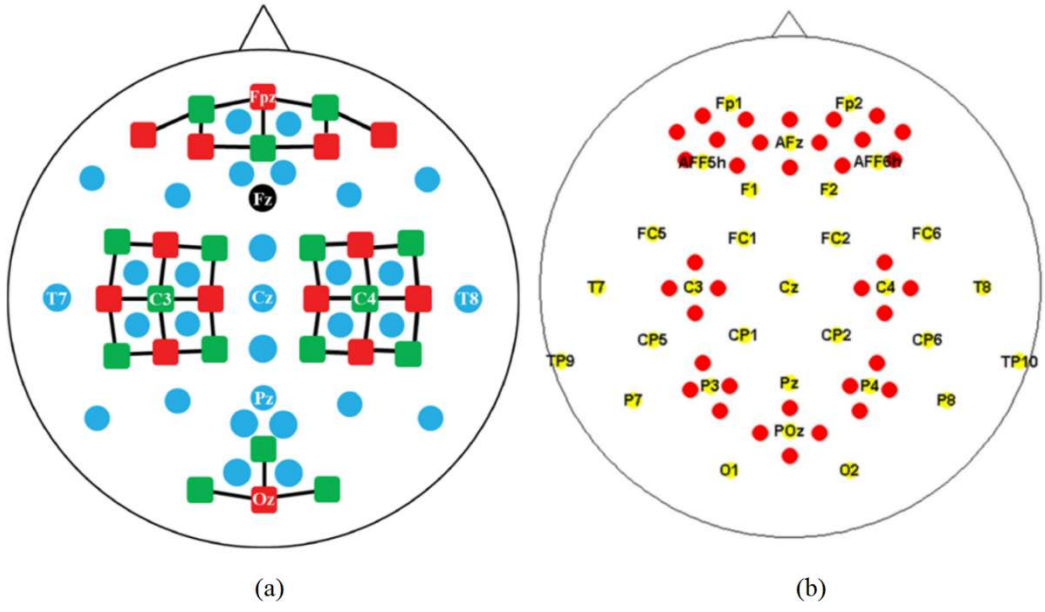
2 Experimental Protocol

2.1 Dataset

The performance of the proposed model was evaluated using two publicly available datasets [24, 25], both of which provide high-quality multimodal foundations by simultaneously acquiring EEG and fNIRS signals.

Dataset A [24] comprises synchronized EEG and fNIRS recordings from 29 subjects (14 males/15 females; 28 right-handed, 1 left-handed; mean age: 28.5 years). EEG signals were recorded via 30 channels at a sampling rate of 1000 Hz and subsequently downsampled to 200 Hz. The fNIRS system consisted of 36 channels (14 sources and 16 detectors), with an original sampling rate of 12.5 Hz downsampled to 10 Hz. Each subject performed three sessions of Mental Arithmetic (MA) tasks, involving continuous "three-digit minus one-digit" subtractions (e.g., 256-7). The Baseline State (BS) served as a period of cognitive rest. Each task was preceded and followed by a 1-minute resting period, with the task repeated 20 times. The spatial topography of EEG electrodes and fNIRS optodes is depicted in Fig. 1.

Dataset B [25] includes synchronized data from 26 subjects (9 males/17 females; mean age: 26.1 years). EEG was recorded through 28 channels (1000 Hz to 200 Hz), while fNIRS utilized 36 channels (10.4 Hz to 10 Hz). The experimental paradigm was Word Generation (WG), where subjects were required to rapidly generate unique words beginning with a specified letter. The Baseline State (BS) involved relaxing while fixating on a central crosshair. The temporal structure of Dataset B is identical to that of Dataset A. The spatial topography of EEG electrodes and fNIRS optodes is depicted in Fig. 1.



(a) (b)

Fig. 1. (a) Layout of EEG electrodes (blue and black circles), fNIRS sources (red squares), and detectors (green squares) in Dataset-A; black solid lines denote fNIRS channels. Source: [24].(b) Layout of EEG electrodes (yellow circles) and fNIRS channels (red circles) in Dataset-B. Source: [25].

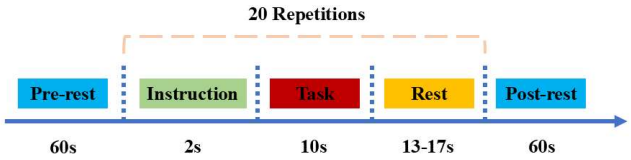


Fig. 2. Timeline of a session of the experiment.

2.2 Pre-processing

EEG Preprocessing. To ensure data quality, identical processing procedures were applied to the EEG data of both datasets. Raw EEG signals were first processed with a zero-phase FIR band-pass filter ranging from 0.5 to 50 Hz to eliminate broadband noise, followed by a 50 Hz notch filter to remove power-line interference. Subsequently, common average reference (CAR) was adopted for global EEG re-referencing. Independent component analysis (ICA) was utilized for artifact removal[26]. Prior to ICA decomposition, the signals were pre-filtered with a 1 Hz high-pass filter to optimize artifact separation performance. A total of 20 independent components were extracted using the default Infomax algorithm. Artifact components were identified via visual inspection. As illustrated in Fig. 3, component ICA-000 exhibited obvious positive activation over the frontal region and presented large sharp deflections in its time-series waveform, which conformed to the typical characteristics of eyeblink artifacts. Since ICA-000 showed such artifact features across most subjects, this component was uniformly discarded during preprocessing, while all remaining components were retained for subsequent analysis.

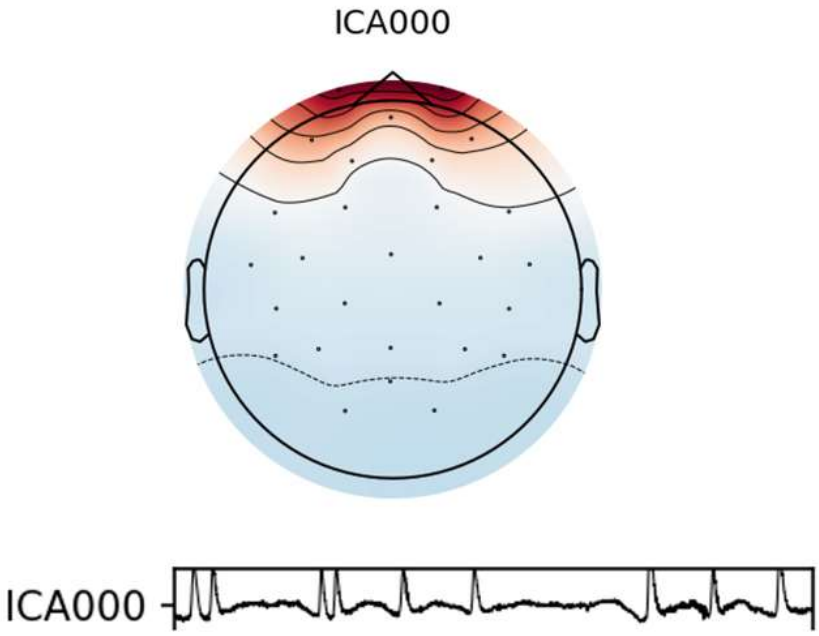


Fig. 3. Topographic map and time series of artifact component ICA-000, showing typical characteristics of ocular artifacts

fNIRS Preprocessing. Raw optical densities were converted into changes in oxyhemoglobin (HbO) and deoxyhemoglobin (HbR) using the modified Beer-Lambert law[**Error! Reference source not found.**]. The raw data of Dataset B were already processed, so this calculation was performed only for Dataset A.The obtained hemodynamic time courses from both datasets were band-pass filtered at 0.01-0.1 Hz to

extract stimulus-evoked responses and suppress physiological drifts and high-frequency instrumental noise.[**Error! Reference source not found.**] Baseline correction was applied by subtracting the average HbO/HbR values measured during the 5 s pre-trial window (−5 s to 0 s).

Data Segmentation. A sliding window approach (window size: 3 s, stride: 1 s) was adopted to balance sample density and temporal resolution. EEG segments were extracted from −2 s to 10 s relative to stimulus onset [13,18,29], which yielded tensors of size 30 × 600 (channels × time) for Dataset A and 28 × 600 (channels × time) for Dataset B. fNIRS data were segmented using the same temporal grid as EEG. However, to account for the hemodynamic delay, each EEG window was aligned with the subsequent 10 s of fNIRS data. This temporal offset structured fNIRS data into tensors of dimensions 11 × 36 × 30 × 2 (window index × channels × time × HbR&HbO) for both datasets.

Spatial Unification. To harmonise spatial granularity, we project both modalities onto a 16 × 16 azimuthal equidistant scalp grid. Cubic-spline interpolation [30] was used to fill inter-electrode voids in the EEG. Remaining vacancies are resolved via nearest-neighbour interpolation, producing tensors of $X_{eeg} \in \square^{16\times16\times600}$.The HbO and HbR values from the fNIRS signals are assigned to all pixels intersecting each source-detector path and processed identically, resulting in tensors of dimensions $X_{fNIRS} \in \square^{16\times16\times30\times2}$.

3 Model

We present a spatio-temporal feature-extraction module that designs a dedicated encoder for each modality to align with its intrinsic physics. This module effectively captures the dominant signatures of both EEG and fNIRS, enabling high-performance hybrid decoding. As illustrated in Fig. 4, EFS-Net comprises three parallel, complementary branches:

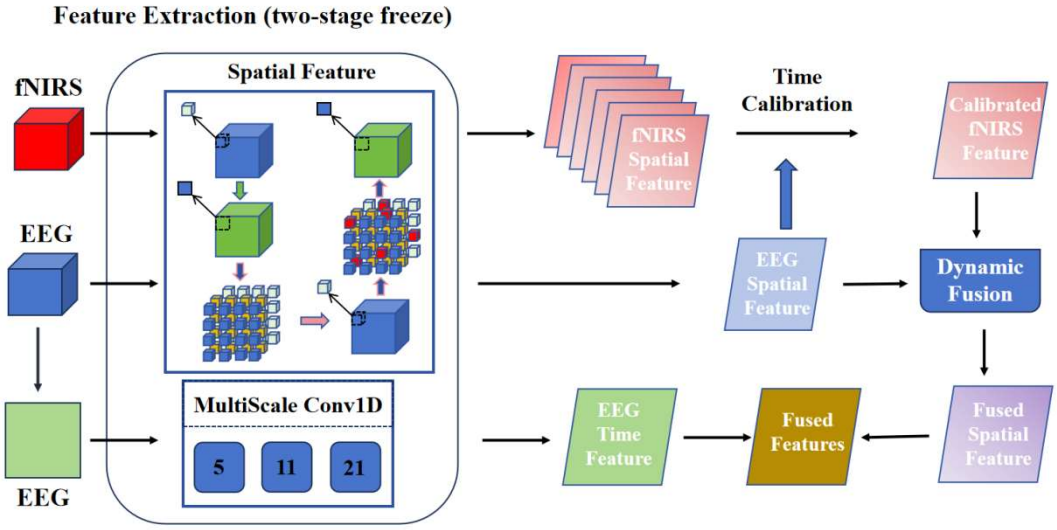


Fig. 4. The framework schematic of our proposed EFS-Net model.

EEG multi-scale temporal branch: Models millisecond-scale transient neural dynamics across

hierarchical time scales to deliver high-resolution temporal features.

EEG spatial branch: Preserves the precise electrode topography and exploits EEG's high temporal fidelity to implicitly calibrate the inherent haemodynamic latency of fNIRS, achieving electrical-to-haemodynamic cross-modal alignment.

fNIRS spatial branch: Targeting the modality's low sampling rate and pronounced latency, it prioritises spatial pattern mining of haemoglobin distributions to compensate for EEG's limited spatial resolution. It dynamically fuses these spatial features with EEG-derived ones to generate a rich, fused spatial representation.

By disentangling and extracting modality-specific features within this novel spatio-temporal framework, EFS-Net preserves the unique physical characteristics of each signal while synergistically integrating and mutually reinforcing complementary cross-modal information.

3.1 MultiScaleEEGEncoder

Multi-scale temporal extraction: To capture the multi-resolution temporal dynamics inherent in neural signals, we employ three parallel 1D convolutional pathways with kernel sizes of 5, 11, and 21 samples, respectively. These pathways are designed to extract features ranging from rapid, millisecond-scale transients to slower, second-scale oscillatory patterns, enabling comprehensive temporal characterization of brain activity. Each pathway adopts a bottleneck architecture, consisting of a 64-dimensional up-projection layer, followed by Layer Normalization, a Gaussian Error Linear Unit (GELU) activation function, and a 32-dimensional down-projection layer. This design effectively reduces computational complexity while preserving and enhancing fine-grained temporal structures. Outputs from the three pathways are concatenated to form a 96-dimensional multi-scale temporal feature descriptor, which encapsulates rich, complementary temporal information for downstream analysis.

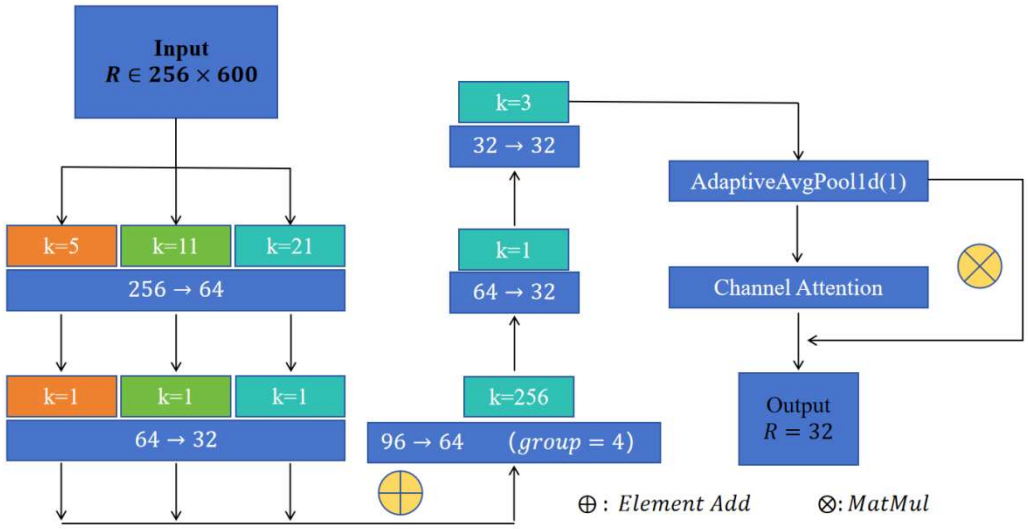


Fig. 5. Multi-scale Temporal Feature Extraction Module

Adaptive enhancement: Global average pooling followed by a two-layer MLP channel-attention block generates a 32-dimensional feature vector that is both temporally discriminative and spatially coherent, providing high-quality EEG embeddings for the downstream cross-modal fusion stage.

3.2 Spatial feature extraction

The spatial feature extraction module consists of multiple cascaded blocks, specifically a spatio-temporal joint convolution block, a spatial feature extraction block, a spatial information aggregation block, a spatio-temporal interaction block, an attention mechanism, and a feature condensation block. Since the spatial lattices of EEG and fNIRS require explicit alignment, the two branches share an identical architectural blueprint; however, their temporal convolution kernels are individually scaled to match the distinct temporal resolutions and sampling rates of the two modalities. Detailed convolution parameters are listed in Table 1.

Table 1. 3-D Convolution Parameter Table

Branch	Layer	Kernel Size	Stride	Channel	Group
EEG	Joint Conv	(3,3,21)	(1,1,5)	8	
	Spatial Conv	(3,3,1)		8	8
	Spatial Fusion	(1,1,1)		16	
	Interact Conv	(3,3,5)	(2,2,4)	16	16
	Feature Refine	(3,3,1)		16	
fNIRS	Joint Conv	(3,3,11)		8	
	Spatial Conv	(3,3,1)		8	8
	Spatial Fusion	(1,1,1)		16	
	Interact Conv	(3,3,5)	(2,2,1)	16	16
	Feature Refine	(3,3,1)		16	

In the first layer of the multimodal spatial-feature extraction model, we instantiate two temporally distinct modelling pathways that leverage the fundamental neurophysiological dichotomy between EEG and fNIRS. To model the millisecond-scale neuro-electric dynamics of EEG, we employ an elongated temporal kernel combined with a supra-unit stride to implement explicit temporal decimation. This operation not only abstracts long-range cortical response envelopes but, more critically, compresses the EEG sequence such that its first-order temporal scale is commensurate with the several-second haemodynamic latency characteristic of fNIRS. Conversely, the fNIRS branch retains its full native temporal range, preserving every slow-evolving sample to faithfully encode the gradual haemodynamic trajectory. The resultant “fast-signal compression versus slow-signal preservation” paradigm ensures temporal-scale coherence prior to any cross-modal interaction.

The second layer introduces spatially separable convolution. First, a grouped convolution partitions the 16 × 16 cortical lattice into local topological clusters, reducing the parameter count from 576 to 72 and suppressing redundant noise. A subsequent point-wise convolution maps 8 channels to 16, enabling free feature mixing across clusters. Finally, a lightweight spatio-temporal interaction layer squeezes redundant activations and further reduces parameters. The resulting joint spatio-temporal tensor preserves modality-specific signal signatures while providing a compact, structurally aligned input to the subsequent dynamic-attention fusion stage, jointly optimizing model compactness and discriminative power.

Dynamic attention: To quantify the importance of each channel, the three-dimensional feature map is squeezed into a channel-wise scalar via adaptive average pooling, as follows:

$$Z(b,c)=\frac{1}{D_1\times D_2\times D_3}\sum_{d_1=1}^{D_1}\sum_{d_2=1}^{D_2}\sum_{d_3=1}^{D_3}X(b,c,d_1,d_2,d_3) \tag{1}$$

Where $X\in\mathbb{R}^{B\times16\times8\times8\times30}$ denotes the input tensor to the attention layer, and $Z\in\mathbb{R}^{B\times16}$ is the squeezed channel-wise feature vector obtained after spatial-temporal aggregation (spatial extents D_1 = D_2 = 8; temporal extent D_3 = 30). The channel weights are subsequently learned dynamically through a two-layer fully-connected network:

$$W_{\text{attn}}(b,c)=\sigma\left(W_2\cdot\text{GELU}\left(W_1\cdot Z(b,:)^T\right)\right) \tag{2}$$

Where $W_1\in\mathbb{R}^{4\times16}$ is the compression matrix (16 → 4 dimensions), $W_2\in\mathbb{R}^{4\times16}$ is the restoration matrix (4 → 16 dimensions), σ(·) notes the Sigmoid function, and GELU indicates the activation function.

Feature modulation:

$$Y(b,c,d_1,d_2,d_3)=X(b,c,d_1,d_2,d_3)\cdot W_{attn}(b,c) \tag{3}$$

Finally, the attention-refined features are fed into a spatial convolution layer to extract location-specific representations.

3.3 Temporal Alignment Module

The EEG and fNIRS features are aligned into a shared space (addressing the heterogeneity of modal feature spaces) using $1\times1\times1$ 3D convolutions (transforming only the channel dimension while preserving spatiotemporal structure).

Cross-Modal Feature Projection. To address the inherent heterogeneity of EEG and fNIRS feature spaces, we align both modalities into a shared latent space using lightweight 1×1×1 3D point convolutions, which transform only the channel dimension while preserving the full spatiotemporal structure of each signal.

Given EEG input features $X_{eeg}\in R^{B\times C\times H\times W\times T}$ (where B = batch size, C = feature channels, 16×16= electrode grid, T = time length), we apply a modality-specific 3D point convolution:

$$X_{EEG}^{P}=Conv3d_{1\times1\times1}(X_{eeg};W_{eeg}) \tag{4}$$

Here, W_{EEG} is a learnable kernel dedicated to EEG (no parameter sharing with fNIRS), which refines channel-wise features for cross-modal compatibility while retaining the original spatiotemporal dimensions ($X_{EEG}^{p}\in R^{B\times C\times H\times W\times T}$).

fNIRS input features incorporate temporal lags to capture delayed hemodynamic responses, denoted as $X_{fNIRS}\in R^{B\times N\times C\times H\times W\times T}$ (where N = number of temporal lags, C = HbO/HbR channels). We apply the same 1×1×1 3D convolution to each lag slice independently, then stack the results along the lag dimension to preserve temporal lag information:

$$X_{fNIRS}^{p}=Stack[Conv3D_{1\times1\times1}(X_{fNIRS,n};W_{fNIRS})]_{n=1}^{N} \tag{5}$$

The kernel W_{fNIRS} is modality-specific (independent of WEEG) to accommodate fNIRS's dual-

signal properties. The output $X^{p}_{fNIRS} \in R^{B\times N\times C\times H\times W\times T}$ maintains the full lag-spatiotemporal structure while optimizing channel features for fusion.

Cross-Modal Similarity Calculation. Given EEG's high temporal resolution, it serves as a temporal reference to quantify the "task relevance" of each fNIRS lag via cosine similarity, thereby mitigating fNIRS's inherent hemodynamic delay.

To isolate channel-level correlations and eliminate spatiotemporal dimension discrepancies, we flatten the spatiotemporal features (excluding the batch/sample dimension) into a single dimension:

$$\tilde{\mathbf{X}}^{p}_{\text{EEG}} = \text{Flatten}(\mathbf{X}^{p}_{\text{EEG}}; \text{dim} = 2), \quad \tilde{\mathbf{X}}^{p}_{\text{fNIRS}} = \text{Flatten}(\mathbf{X}^{p}_{\text{fNIRS}}; \text{dim} = 3) \tag{6}$$

This collapses the H×W×T spatiotemporal structure into a single HWT dimension. We then apply L2-normalization along the "sample-channel" axes to ensure cosine similarity reflects pure angular distance, independent of feature magnitude:

$$\tilde{\mathbf{X}}^{p}_{\text{EEG}} = \text{L2-Norm}(\tilde{\mathbf{X}}^{p}_{\text{EEG}}), \quad \tilde{\mathbf{X}}^{p}_{\text{fNIRS}} = \text{L2-Norm}(\tilde{\mathbf{X}}^{p}_{\text{fNIRS}}) \tag{7}$$

For efficient batch-wise computation, we employ Einstein summation to calculate similarity across all samples and fNIRS lags:

$$\mathbf{S} = \frac{1}{HWT} \cdot \text{Einsum}\left(\tilde{\mathbf{X}}^{p}_{\text{EEG}}, \tilde{\mathbf{X}}^{p}_{\text{fNIRS}}; \ "bcs, bncs \to bn"\right) / \tau \tag{8}$$

Here, $S \in R^{B\times N}$ is the similarity matrix, where $S_{b,n}$ denotes the average cosine similarity between the n-th fNIRS lag and the b-th EEG sample. The $1/HWT$ term normalizes by spatiotemporal length for cross-sample comparability. A temperature parameter τ (default = 0.07) modulates similarity discriminability: smaller τ amplifies differences between high- and low-similarity lags, enhancing weight differentiation.

Calibrated Weight Generation. To generate robust lag-specific weights, we fuse data-driven similarity with learnable prior knowledge, balancing dynamic data patterns and stable initialization.

We introduce a trainable lag prior $\pi \in R^{N}$, initialized as $1/N$ to ensure uniform initial contribution. To integrate prior knowledge with similarity scores, we apply a log transformation to π_n (with 10^{-8} epsilon to avoid numerical underflow) and add it to the similarity matrix:

$$\mathbf{Z} = \mathbf{S} + \log(\pi_n + 10^{-8}) \tag{9}$$

This aligns the scale of the prior (0-1) with similarity scores (real numbers), ensuring the prior guides weight generation without initial bias toward any lag. The resulting Z∈RB×N integrates both data-driven patterns and prior constraints.

Softmax Normalization.We convert Z into probabilistic weights via Softmax along the lag dimension:

$$\mathbf{W} = \text{Softmax}(\mathbf{Z}; \text{dim} = 1) \Rightarrow W_{b,n} = \frac{\exp(Z_{b,n})}{\sum_{n'=1}^{N} \exp(Z_{b,n'})} \tag{10}$$

This enforces $\sum_{n=1}^{N} W_{b,n} = 1$ (probabilistic normalization), where lags with higher EEG similarity and

larger prior π_n receive larger weights, directly determining their contribution to the fused fNIRS feature. This design ensures temporal consistency between EEG and fNIRS, laying the foundation for effective cross-modal fusion.

Module parameters are trained in synergy with a two-stage strategy ("learn features first, then optimize weights"):

Stage 1 (Full Training): Train all parameters (including $\mathbf{W}_{\text{proj}}^{\text{EEG}}, \mathbf{W}_{\text{proj}}^{\text{fNIRS}}, \alpha$) to let projectors learn effective feature mappings and let the lag prior adapt to the data distribution.

Stage 2 (Weight Fine-Tuning): Freeze feature extractors and classifiers, only training module-specific parameters ($\mathbf{W}_{\text{proj}}^{\text{EEG}}, \mathbf{W}_{\text{proj}}^{\text{fNIRS}}, \alpha$). This precisely optimizes lag weight allocation without compromising high-quality spatiotemporal features.

3.4 Feature Fusion

This section presents the unified framework for feature refinement, adaptive modality weighting, and multi-objective optimization, ensuring robust cross-modal integration.

Feature refinement is first implemented via a stacked transformation of linear projection → batch normalization (BN) → GELU activation, which standardizes feature distributions and enhances discriminative capacity.

To adaptively assign contributions from each modality, a spatial dynamic weighting mechanism is introduced. Specifically, a bimodal joint feature vector is constructed as:

$$X_{jo\,\text{int}} = \left[X'_{eeg} \oplus X'_{fNIRS} \right] \in R^{B\times 2D} \tag{11}$$

where X'_{eeg} and X'_{fNIRS} denote the distilled spatial signatures of EEG and fNIRS, respectively. A two-layer multi-layer perceptron (MLP) with Sigmoid activation is then applied to the joint feature to generate a modality-specific weight $w \in [0,1]$ for EEG; the complementary weight for fNIRS is naturally defined as $1-w$:

$$\begin{aligned} Z_1 &= GELU(BN(X_{jo\,\text{int}} W_1 + \mathrm{b}_1)) \\ w &= \sigma(Z_1 W_2 + b_2) \in \square^{B\times 1} \end{aligned} \tag{12}$$

where $\sigma(\cdot)$ denotes the Sigmoid function.

The refined unimodal features are fused according to the learned weights to form the final multimodal representation:

$$X_{fused} = X'_{eeg} \square\ w + X'_{fNIRS} \square\ (1-w) \in R^{B\times D} \tag{13}$$

where $\square$ denotes element-wise multiplication. This linear fusion strategy facilitates gradient propagation while enforcing temporal consistency between EEG and fNIRS.

Regularization and Multi-Objective Optimization. To mitigate excessive reliance on a single modality and promote balanced feature learning, a squared-weight regularization term is introduced:

$$L_{reg} = \lambda \frac{1}{B} \sum_{b=1}^{B} (w_b - 0.5)^2 \tag{14}$$

where λ controls the regularization strength. The overall objective function for the fusion module

is a linear combination of classification loss and regularization loss, enabling multi-objective optimization:

$$L_{total} = L_{cls}(X_{fused}, y) + L_{reg} \tag{15}$$

where $L_{cls}(\cdot)$ represents the cross-entropy classification loss.

Model parameters $\theta = \{W_r, b_r, W_1, b_1, W_2, b_2\}$ are optimized via gradient descent. The gradient flow with respect to the distilled unimodal features is calculated as:

$$\frac{\partial L_{total}}{\partial X'_{eeg}} = \frac{\partial L_{cls}}{\partial X_{fused}} \cdot w, \frac{\partial L_{total}}{\partial X'_{fNRIS}} = \frac{\partial L_{cls}}{\partial X_{fused}} \cdot (1-w) \tag{16}$$

where $w(1-w)$ accounts for the derivative of the Sigmoid function, indicating that gradient contributions are jointly modulated by both the classification loss and the regularization constraint.
The gradient structure ensures that modalities with larger weights receive greater updates, prioritizing the optimization of more reliable signals. For W2, the gradient is:

$$\frac{\partial L_{total}}{\partial W_2} = Z_1^T[\frac{\partial L_{cls}}{\partial X_{fused}} \cdot (X'_{eeg} - X'_{fNIRS}) \cdot w(1-w) + 2\lambda(w-0.5)] \tag{17}$$

This gradient is modulated by both classification and regularization losses, enforcing balanced cross-modal learning.

3.5 Loss Function Composition

Since the model comprises multiple complementary modules, the total loss is formulated as a weighted combination of three components:

$$L_{total} = L_{main} + L_{aux} + L_{reg} \tag{18}$$

where L_{main} denotes the primary classification loss, L_{aux} the auxiliary loss from the EEG branch, and L_{reg} the squared-weight regularization term.

Main-Task Loss. The main-task loss is a weighted label-smoothing cross-entropy, designed to enhance generalization and handle class imbalance. Building upon standard cross-entropy, we incorporate both label smoothing and per-class weighting.

For binary classification with ground-truth label $y \in \{0,1\}$, we construct a soft label distribution with smoothing factor ε:

$$(q(k|\ y) = \begin{cases} 1-\varepsilon & \text{if } k = y \\ \varepsilon & \text{if } k \neq y \end{cases}) \tag{19}$$

Incorporating the class-weight vector $w = [w_0, w_1]$, the primary classification loss is expressed as:

$$L_{main} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{k=0}^{1} w_k \cdot q(k \mid y_b) \cdot \log(p(k \mid f_{fused,b})) \tag{20}$$

where B denotes the batch size, and $p(k \mid f_{fused,b})$ is the primary classifier's predicted probability for the b-th sample.The term $f_{fused,b}$ denotes the tri-branch fused feature.

Auxiliary-Task Loss. The auxiliary loss is a weighted label-smoothing cross-entropy computed on unimodal EEG features, which enforces discriminative EEG representations and facilitates cross-modal alignment:

$$L_{aux} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{k=0}^{1} w_k \cdot q(k \mid y_b) \cdot \log(p(k \mid f_{eeg,b})) \tag{21}$$

Here, $f_{eeg,b}$ denotes the integrated EEG feature (combining temporal and spatial information) for the b-th sample, which acts as a regularization signal to stabilize multimodal training.

4 Experimental

To accommodate sliding-window segmentation and the inherent hemodynamic latency of fNIRS, all evaluations were conducted under a strict subject-specific paradigm. Inter-session robustness was quantified using a leave-one-session-out cross-validation scheme: each of the three recording sessions was iteratively withheld as an independent test set, while the remaining two were pooled for model development.

The training pipeline adopted a dual-stage strategy to balance discriminative capability with generalization stability, implementing fine-grained control over the learning rates:

Discriminative Pre-training: The pooled data were partitioned into training and validation subsets at a 4:1 ratio. The model was trained for a maximum of 200 epochs, using the Adam optimizer with an initial learning rate set to 1×10^{-3} and a weight decay of 3×10^{-4} to constrain model complexity. An early stopping mechanism was introduced—training terminated if the validation loss did not decrease for 50 consecutive epochs—effectively mitigating the risk of overfitting.

Regularized Fine-tuning: In this stage, all feature extraction layers were frozen to preserve the underlying representations learned in the first stage. Only fusion and dynamic weighting parameters were updated using the combined training and validation dataset. The Adam optimizer was adopted, with the learning rate lowered to 1×10^{-4} for subtle parameter updates. This avoids drastic gradient changes that may damage stable feature manifolds obtained in the prior stage. The process stops after 50 epochs or when the training loss falls below the minimum validation loss from the first stage, so as to achieve optimal parameter adjustment of fusion modules.

4.1 Overall performance

The overall performance of EFS-Net was benchmarked against multiple state-of-the-art baseline models, with results reported in Table 2.

Table 2. SOTA Comparison: EEG-fNIRS Fusion Performance on Task

Methods	**Signal**	**WG**			
		Accuracy(%)	**S.D.(%)**	**95%CI(%)**	***p*-value**
Conformer[31]	EEG	74.26	10.44	[70.44,78.48]	0.0038**
MSCFormer[10]	EEG	73.95	10.33	[69.78,78.12]	0.0004**
TCANet[9]	EEG	71.94	10.82	[67.57,76.31]	<0.0001***
fNIRSNet[32]	fNIRS	61.53	9.38	[57.74,65.32]	<0.0001***
FGA-Net[14]	EEG+fNIRS	70.33	9.93	[66.32,74.34]	<0.0001***
DUAL-EEGNet[33]	EEG+fNIRS	73.53	10.34	[69.35,77.71]	0.00012**

DARNet[15]	EEG+fNIRS	73.32	9.37	[69.53,77.11]	0.0002**
IIMCNet[20]	EEG+fNIRS	72.08	9.51	[68.24,75.92]	<0.0001***
STA-Net[13]	EEG+fNIRS	74.58	8.98	[70.95,78.21]	0.018*
Model	EEG+fNIRS	77.71	8.23	[74.39,81.03]	-
		MA			
		Accuracy(%)	**S.D.(%)**	**95%CI(%)**	***p*-value**
Conformer[31]	EEG	78.74	10.60	[74.88,82.60]	0.0023**
MSCFormer[10]	EEG	77.12	9.83	[73.54,80.70]	0.0005**
TCANet[9]	EEG	76.67	10.46	[72.86,80.48]	0.0025**
fNIRSNet[32]	fNIRS	72.11	10.01	[68.30,75.92]	<0.0001***
FGA-Net[14]	EEG+fNIRS	75.58	10.04	[71.93,79.23]	0.0017**
DUAL-EEGNet[33]	EEG+fNIRS	77.17	9.75	[73.62,80.72]	0.0029**
DARNet[15]	EEG+fNIRS	78.21	9.63	[74.71,81.71]	0.0038**
IIMCNet[20]	EEG+fNIRS	77.66	9.92	[74.05,81.27]	0.0027**
STA-Net[13]	EEG+fNIRS	80.69	9.05	[77.40,83.98]	0.2390
Model	EEG+fNIRS	81.69	9.49	[78.24,85.14]	-

Among Experimental Settings and Methodological Rigor. To ensure the impartiality of the comparative analysis and the reproducibility of the experimental results, we established a rigorous evaluation protocol:

Diversity and Representativeness of Baselines: The experimental setup incorporates a diverse array of state-of-the-art models. The baseline library encompasses unimodal paradigms (Conformer, MSCFormer, TCANet, and fNIRSNet), methods utilizing 3D spatial projection (FGA-Net and STA-Net), and recent multimodal fusion paradigms that process raw channel-space data (IIMCNet, DUAL-EEGNet, and DARNet). This comprehensive coverage of both feature-level and decision-level fusion strategies ensures a thorough assessment across varying neural imaging processing paradigms.

Consistency in Training Protocols: To eliminate bias introduced by the experimental environment, all baseline models were trained within the identical dual-stage training framework as EFS-Net. This process strictly unified data partitioning strategies and early stopping criteria, ensuring that all models converged under equivalent conditions.

Rigorous Hyperparameter Configuration: All baseline models adopt identical hyperparameter optimization configurations with unified learning rates and optimizer settings, eliminating performance discrepancies caused by differing hyperparameters. Hence, performance gaps originate solely from architectural distinctions, complying with the single-variable principle for controlled comparison. Specifically, the Adam optimizer is adopted with an initial learning rate of 1×10^{-3} and a weight decay of 3×10^{-4} to regularize model parameters.

Multidimensional Quantitative Evaluation System: Beyond base accuracy metrics, this study integrates standard deviation (S.D.), 95% confidence intervals (95% CI), and p-value statistical analysis to construct a multidimensional evaluation system ranging from mean performance to statistical significance, ensuring the objectivity and robustness of performance improvements.

Experimental Results and Analysis. Table 2 presents the quantitative performance of EFS-Net compared against baseline models on both the WG and MA datasets.

Experimental results demonstrate that EFS-Net achieves superior classification accuracy against all

competing methods across both classification tasks. On the WG dataset, EFS-Net obtains an accuracy of 77.71%, surpassing the second-best STA-Net (74.58%) by 3.13 percentage points; on the MA dataset, its accuracy reaches 81.69%, which verifies its powerful multimodal recognition capability. Further statistical significance tests based on p-values confirm that the performance of EFS-Net is statistically significantly different from that of existing algorithms.

4.2 Ablation Study

Table 3. Single-Branch Performance and Multi-Branch Ablation Results of EFS-Net

Single-Branch	**WG**		**MA**	
	Accuracy(%)	**S.D.(%)**	**Accuracy(%)**	**S.D.(%)**
EEG-time	73.07	8.66	75.15	8.69
fNIRS	64.11	10.58	70.38	10.27
EEG-spatial	62.94	8.17	67.79	8.23
Multi-Branch				
w/o EEG-time	67.68	9.37	75.17	9.73
w/o fNIRS	73.34	9.31	76.02	8.76
w/o EEG-spatial	66.46	11.95	74.53	10.01
Model	77.71	8.23	81.69	9.49

Component-wise Ablation: Branch Analysis.To systematically explore the contribution of each component in the proposed EFS-Net, we conducted comprehensive ablation experiments on the WG and MA datasets. The results of single-branch performance and multi-branch ablation are presented in Table 3.

First, we evaluated the performance of each individual branch to establish baseline results. As listed in Table 3, the EEG temporal branch achieved the highest accuracy among all single modalities on both datasets, reaching 73.07% on WG and 75.15% on MA. This demonstrates that temporal features derived from EEG signals possess the strongest discriminative ability for the given task. By contrast, the fNIRS branch (64.11% on WG, 70.38% on MA) and EEG spatial branch (62.94% on WG, 67.79% on MA) yielded inferior performance when used alone. It indicates that relying solely on a single modality or EEG spatial features cannot fully characterize the complex properties of the task.

Subsequently, multi-branch ablation experiments were carried out to analyze how the removal of each module affects the overall model performance. The experimental results reveal that every branch plays an essential role, and consistent performance trends are observed across the two datasets.

When the EEG temporal branch (w/o EEG-time) was removed, the accuracy dropped sharply from 77.71% to 67.68% on the WG dataset and decreased from 81.69% to 75.17% on the MA dataset. This highlights the dominant contribution of this branch, with a more pronounced impact on the WG dataset.

Eliminating the fNIRS branch (w/o fNIRS) led to a moderate and steady performance decline on both datasets. The accuracy fell from 77.71% to 73.34% on WG and from 81.69% to 76.02% on MA. This verifies that fNIRS features provide complementary physiological information and effectively improve the model robustness.

After removing the EEG spatial branch (w/o EEG-spatial), the model accuracy decreased remarkably, from 77.71% to 66.46% on WG and from 81.69% to 74.53% on MA. This further proves that spatial patterns in EEG signals are critical for decoding tasks, and this branch also acts as a core part to anchor the correlation between EEG and fNIRS.

Notably, the complete multi-branch EFS-Net outperformed all single-branch models and ablated variants on both datasets and achieved the optimal performance. Its accuracy reached 77.71% (standard deviation = 8.23%) on WG and 81.69% (standard deviation = 9.49%) on MA. Compared with the best single-branch baseline (the EEG temporal branch), the full model achieved relative improvements of 4.64% and 6.54% on the WG and MA datasets, respectively, which fully reflects the synergistic effect of multi-modal and multi-view feature fusion.

Table 4. Module Ablation

Branch Ablation	WG		MA	
	Accuracy(%)	**S.D.(%)**	**Accuracy(%)**	**S.D.(%)**
w/o L_{aux}	76.58	8.87	80.91	9.54
w/o $L_{fusion-reg}$	77.06	8.37	81.27	8.48
w/o Calibrated Weight	76.96	8.95	78.36	7.29
w/o Temporal Alignment	75.31	8.45	79.35	8.61
Full Model	77.71	8.23	81.69	9.49

Module Ablation Analysis. The module ablation experiments systematically quantify the marginal contributions of each core component. Removing the temporal alignment module leads to performance drop, with accuracy decreasing to 75.31% on the WG dataset and 79.35% on the MA dataset, respectively. This result quantitatively demonstrates the central role of this module in compensating for the inherent 4-8 second hemodynamic latency of fNIRS signals and resolving cross-modal temporal asynchrony, which is one of the major bottlenecks in multimodal brain-computer interfaces (BCIs).

After removing the calibrated weight mechanism, the accuracy dropped sharply from 81.69% to 78.36% on the MA dataset and to 76.96% on the WG dataset. This indicates that adaptive weight allocation can effectively balance the contributions of different modalities and mitigate the model's over-reliance on a single modality.

Excluding the auxiliary loss (L_{aux}) reduces the accuracy to 76.58% and 80.91% on the two datasets, respectively. It shows that supervising single-branch learning can enhance the discriminability of features, provide essential regularization constraints for multimodal fusion, and improve training stability. Removing the fusion regularization term ($L_{fusion-reg}$) decreases the accuracy to 77.06% and 81.27%, respectively, verifying its function in preventing modal collapse and constructing a robust multimodal feature space.

The full model achieves the optimal performance on both datasets. In conclusion, EFS-Net realizes the effective fusion of high-temporal-resolution EEG and high-spatial-resolution fNIRS signals through the synergistic integration of temporal alignment, adaptive weight allocation, and multi-objective optimization.

Table 5. Classification Accuracy with Different fNIRS Chromophore Combinations

Signal	WG		MA	
	Accuracy(%)	**S.D.(%)**	**Accuracy(%)**	**S.D.(%)**
EEG	73.34	9.31	76.02	8.76
EEG+HbR	77.50	8.44	80.87	8.89
EEG+HbO	77.53	9.04	81.12	9.71
EEG+HbO+HbR	77.71	8.23	81.69	9.49

As shown in Table 5, incorporating both fNIRS chromophores only yields a slight improvement in decoding accuracy compared with using a single chromophore. Given the relatively large standard deviations across groups, we performed paired statistical tests for rigorous evaluation. The results demonstrate no statistically significant differences in accuracy among the EEG+HbR, EEG+HbO and combined EEG+HbO+HbR models. This finding is consistent with the principles of neurovascular physiology: HbO and HbR exhibit antiphasic responses to tasks and deliver comparable decoding performance when used individually. Therefore, the core value of multi-chromophore fusion lies not in boosting accuracy, but in providing structural redundancy to achieve a physiologically complete characterization of neurovascular coupling.

4.3 Feature Visualisation

The fNIRS Temporal Alignment Module is a core component of EFS-Net. To interpret its underlying mechanism, we computed subject- and label-specific temporal attention maps. During testing, frame-level calibration weights were extracted for each subject and averaged across three sessions to obtain stable weight distributions.

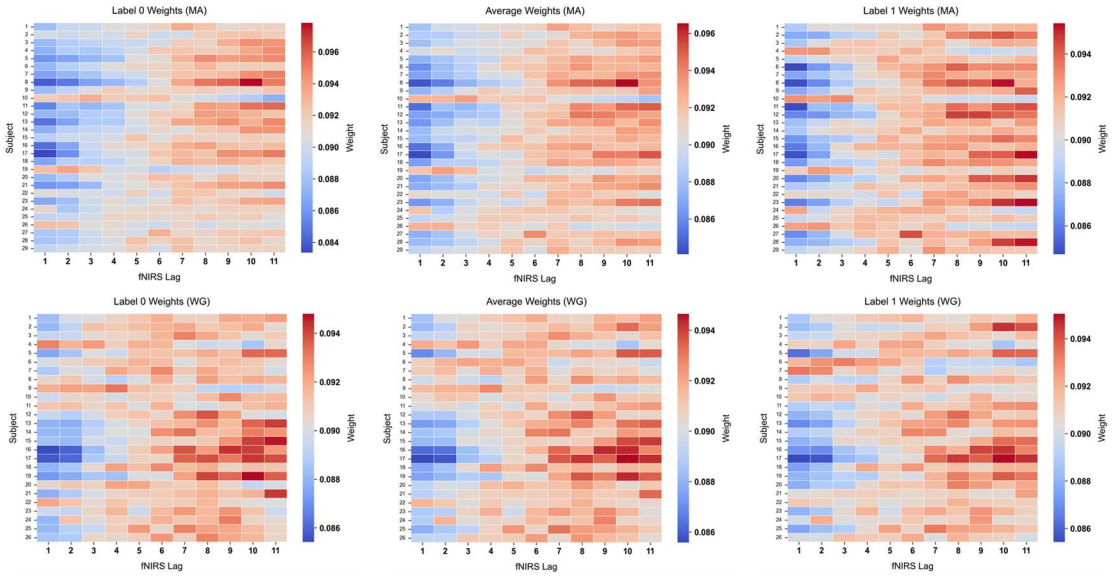


Fig. 6. Temporal attention weight heatmaps. The heatmaps visualize calibration weights assigned by the fNIRS temporal alignment module across subjects (rows) and fNIRS time lags (columns, 1-11 s). Red indicates higher weights, corresponding to more reliable cross-modal correlations. Results are shown for the MA (top) and WG (bottom) datasets, including Label 0 weights (left), cross-session averaged weights (middle), and Label 1 weights (right). The module adaptively models subject-specific and task-dependent hemodynamic response latencies.

Visualization of the heatmaps (Fig. 6) reveals two key observations:

Inter-subject heterogeneity: Each subject shows a unique weight pattern, consistent with the well-known variability in neurovascular latency across individuals.

Intra-subject task dependence: While the overall weight distribution remains similar between baseline and task conditions, the peak latency and spread differ slightly, indicating adaptive weighting rather than fixed time shifts.

During task blocks, the weights rise smoothly and then decay gradually, closely matching the slow hemodynamic response process of HbO/HbR. Contiguous time windows tend to be assigned high or low weights synchronously, confirming that the calibration kernel follows the low-pass characteristics of neurovascular signals rather than overfitting to fast EEG transients.

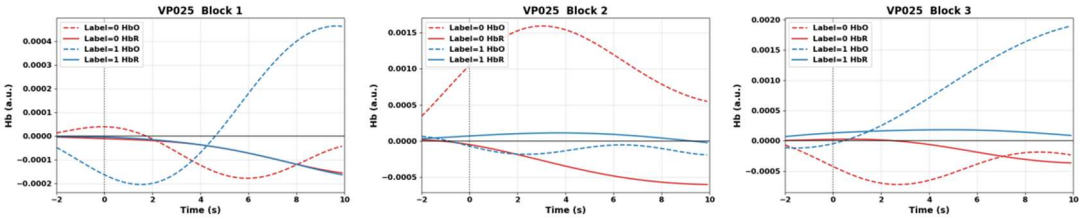


Fig. 7. HbR/HbO curves under task/baseline states across 3 sessions

Session-wise analysis (Fig. 7) further shows that the dominant weight window shifts slightly across runs, reflecting subtle fluctuations in habituation and attention that fixed-latency methods cannot capture.

We select Participant 25 from the WG dataset as a typical case to elaborate the working mechanism of the temporal calibration module. The fNIRS signals from all prefrontal channels are averaged, and baseline-corrected signal variations are calculated for task and resting states. As illustrated in Fig. 7 HbO shows larger amplitude variations than HbR. The two signals become anti-correlated approximately 2 seconds after trial onset, which is consistent with the classic rules of neurovascular coupling.

Table 6. Temporal Attention Weights for Participant 25's fNIRS Signals Across Three Sessions on WG (Values Normalized to 100%)

fNIRS	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}	W_{11}
S1	8.933	8.873	8.953	9.087	9.238	9.020	9.197	9.247	9.107	9.164	9.175
S2	8.518	8.668	8.917	9.117	9.107	8.929	9.483	9.333	9.309	9.348	9.265
S3	8.904	9.066	8.885	9.015	9.382	9.253	9.254	9.112	9.250	9.010	8.866

Table 6 summarizes the temporal attention weights of eleven 3-second windows (W1-W11) obtained from three experimental sessions. All weight values fluctuate around the theoretical average of 9.09%. Combined with the weight distribution in Figure 6, the weights tend to be relatively lower for early windows and slightly higher for later windows, while the peak positions of weights differ across experimental sessions.

This result is consistent with both the study design and physiological characteristics. A highly overlapping sliding time window strategy was adopted, with approximately two-thirds overlap between adjacent segments, resulting in continuous and smooth changes in weights over time. Furthermore, the hemodynamic response itself is a slow and sustained process. Accordingly, the weights do not exhibit abrupt peaks, but rather a gradual increase and progressive concentration in later windows, effectively matching individual hemodynamic delay characteristics.

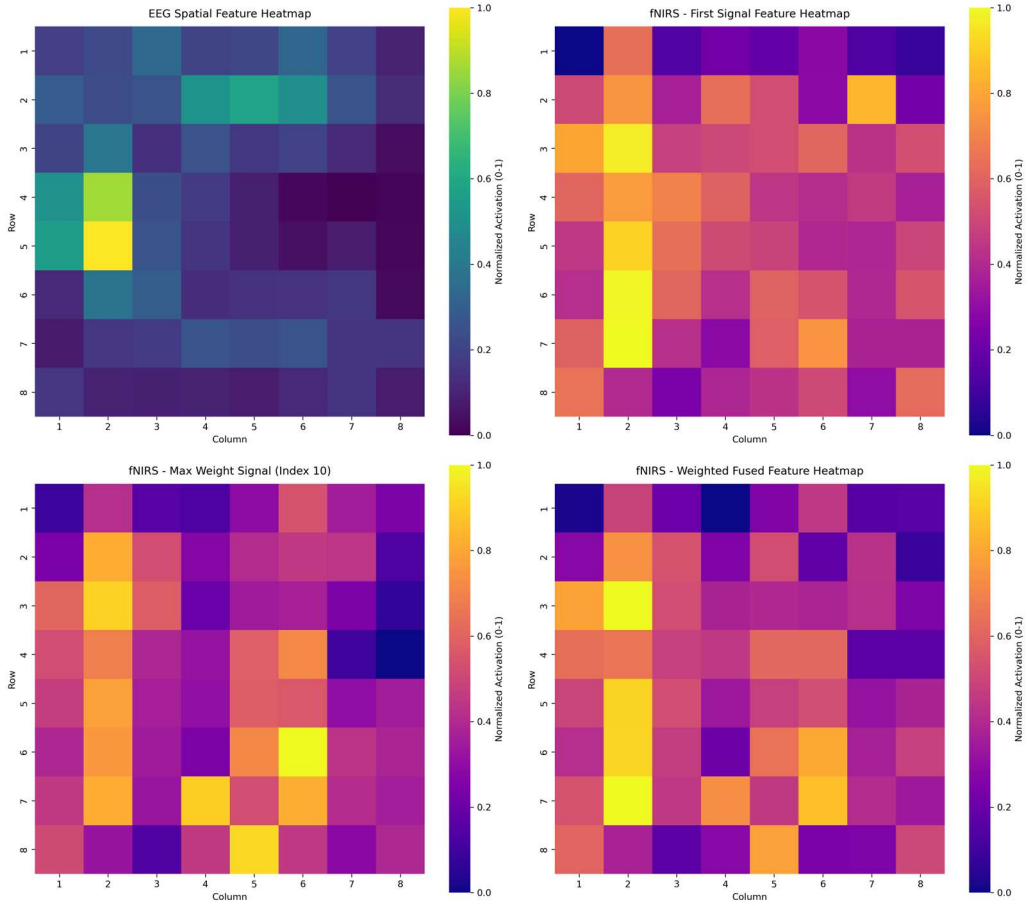


Fig. 8. fNIRS Feature Visualization (Top-left: EEG spatial feature heatmap; Top-right: raw fNIRS feature heatmap; Bottom-left: EEG-guided calibrated fNIRS feature heatmap; Bottom-right: weighted fused fNIRS feature heatmap, demonstrating enhanced spatial consistency between fNIRS and EEG after calibration and fusion).

Fig. 8 visualizes the spatial feature maps of EEG and fNIRS across different processing stages. The top-left EEG spatial heatmap exhibits distinct left-hemisphere dominance, serving as the spatial reference. The top-right raw fNIRS feature map displays scattered activation patterns with limited spatial alignment to EEG. In contrast, the bottom-left map, guided by high-confidence (maximum-weight) fNIRS frames, shows concentrated and focal activations, indicating effective mitigation of neurovascular latency. Finally, the bottom-right weighted fused feature map demonstrates strong spatial coherence with the EEG template, confirming that the calibration module aligns fNIRS spatial features to the underlying neural topography. This visualization validates that the proposed temporal-calibration and fusion strategy successfully resolves spatial misregistration and enforces consistent spatial representation between EEG and fNIRS.

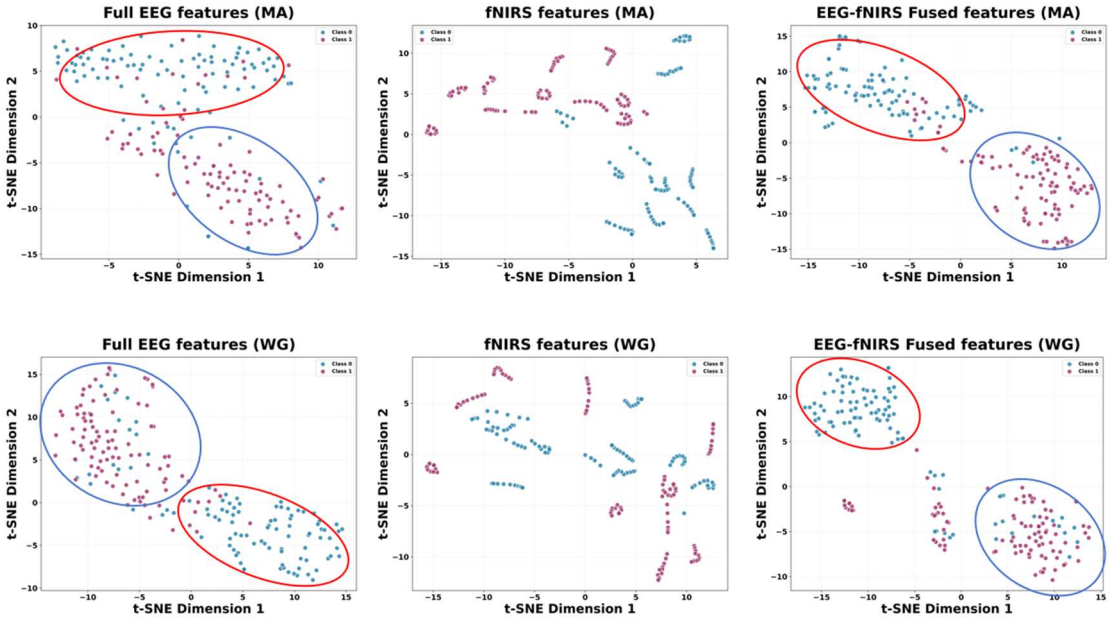


Fig. 9. Comparison before and after applying temporal weighting

Feature Analysis. Taking the first session data of Subject 1 from both datasets as an example, we visualize raw EEG features, raw fNIRS features and their fused representations within the fusion block, as illustrated in Fig. 9 When classification is performed using EEG alone, the centroids of different classes lie close to one another with partial overlapping. By contrast, fNIRS features are widely scattered and show unsatisfactory separability across classes. After feature fusion, samples belonging to the same category cluster much more compactly and barely overlap between different classes, which verifies that fused features greatly improve the discriminative capability for classification tasks.

4.4 Computational Complexity Analysis

Table 7. Computational overhead comparison

Model	Signal	Params (M)	FLOPs (GMac)	Inference time (ms/sample)
STA-Net	EEG+fNIRS	3.33	0.11	28.68
fNIRSNet	fNIRS	0.0005	0.008	0.3
Conformer	EEG	0.58	0.043	4.73
Model	EEG+fNIRS	17.34	0.17	18.82

As shown in Table 7, single-modal models obtain advantages in parameter quantity and inference speed owing to limited input information, yet their classification accuracy is inferior to multi-modal alternatives. The proposed model adopts a three-branch decoupled architecture for separate feature extraction of EEG and fNIRS signals, which inevitably increases total parameters. Besides, the temporal delay calibration module brings extra computational cost, leaving room for lightweight optimization in practical real-time deployment. Compared with the dual-modal STA-Net, our method possesses more parameters and higher theoretical computational complexity while achieving faster inference benefiting from the inherent parallelism of the three-branch design. Future work will adopt lightweight strategies to further reduce model overhead.

5 Conclusion

We propose EFS-Net, an end-to-end EEG-fNIRS multimodal fusion framework to tackle the spatio-temporal mismatch between electrophysiological and hemodynamic signals. Composed of a multi-scale temporal encoder, an EEG-driven spatial calibration branch and a dynamically aligned fNIRS spatial pathway, the model adaptively compensates for subject-specific neurovascular delays and realizes cross-modal spatial consistency across different sessions of each individual subject. Evaluated under the intra-subject training protocol on two public datasets, our method achieves state-of-the-art classification accuracies of 77.71% on the WG dataset and 81.69% on the MA dataset, surpassing existing unimodal and multimodal fusion algorithms. Ablation experiments and visualization results verify the essential contribution of each module; specifically, the temporal alignment module effectively reduces feature mismatches originating from signal latency. Compared with straightforward feature concatenation or late fusion, elaborate spatio-temporal alignment is critical for exploiting the complementary merits of EEG and fNIRS in subject-specific decoding tasks.

6 Limitation and Future Work

Experimental results have verified the effectiveness of EFS-Net for multimodal EEG-fNIRS decoding, yet several limitations remain to be addressed. The temporal calibration module, designed to adapt to subject-specific hemodynamic latencies, introduces extra computational overhead and thus needs further optimization for real-time inference. Besides, the three-branch architecture facilitates full extraction of discriminative multimodal features but leads to a relatively large number of model parameters.

Future work will optimize the proposed algorithm from multiple dimensions. On the one hand, we will restructure the computational logic of the calibration module and employ mainstream lightweight techniques to reduce model size and inference latency while preserving classification accuracy; advanced interaction strategies will also be explored to further boost the model’s real-time performance. On the other hand, we will collect more datasets covering various experimental tasks to promote the practical deployment of the algorithm on portable wearable devices.

Declaration of Interest Statement

The authors declare no competing interests.

Acknowledgements

This work was supported by Brain Science and Brain-like Intelligence Technology-
National Science and Technology Major Project 2022ZD0208900,in part by Shanghai Municipal Scien ce and Technology Major Project under Grant 2021SHZDZX;This research is also supported by key co re technologies under Grant BE2022064-1,in part by the Lingang Laboratory under Grant No.LGL8998.

References

1.Wei Liang, Jing Jin*, Ian Daly, Hao Sun, Xingyu Wang, Andrzej Cichocki, Novel Channel Selection Model Based on Graph Convolutional Network for Motor Imagery, Cognitive Neurodynamics, 2023, 17: 1283-1296, DOI: 10.1007/s11571-022-09892-1

2.J. Jin et al., “Cross stimulus transfer learning framework using common period repetition components for fast calibration of SSVEP based BCIs,” IEEE Internet of Things Journal, 2024.
3.J. Jin et al., “A Growing Bubble Speller Paradigm for Brain-Computer Interface Based on Event-related Potentials,” IEEE Transactions on Biomedical Engineering, 2024.
4.R. Xu et al., “Multi-Scale Pyramid Squeeze Attention Similarity Optimization Classification Neural Network for ERP Detection,” Neural Networks, vol. 184, p. 107124, 2025.
5.X. Si, H. He, J. Yu, and D. Ming, “Cross-subject emotion recognition brain-computer interface based on fNIRS and DBJNet,” Cyborg and Bionic Systems, vol. 4, p. 0045, 2023.
6.X. Zhang, H. Zheng, J. Li, Z. Xie, H. Sun, and H. Wang, “Safety Decision-Making for Autonomous Vehicles Integrating Passenger Physiological States by fNIRS,” Cyborg and Bionic Systems, vol. 6, p. 0205, 2025.
7.N. Naseer and K.-S. Hong, “fNIRS-based brain-computer interfaces: a review,” Frontiers in human neuroscience, vol. 9, p. 3, 2015.
8.W. Liang, R. Xu, X. Wang, A. Cichocki, and J. Jin, “Enhancing robustness of spatial filters in motor imagery based brain-computer interface via temporal learning,” Journal of Neuroscience Methods, vol. 418, p. 110441, 2025.
9.W. Zhao, H. Lu, B. Zhang, X. Zheng, W. Wang, and H. Zhou, ‘TCANet: a temporal convolutional attention network for motor imagery EEG decoding’, Cognitive Neurodynamics, vol. 19, no. 1, p. 91, Jun. 2025.
10.W. Zhao, B. Zhang, H. Zhou, D. Wei, C. Huang, and Q. Lan, ‘Multi-scale convolutional transformer network for motor imagery brain-computer interface’, Scientific Reports, vol. 15, no. 1, p. 12935, Apr. 2025.
11.J. Li, J. Li, X. Wang, X. Zhan, and Z. Zeng, “A domain generalization and residual network-based emotion recognition from physiological signals,” Cyborg and Bionic Systems, vol. 5, p. 0074, 2024.
12.J. Ju, A. G. Feleke, L. Luo, and X. Fan, “Recognition of drivers’ hard and soft braking intentions based on hybrid brain-computer interfaces,” Cyborg and Bionic Systems, 2022.
13.M. Liu et al., “STA-Net: Spatial-temporal alignment network for hybrid EEG-fNIRS decoding,” Information Fusion, vol. 119, p. 103023, 2025.
14.Y. Kwak, W.-J. Song, and S.-E. Kim, “FGANet: fNIRS-guided attention network for hybrid EEG-fNIRS brain-computer interfaces,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30, pp. 329-339, 2022.
15.M. Meng, D. Hao, Y. Gao, L. Li, M. Wang, and X. Xi, “Decouple and Rebalance: Towards effective joint learning for EEG-fNIRS integration,” Neurocomputing, vol. 638, p. 130128, 2025.
16.H. Xu, Q. She, M. Meng, Y. Gao, and Y. Zhang, “EFDFNet: A multimodal deep fusion network based on feature disentanglement for attention state classification,” Biomedical Signal Processing and Control, vol. 109, p. 108042, 2025.
17.X. Si, S. Zhang, Z. Yang, J. Yu, and D. Ming, “A bidirectional cross-modal transformer representation learning model for EEG-fNIRS multimodal affective BCI,” Expert Systems with Applications, vol. 266, p. 126081, 2025.
18.Y. Qin et al., “ECA-FusionNet: a hybrid EEG-fNIRS signals network for MI classification,” Journal of Neural Engineering, vol. 22, no. 1, p. 016030, 2025.
19.X. Shi, H. Wang, B. Li, Y. Qin, C. Peng, and Y. Lu, “Fusion analysis of EEG-fNIRS multimodal brain signals: a multitask classification algorithm incorporating spatial-temporal convolution and dual attention mechanisms,” IEEE Transactions on Instrumentation and Measurement, 2025.
20.X. Yuan, Y. Zhang, and P. Rolfe, “IIMCNet: Intra-and Inter-modality Correlation Network for Hybrid EEG-fNIRS Brain-Computer Interface,” IEEE Journal of Biomedical and Health Informatics, 2025.

21.C. Zhang et al., “Comparing multi-dimensional fNIRS features using bayesian optimization-based neural networks for mild cognitive impairment (MCI) detection,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 1019-1029, 2023.

22.Z. Chen, C. Gao, T. Li, X. Ji, S. Liu, and M. Xiao, “Open access dataset integrating EEG and fNIRS during Stroop tasks,” Scientific Data, vol. 10, no. 1, p. 618, 2023.

23.S. M. Hosni, S. B. Borgheai, J. Mclinden, and Y. Shahriari, “An fNIRS-based motor imagery BCI for ALS: a subject-specific data-driven approach,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28, no. 12, pp. 3063-3073, 2020.

24.J. Shin et al., ‘Open Access Dataset for EEG+NIRS Single-Trial Classification’, IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 10, pp. 1735-1745, Oct. 2017.

25.J. Shin, A. von Lühmann, D.-W. Kim, J. Mehnert, H.-J. Hwang, and K.-R. Müller, ‘Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset’, Scientific Data, vol. 5, no. 1, p. 180003, Feb. 2018.

26.T.-W. Lee, M. Girolami, and T. J. Sejnowski, “Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources,” Neural computation, vol. 11, no. 2, pp. 417-441, 1999.

27.W. B. Baker, A. B. Parthasarathy, D. R. Busch, et al., Modified Beer-Lambert law for blood flow, Biomedical Optics Express, 2014, 5(11): 4053, DOI: 10.1364/BOE.5.004053.

28.I. W. Selesnick and C. S. Burrus, “Generalized digital Butterworth filter design,” IEEE Transactions on signal processing, vol. 46, no. 6, pp. 1688-1694, 2002.

29.Q. He, L. Feng, G. Jiang, and P. Xie, “Multimodal multitask neural network for motor imagery classification with EEG and fNIRS signals,” IEEE Sensors Journal, vol. 22, no. 21, pp. 20695-20706, 2022.

30.S. McKinley and M. Levine, “Cubic spline interpolation,” College of the Redwoods, vol. 45, no. 1, pp. 1049-1060, 1998.

31.Y. Song, Q. Zheng, B. Liu, and X. Gao, “EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 31, pp. 710-719, 2023, doi: 10.1109/TNSRE.2022.3230250.

32.Z. Wang, J. Fang, and J. Zhang, “Rethinking delayed hemodynamic responses for fNIRS classification,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 4528-4538, 2023.

33.Y. Li, X. Zhang, and D. Ming, “Early-stage fusion of EEG and fNIRS improves classification of motor imagery,” Frontiers in Neuroscience, vol. 16, p. 1062889, 2023.