Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

EFRM: A Multimodal EEG–fNIRS Representation-learning Model for few-shot brain-signal classification

Euijin Jung [a], Jinung An [a,b,*]

[a] Division of Intelligent Robot, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, 42988, South Korea
[b] Department of Interdisciplinary Studies, Graduate School, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, 42988, South Korea

ARTICLE INFO

Keywords:
EEG
fNIRS
Multimodal representation learning
Transfer learning
Few-shot learning

ABSTRACT

Recent advances in brain signal analysis highlight the need for robust classifiers that can be trained with minimal labeled data. To meet this demand, transfer learning has emerged as a promising strategy: large-scale unlabeled data is used to train pre-trained models, which are later adapted with minimal labeled data. However, while most existing transfer learning studies focus primarily on electroencephalography (EEG) signals, their generalization to other brain signal modalities such as functional near-infrared spectroscopy (fNIRS) remains limited. To address this issue, we propose a multimodal representation model compatible with EEG-only, fNIRS-only, and paired EEG–fNIRS datasets. The proposed method consists of two stages: a pre-training stage that learns both modality-specific and shared representations across EEG and fNIRS, followed by a transfer learning stage adapted to specific downstream tasks. By leveraging the shared domain across EEG and fNIRS, our model outperforms single-modality approaches. We constructed pre-training datasets containing approximately 1250 h of brain signal recordings from 918 participants. Unlike previous multimodal approaches that require both EEG and fNIRS data for training, our method enables adaptation to single-modality datasets, enhancing flexibility and practicality. Experimental results demonstrate that our method achieves competitive performance in comparison with state-of-the-art supervised learning models, even with minimal labeled data. Our method also outperforms previously pre-trained models, showing especially significant improvements in fNIRS classification performance.

1. Introduction

Recently, various deep learning-based methods for brain signal analysis have been proposed. However, most existing approaches [1–4] require large-scale labeled datasets to attain high classification performance. Since collecting sufficient labeled brain signal data often requires a tedious and repetitive process, recent research [5–10] has introduced a pre-training and fine-tuning approach that builds a representation model using large-scale unlabeled brain signal datasets and adapts it to specific tasks. This approach enables high classification performance even with a limited number of labeled samples.

Although pre-trained models [5–10] have shown promising results, they have been predominantly designed for electroencephalography (EEG) data, limiting their generalizability to other brain signal modalities. Non-invasive brain signals include not only EEG, which captures neuronal electrical activity [11], but also functional near-infrared spectroscopy (fNIRS), which measures hemodynamic responses using near-infrared light [11]. Both EEG and fNIRS play a vital role in Brain–Computer Interface (BCI) applications [12–15], and recent studies [16–19] have highlighted the importance of multimodal brain signal analysis, particularly using simultaneously recorded paired EEG and fNIRS data. However, despite the increasing demand for methods that can effectively handle multimodal brain signals under limited labeled data conditions, existing research [5–10] has yet to provide a comprehensive solution to this problem.

To address these limitations, we introduce the multimodal EEG–fNIRS Representation learning Model (EFRM), a novel framework designed to enhance brain signal analysis across multiple modalities. The proposed method consists of two primary training stages: (1) a pre-training stage, where the model is trained on large-scale, publicly available multimodal datasets without labels, and (2) a transfer learning stage, where the model is optimized for specific downstream tasks using only a small number of labeled samples. During the pre-training stage, the model extracts modality-specific features based on a Masked Autoencoder (MAE) [20] while simultaneously extracting shared domain between EEG and fNIRS through contrastive learning [21]. The

* Corresponding author.
E-mail address: robot@dgist.ac.kr (J. An).

https://doi.org/10.1016/j.compbiomed.2025.111292
Received 27 March 2025; Received in revised form 12 October 2025; Accepted 9 November 2025
Available online 11 November 2025

MAE-based approach applies random masking and reconstructs missing segments during training. This process enables the model to interpret the source data accurately and extract the modality-specific features effectively. The contrastive learning ensures that embeddings from simultaneously recorded multimodal brain signals are mapped closer in the latent space, while non-corresponding samples are pushed apart. By jointly leveraging MAE and contrastive learning, the proposed model captures both independent and complementary features across modalities.

The primary contributions of this study are summarized as follows: (1) We propose the first multimodal representation learning model designed to extract both modality-specific and shared domain representations from EEG and fNIRS signals. The model is pre-trained on a large-scale dataset comprising 1250 h of brain signal recordings from 918 participants, enabling self-supervised learning effectively. (2) Unlike existing multimodal supervised learning approaches [16–19] that require paired EEG–fNIRS data, our method is applicable to EEG-only, fNIRS-only, and paired EEG–fNIRS scenarios, addressing the challenge of acquiring paired multimodal datasets. (3) Quantitative evaluations demonstrate that our model achieves performance comparable to state-of-the-art supervised learning approaches [16,18,22,23] while requiring significantly fewer labeled samples. Moreover, it outperforms most existing pre-trained models [7,8,20], particularly by improving fNIRS classification performance through shared domain learning. (4) Finally, we demonstrate the importance of the shared domain by showing that increasing its amount during pre-training leads to improved downstream classification performance.

1.1. Related work

1.1.1. Representation learning for bio-signal data analysis

A foundation model is a large-scale, generalized pre-trained model built on unlabeled data, designed to transfer knowledge effectively to various downstream tasks with minimal labeled data through finetuning. Recent studies have demonstrated the effectiveness of transfer learning in minimizing training data requirements across multiple modalities, including natural language [24], image [20], speech [25], and multimodal data [26]. This approach has also been applied to bio-signals such as EEG [5–7,9,10], electrocardiography (ECG) [27], and multimodal physiological signals [8], enabling label-efficient learning. For EEG, a self-supervised contrastive learning method [7] that trains a model to distinguish between similar and dissimilar EEG samples has been introduced. This approach has successfully transferred learned knowledge to sleep stage classification, achieving high classification performance with minimal labeled data. Another notable model, BENDR [5], a transformer-based pre-trained model for EEG, demonstrated generalizability across five datasets, including motor imagery and sleep stage classification. Additionally, Pulver et al. [6] developed a masked autoencoder-based pre-trained model using emotional EEG data, which was later applied to cognitive load classification. Similarly, Chien et al. [9] proposed MAEEG, a masked autoencoder model for EEG representation learning, which reconstructs masked EEG signals to learn meaningful features and demonstrated improved performance on sleep stage classification tasks. Banville et al. [10] leveraged three pretext tasks (relative positioning, temporal shuffling, and contrastive predictive coding) for self-supervised EEG representation learning, demonstrating that frozen embeddings with a linear probe deliver label-efficient gains on sleep-stage classification and pathology detection. In the case of ECG, a masked autoencoder-based model [27] has been utilized to construct a representation learning model, allowing for robust classification of hypertrophic cardiomyopathy and ST-elevation myocardial infarction detection with minimal labeled data. Regarding multimodal bio-signal pre-training, the only existing study [8] introduced a model trained on multimodal physiological signals (electroencephalography (EEG), electrooculography (EOG), electrocardiography (ECG), electromyography (EMG), and respiratory channels) collected during polysomnography, which was then transferred for sleep stage classification. However, no prior research has introduced a multimodal representation learning model specifically for non-invasive brain signals such as EEG and fNIRS. In this study, we propose the first multimodal representation learning model designed for EEG and fNIRS analysis, leveraging transfer learning to enhance label efficiency and generalizability across downstream tasks.

1.1.2. EEG-fNIRS based multimodal brain signal analysis

EEG measures electrical activity generated by neuronal firing, offering high temporal resolution but limited spatial resolution [11]. In contrast, fNIRS measures hemodynamic responses using near-infrared light [11], offering higher spatial resolution but lower temporal resolution than EEG. Due to their complementary temporal and spatial characteristics, numerous studies [16–19,28–30] have explored the integration of these modalities to enhance brain signal analysis. EF-Net [18], as a CNN-based model, was designed to simultaneously analyze EEG's temporal dynamics and fNIRS's spatial patterns, improving mental state classification. To address the inherent spatial and temporal discrepancies between EEG and fNIRS, STA-Net [19] introduced fNIRS-Guided Spatial Alignment(FGSA) and EEG-Guided Temporal Alignment(EGTA) layers to enable more effective fusion of the two signals. Additionally, a bimodal deep learning model [16] employing CNN-GRU architectures successfully integrated EEG and fNIRS data, enhancing classification performance in both overt and imagined speech tasks. Li et al. [17] proposed an early-stage EEG–fNIRS fusion model to minimize data loss during feature extraction, improving classification accuracy. E-FNet [28] leverages a (2 + 1)D CNN architecture and a 4D tensor representation to effectively integrate the spatiotemporal characteristics of the EEG and fNIRS, thereby achieving superior accuracy and computational efficiency in comparison with existing models. M2NN [29] employs separate CNN-based spatial–temporal feature learning modules for EEG and fNIRS, integrating information through a multimodal feature fusion module. Additionally, it leverages a multitask learning (MTL) module to enhance the recognition accuracy of motor imagery tasks. FGANet [30] addresses the limitations of late fusion methods in existing hybrid EEG–fNIRS approaches, which are limited in their ability to extract correlated features between the two signals. To overcome this issue, an fNIRS-guided attention mechanism is introduced to emphasize the critical spatial features of EEG signals, while a prediction weight adjustment method is employed to mitigate performance degradation caused by the inherent delay in fNIRS responses. Despite their effectiveness, existing EEG–fNIRS multimodal models primarily rely on supervised learning, which requires a substantial amount of labeled data. Furthermore, these approaches depend on the simultaneous acquisition of EEG and fNIRS signals, limiting their flexibility in real-world applications where multimodal data collection may not always be feasible. In contrast, we propose a model that is designed to perform robustly with minimal labeled data and generalizes even with a single available modality (EEG or fNIRS), addressing the key limitations of previous multimodal approaches.

1.1.3. Neurovascular coupling

Neurovascular coupling (NVC) refers to the process in which neural activity in specific brain regions induces localized increases in blood flow [31]. Recent studies [32–35] have confirmed the presence of NVC by examining correlations between paired EEG and fNIRS signals. Lin et al. [32] demonstrated significant correlations between EEG (theta, alpha, and beta bands) and oxygenated hemoglobin (HbO) levels of fNIRS during motor execution tasks. Similarly, Blanco et al. [33] conducted experiments with 18 participants performing motor imagery tasks, revealing strong associations between EEG (beta and gamma bands) and HbO levels. During resting-state conditions, EEG (delta and theta bands) exhibited high correlations with HbO levels. Beyond motor-related tasks, Chen et al. [34] extended NVC analysis to sensory

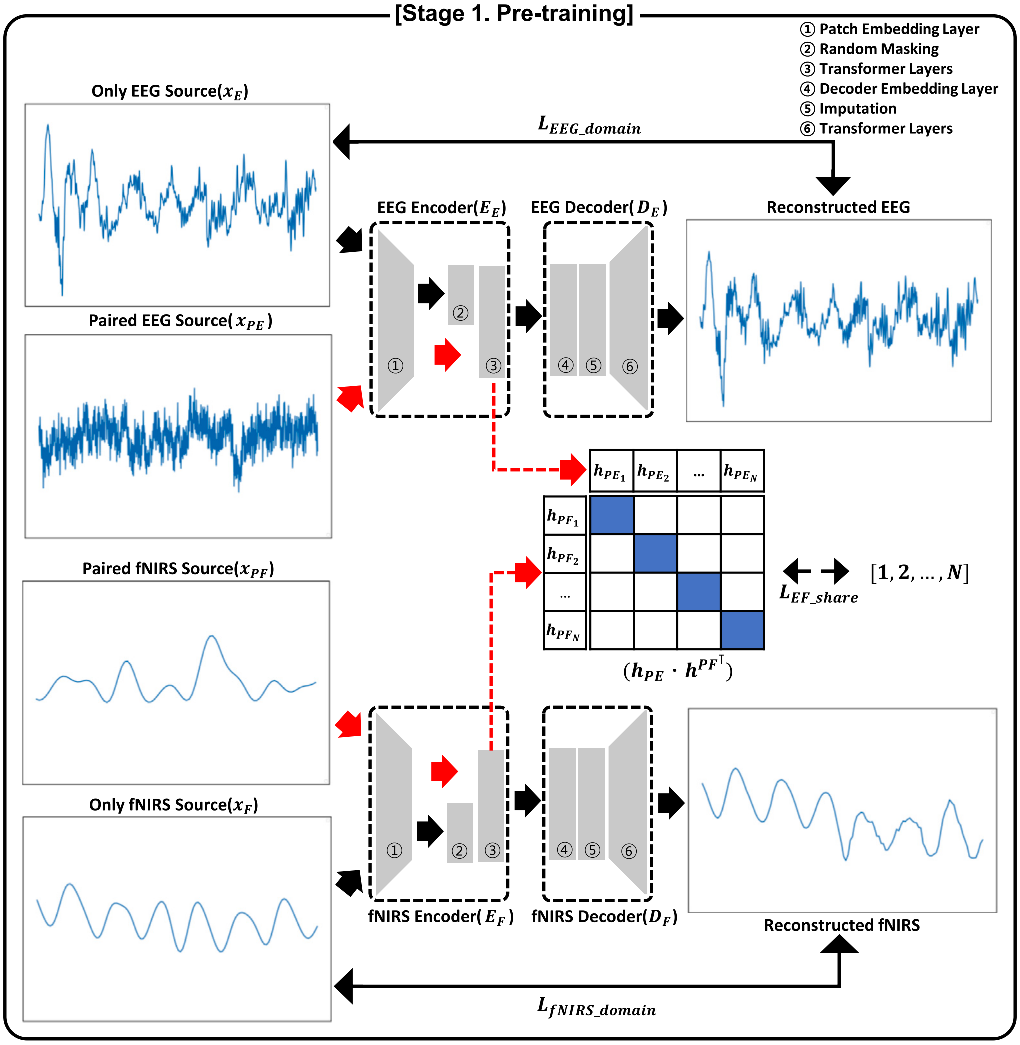

Fig. 1. The pre-training process of the proposed EEG–fNIRS multimodal representation learning model, demonstrating how modality-specific and shared domain representations are learned.

processing, identifying significant EEG–fNIRS correlations in response to visual and auditory stimuli. These studies support the robustness of NVC across different cognitive processes. Based on the established correlation between EEG and fNIRS, Li et al. [35] introduced a cross-modality learning approach that enables the transformation of EEG into fNIRS. This method generates synthetic EEG–fNIRS pairs, addressing challenges associated with acquiring real paired datasets. In our study, we also propose a method that extracts and maximizes shared information between EEG and fNIRS based on their confirmed correlation, enhancing multimodal representation learning for brain signal analysis.

2. Method

In this study, we propose the multimodal EEG–fNIRS Representation-learning Model (EFRM), which comprises two training stages: a pre-training stage for building the representation learning model and a transfer learning stage for adapting the pre-trained model to downstream tasks. As shown in Fig. 1, the pre-training stage employs two self-supervised learning approaches using EEG-only (x_E), fNIRS-only (x_F), and paired EEG (x_{PE})–fNIRS (x_{PF}) datasets to extract three feature domains: the EEG-specific domain, the fNIRS-specific domain, and the shared EEG–fNIRS domain. The modality-specific domains are learned using a masked autoencoder, which reconstructs randomly masked features, enabling the encoder to develop strong general feature extraction capabilities for each modality. Meanwhile, contrastive learning is applied to maximize shared representations between x_{PE} and x_{PF} by aligning their corresponding embeddings while distancing non-corresponding pairs, following pre-training strategies established in image and language contrastive learning [21]. As shown in Fig. 2, the transfer learning stage adapts the pre-trained model for downstream tasks, supporting both single-modality (x_E or x_F) and multimodal input (x_{PE} and x_{PF}). This approach mitigates the limitations posed by conventional multimodal supervised learning, which requires simultaneous data from both modalities. To further improve label efficiency, a linear probing method is employed, enabling effective adaptation to specific tasks with only a small amount of labeled data. The detailed pre-training and transfer learning procedures are described in the following sections. Also, the detailed model architecture is provided in Table 1.

2.1. Pre-training stage

As shown in Fig. 1, the pre-training process involves learning three domains: the EEG domain, the fNIRS domain, and the shared EEG–fNIRS domain. We extract the EEG representation by feeding x_E (1×24 channels $\times$ 1024 time points) into a three-stage EEG encoder (E_E). We adopt a patch-based masked-autoencoding paradigm: before feature extraction, x_E is partitioned into fixed-size patches that are linearly projected into patch tokens. A patch is a contiguous, non-overlapping slice of x_E with size 1 along the channel axis by 32 along the time axis. Using a patch size of 1 × 32 with a stride of 1 × 32, x_E is partitioned into a grid of 24 rows and 32 columns. Each 1 × 32 patch is then mapped by the patch-embedding layer to a fixed-dimensional token (24 × 32). The choice of the 1 × 32 patch size was determined by an ablation study (Section 4.3.1). Positional information is then added to the extracted embeddings to preserve the relative positions of the vectors. Next, a random masking step is applied, where 50% of the compressed features are masked—a ratio chosen by considering both performance and computational efficiency, as determined through an ablation study (Section 4.3.2). Here, we define the masking region $m \in \{0,1\}^d$ as a binary mask over the feature axis of x_E (where d denotes the feature dimension of x_E; see Table 1 for architectural details). It selects 50% of positions uniformly at random. Applied to x_E, this selection reduces a 768 × 768 embedding to 384 × 768.

The embeddings are then processed through a transformer layer, capturing long-range dependencies using self-attention mechanisms. This process is repeated across 12 transformer blocks, each employing a 12-head multi-head attention mechanism, following the architecture used in ViT [36]. The encoded EEG features are subsequently fed into the EEG decoder (D_E), which reconstructs the masked regions through: (1) a decoder embedding layer applies a linear transformation, (2) imputation is performed by inserting random values into masked regions, followed by the addition of positional information, and (3) a transformer layer predicts the masked segments. The transformer layer consists of 8 transformer blocks, maintaining the same structure as the transformer blocks in E_E. The difference between the reconstructed EEG and x_E is minimized to effectively learn the EEG domain representations. The objective function is formally defined as:

$$L_{\text{EEG\_domain}} = \mathbb{E}_{x_E} \left\| \left(D_E(E_E(x_E)) - x_E \right) \times m \right\|_2^2 . \tag{1}$$

The fNIRS domain is extracted following the same methodology as EEG domain training, where x_F (2 (oxygenated and deoxygenated hemoglobin) × 64 channels × 128 time points) is processed through a fNIRS encoder (E_F) and a decoder (D_F). The structure of E_F follows patch embedding, random masking, and transformer layers for long-range dependency learning. Additionally, the patch size in the patch embedding step is set to 1 × 32, where 1 represents the channel axis and 32 corresponds to the time axis (approximately 2 s), based on ablation study results. Applying a patch-embedding layer with kernel size 1 × 32 to x_F yields a token sequence of shape 64 × 4. Except these modifications, the architecture of E_F follows that of E_E. Similarly, D_F reconstructs the masked regions using embedding, imputation, and transformer layers, with an objective function minimizing the difference between the reconstructed fNIRS and x_F. The objective function is formally defined as:

$$L_{\text{fNIRS\_domain}} = \mathbb{E}_{x_F} \left\| \left(D_F(E_F(x_F)) - x_F \right) \times m \right\|_2^2 . \tag{2}$$

The learning procedure for the EEG–fNIRS shared domain is formulated over inputs x_{PE} and x_{PF}, where x_{PE} has the same dimensionality as x_E and x_{PF} has the same dimensionality as x_F. In this process, x_{PE} and x_{PF} are separately input into E_E and E_F, respectively. Each signal passes through the patch embedding layer, the transformer layer without masking, and global average pooling, where they are transformed into embedding vectors (h_{PE} and h_{PF}). Specifically, we obtain pooled embeddings $h_{PE} \in \mathbb{R}^{N \times 768}$ and $h_{PF} \in \mathbb{R}^{N \times 768}$, where N denotes the mini-batch size. To maximize shared information between the two embeddings, contrastive learning is applied. This ensures that the embeddings of x_{PE} and x_{PF} are mapped closer together in the latent space, while non-corresponding embeddings are pushed apart. To achieve this, the extracted h_{PE} and h_{PF} from each modality undergo L2 normalization and matrix multiplication, producing two $N \times N$ batch–batch embedding matrices (h_{PEF1} and h_{PEF2}), defined as:

$$\|h_{PE}\| = \text{norm}_{\ell_2}(h_{PE}), \qquad \|h_{PF}\| = \text{norm}_{\ell_2}(h_{PF}) \tag{3}$$

$$h_{\text{PEF1}} = \|h_{PE}\| \, \|h_{PF}\|^{\top}, \qquad h_{\text{PEF2}} = \|h_{PF}\| \, \|h_{PE}\|^{\top} \tag{4}$$

Through this process, the obtained embeddings are used to compute the cross-entropy loss, defined as follows (where y denotes the ground-truth indices (in a mini-batch of size N) that distinguish *paired* from *unpaired* embeddings).

$$\text{L}_{\text{CE}}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} y_i \log\left(\hat{y}_i\right) \tag{5}$$

$$L_{\text{EF\_share1}} = \text{L}_{\text{CE}}\left(y,\ h_{\text{PEF1}}\right), \qquad L_{\text{EF\_share2}} = \text{L}_{\text{CE}}\left(y,\ h_{\text{PEF2}}\right) \tag{6}$$

$$L_{\text{EF\_share}} = \frac{L_{\text{EF\_share1}} + L_{\text{EF\_share2}}}{2}, \tag{7}$$

Minimizing this objective pulls paired embeddings together and pushes unpaired embeddings apart. This contrastive learning strategy helps the model better capture mutual information between the two modalities, allowing it to extract shared features that may not be easily identifiable within a single modality. The defined loss function is employed for extracting the shared domain between EEG and fNIRS.

The overall pre-training loss function is formulated as the sum of three components: $L_{\text{EEG\_domain}}$, $L_{\text{fNIRS\_domain}}$, $L_{\text{EF\_share}}$. This combined loss function ensures that the model effectively learns both modality-specific and shared representations. The complete formulation of the pre-training loss function is defined as:

$$L_{\text{pre-train}} = L_{\text{EEG\_domain}} + L_{\text{fNIRS\_domain}} + L_{\text{EF\_share}}. \tag{8}$$

2.2. Transfer learning stage

As shown in Fig. 2, the proposed method provides an optimization strategy tailored to each modality configuration. The features of x_E and x_F are extracted via E_E and E_F, respectively. For x_{PE} and x_{PF}, features extracted from both encoders are aggregated through summation. The extracted features are then compressed using a newly introduced class-specific fully connected layer, which adjusts the feature dimension based on the number of target classes in the downstream task. In the proposed method, both EEG and fNIRS features are ultimately compressed into a vector of size N × 768, where N is the batch size and 768 is the embedding dimension. This vector is then passed through the fully connected layer and projected to binary or ternary outputs, depending on the number of classes in the downstream task. To minimize the difference between the predicted output and the ground truth labels (y), the model is trained using a cross-entropy loss function, defined as follows:

$$L_c = \text{L}_{\text{CE}}\left(y,\ \text{FC}_{\text{class}}(E(x_f))\right). \tag{9}$$

During this process, a linear probing approach is employed, where the parameters of the pre-trained layers remain frozen, and only the newly introduced linear layer is trained.

3. Experimental settings

3.1. Pre-training datasets

To construct the proposed pre-training model, three types of datasets were required: EEG-only, fNIRS-only, and paired EEG–fNIRS recordings. To develop a representation learning model adaptable to various downstream tasks, we utilized nine publicly available datasets covering diverse tasks, including data from 918 participants and totaling approximately 1250 h of recorded brain signals. All EEG datasets were processed through a common preprocessing pipeline, including resampling to 128 Hz, segmenting into fixed 8-second windows, applying a band-pass filter (0.5–50 Hz), and performing normalization (mean = 0, standard deviation = 1) for each channel. fNIRS data preprocessing involved converting raw signals into oxygenated (HbO) and

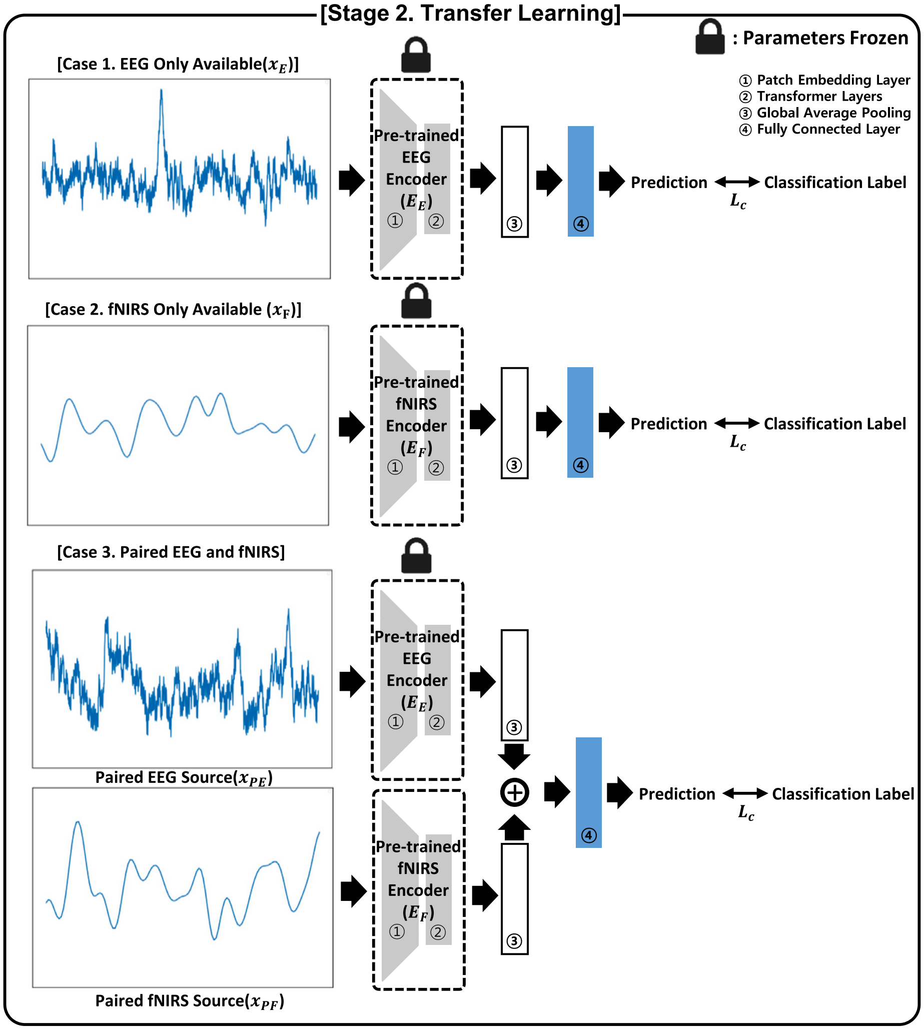
[Stage 2. Transfer Learning]
: Parameters Frozen
① Patch Embedding Layer
② Transformer Layers
③ Global Average Pooling
④ Fully Connected Layer
[Case 1. EEG Only Available(x_E)]
Pre-trained EEG Encoder (E_E)
① ②
③
④
Prediction
Classification Label
L_c
[Case 2. fNIRS Only Available (x_F)]
Pre-trained fNIRS Encoder (E_F)
① ②
③
④
Prediction
Classification Label
L_c
[Case 3. Paired EEG and fNIRS]
Pre-trained EEG Encoder (E_E)
① ②
③
Paired EEG Source(x_{PE})
Pre-trained fNIRS Encoder (E_F)
① ②
③
Paired fNIRS Source(x_{PF})
⊕
④
Prediction
Classification Label
L_c

Fig. 2. The transfer learning process to transfer the pre-trained knowledge for downstream tasks.

deoxygenated hemoglobin (HbR) concentrations using the Modified Beer–Lambert Law (MBLL). Additional preprocessing steps included resampling to 16 Hz, segmenting into 8-second windows, applying a band-pass filter (0.01–0.1 Hz), and performing normalization(mean = 0, standard deviation = 1) for each channel. In the proposed method, EEG datasets were standardized to 24 channels, while fNIRS datasets were set to 64 HbO and 64 HbR channels. These channel configurations were determined based on the number of available channels across both pre-training and downstream task datasets. If a dataset contained more channels than required, we randomly selected the required N consecutive channels. If a dataset had fewer channels, existing channels were duplicated to match the required input dimensions. All datasets used during the pre-training stage are summarized in Table 2, with detailed descriptions following.

EEG-only Dataset: We used three publicly available EEG datasets, totaling 868 h of data from 766 participants. The Temple University Seizure Corpus (TUSZ) [37] contains EEG recordings of both seizure events and resting states. The dataset spans approximately 786 h, with seizure events accounting for 6.8% of the data. The Motor Movement and Imagery Dataset [38] includes EEG recordings from 109 participants performing various motor execution and imagery tasks, such as grasping and releasing with either hand, mentally imagining these movements, and performing bilateral hand or foot movements. The SEED dataset [39,40] captures EEG responses associated with positive, negative, and neutral emotional states from 15 participants. During data collection, participants watched emotionally evocative video clips, followed by self-assessments of their emotional states.

fNIRS-only Dataset: We used five publicly available fNIRS datasets, comprising a total of 364 h of data from 123 participants. The Speech Perception A Dataset [41] includes fNIRS recordings from 38 participants, analyzing the relationship between temporal lobe activity and speech perception accuracy. Participants repeatedly listened to spoken sentences while brain activity was recorded. The Speech Perception B Dataset [42] extends this study by including both normal-hearing individuals and tinnitus patients, examining the effects of tinnitus on functional brain connectivity. The Neurofeedback Dataset [43] consists of fNIRS recordings from 19 participants, including individuals with ADHD and neurotypical controls, undergoing neurofeedback training and resting-state measurements. During the neurofeedback task, participants regulated prefrontal oxygenation levels in response to real-time visual feedback displayed on a monitor in a dark environment. The Finger and Foot Tapping Dataset [44] captures fNIRS signals from 30 participants performing both hands finger and foot tapping tasks to analyze motor-related brain activity. The Motor Observation and Execution Dataset [45] includes fNIRS recordings of participants during motor observation and execution tasks. Motor observation data were collected while participants watched videos of hand movements, categorized into simple and complex actions. Motor execution data were recorded as they physically performed the observed movements.

Paired EEG–fNIRS Dataset: We used a publicly available paired EEG–fNIRS dataset [46,47] for motor imagery tasks, where participants

Table 1
Model architecture of the proposed EEG-fNIRS multimodal representation learning model.

Modality	Encoder/Decoder	Layer	Operation type	Input dimension	Output dimension	Key parameters
EEG	EEG encoder	Patch embedding layer	Conv2d	$(N, 1, 24, 1024)$	$(N, 768, 24, 32)$	Patch size = $(1, 32)$, Stride = $(1, 32)$, Dim = 768
			Flatten + Identity norm	$(N, 768, 24, 32)$	$(N, 768, 768)$	–
		Positional embedding	Sin–Cos positional embedding	$(N, 768, 768)$	$(N, 768, 768)$	–
		Random masking	Random masking	$(N, 768, 768)$	$(N, 384, 768)$	Masking ratio = 0.5
		Transformer blocks	Add class token	$(N, 384, 768)$	$(N, 385, 768)$	–
			Multi-head self attention	$(N, 385, 768)$	$(N, 385, 768)$	# of blocks = 12, # of heads = 12, MLP ratio = 4
			LayerNorm	$(N, 385, 768)$	$(N, 385, 768)$	Dim = 768
	EEG decoder	Decoder embedding layer	Linear	$(N, 385, 768)$	$(N, 385, 512)$	Dim = 512
		Imputation	Impute mask token	$(N, 385, 512)$	$(N, 769, 512)$	–
		Positional embedding	Sin–Cos positional embedding	$(N, 769, 512)$	$(N, 769, 512)$	–
		Transformer blocks	Multi-head self attention	$(N, 769, 512)$	$(N, 769, 512)$	# of blocks = 8, # of Heads = 16, MLP ratio = 4
			Linear	$(N, 769, 512)$	$(N, 769, 32)$	Dim = 32
			Remove class token	$(N, 769, 32)$	$(N, 768, 32)$	–
			Reshape	$(N, 768, 32)$	$(N, 1, 24, 1024)$	–
fNIRS	fNIRS encoder	Patch embedding layer	Conv2d	$(N, 2, 64, 128)$	$(N, 768, 64, 4)$	Patch size = $(1, 32)$, Stride = $(1, 32)$, Dim = 768
			Flatten + Identity norm	$(N, 768, 64, 4)$	$(N, 256, 768)$	–
		Positional embedding	Sin–Cos positional embedding	$(N, 256, 768)$	$(N, 256, 768)$	–
		Random masking	Random masking	$(N, 256, 768)$	$(N, 128, 768)$	Masking ratio = 0.5
		Transformer blocks	Add class token	$(N, 128, 768)$	$(N, 129, 768)$	–
			Multi-head self attention	$(N, 129, 768)$	$(N, 129, 768)$	# of blocks = 12, # of heads = 12, MLP ratio = 4
			LayerNorm	$(N, 129, 768)$	$(N, 129, 768)$	Dim = 768
	fNIRS decoder	Decoder embedding layer	Linear	$(N, 129, 768)$	$(N, 129, 512)$	Dim = 512
		Imputation	Impute mask token	$(N, 129, 512)$	$(N, 257, 512)$	–
		Positional embedding	Sin–Cos positional embedding	$(N, 257, 512)$	$(N, 257, 512)$	–
		Transformer blocks	Multi-head self attention	$(N, 257, 512)$	$(N, 257, 512)$	# of blocks = 8, # of heads = 16, MLP ratio = 4
			Linear	$(N, 257, 512)$	$(N, 257, 64)$	Dim = 64
			Remove class token	$(N, 257, 64)$	$(N, 256, 64)$	–
			Reshape	$(N, 256, 64)$	$(N, 2, 64, 128)$	–

Table 2
Summary of pre-training and downstream task datasets.

Type	Modality	Task	# of subjects	# of channels	Recording time (h)	Age	Sex (%) (F/M)
Pre-training	EEG-only	Seizure events [37]	642	25∼129	786	51.6	51/49
		Motor movement and imagery [38]	109	64	48	–	–
		Emotion states [39,40]	15	62	34	23.3 ± 2.4	53/47
	fNIRS-only	Speech perception A [41] (Normal Subjects)	38	14	38	24.97	67/33
		Speech perception B [42] (Normal and Tinnitus Patients)	18	44	5.75	Normal: 25.4 ± 7.3, Patients: 48.6 ± 16.0	39/61
		Neurofeedback training [43]	19	44	312	30.4 ± 9.3	32/68
		Finger and foot tapping [44]	30	20	6.3	23.4 ± 2.5	43/57
		Motor observation and execution [45]	18	45	1.75	33.5 ± 15.5	57/43
	Paired EEG & fNIRS	Motor imagery [46,47]	29	EEG: 32, fNIRS: 36	15.5	28.5 ± 3.7	52/48
Downstream task	EEG-only	Sleep stages [48]	78	1	720	58.8 ± 22.2	53/47
	fNIRS-only	Mental arithmetic [49]	8	52	4	26.0 ± 2.8	63/37
	Paired EEG & fNIRS	Alertness and sleep [50]	11	EEG: 18, fNIRS: 38	35	27.6	32/68

performed imagined hand movements in response to visual (left/right arrows) and auditory cues. EEG and fNIRS signals were simultaneously recorded using electrodes and optodes embedded in the same recording cap. Participants sat in a well-lit room, facing a 50-inch screen while following task instructions.

3.2. Downstream task datasets

To evaluate the modality-specific adaptability of the pre-trained model, we assessed its performance on downstream tasks using different modality configurations: (1) an EEG-only dataset, (2) a fNIRS-only dataset, and (3) a paired EEG–fNIRS dataset. The downstream task datasets were preprocessed using the same procedures as in pre-training to maintain consistency. Table 2 summarizes all datasets used in the transfer learning stage; detailed descriptions follow.

EEG-only Dataset: We used EEG-only data [48] collected from 78 participants to analyze brain activity across different sleep stages. Sleep, typically divided into five stages, was simplified into binary (alertness, sleep) and ternary (alertness, Non-REM, REM) classifications to mitigate the risk of overfitting, which can occur when training data is limited. The data was recorded as single-channel EEG based on the Fpz-Cz potential difference. In addition to EEG-only downstream evaluation, this single-channel dataset was used to assess the pre-trained model's transferability to a single-channel setting.

fNIRS-only Dataset: We used fNIRS-only data [49] collected from 8 participants who performed a mental arithmetic task and resting state sessions. Each participant was monitored for approximately 30 min, resulting in 4 h of fNIRS recordings across 52 channels. During the experiment, participants viewed simple addition or subtraction problems on a screen and performed mental arithmetic tasks, interspersed with resting periods.

Paired EEG–fNIRS Dataset: Simultaneous EEG and fNIRS data were recorded from 11 participants, capturing alertness and sleep states. During the experiment, participants were exposed to auditory beep sounds, presented randomly to either the left or right side. If a participant responded, the trial was labeled as alertness, whereas no response was labeled as sleep. The study protocol was reviewed and approved by the Institutional Review Board (IRB) of DGIST (DGIST-171011-HR-035-01). All participants were informed of the study's purpose and provided written consent prior to participation. All procedures adhered to the ethical guidelines set by the IRB. Further details regarding the data collection process can be found in our previous study [50].

3.3. Training setting and baseline methods

The detailed configurations for constructing the proposed pre-trained model and optimizing it for downstream tasks are as follows. For pre-training, the initial learning rate was set to 0.0001 and progressively reduced using cosine annealing scheduling. The model was trained for 1,500,000 iterations over eight days on an NVIDIA RTX A6000 GPU with a batch size of 32. For downstream task optimization, linear probing was performed with an initial learning rate of 0.0001, which was gradually reduced over 200 iterations. All training and testing processes were implemented using the PyTorch deep learning framework. The weighted Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$) was used for both the pre-training and transfer learning stages. Our code is available at https://github.com/EuijinMisp/EFRM-A-Multimodal-EEG-fNIRS-Representation-learning-Model. To quantitatively evaluate the proposed method, we used accuracy and F1-score as performance metrics. Accuracy provides an overall measure of correct predictions, while F1-score accounts for class imbalance by considering both precision and recall. Using both metrics ensures a more comprehensive evaluation, capturing overall performance while addressing potential biases in class distributions. To validate the effectiveness of the proposed approach, we conducted experiments comparing it with state-of-the-art supervised and self-supervised learning models. Comparisons with supervised learning methods allowed us to assess its ability to reduce dependency on labeled data. The specific baseline methods for each modality are detailed below.

Baseline Methods for EEG-only Data: Among the comparison methods, self-supervised learning models were pre-trained solely on EEG-only datasets before being optimized for downstream tasks. The details of these methods are introduced below. MAE (Masked Autoencoder) [20] pre-trains an encoder and decoder by randomly masking parts of the input data and reconstructing them. The rationale for selecting MAE, originally proposed for the image domain, as a comparison method in this experiment is as follows: (1) Several existing pre-training models [6,9] for EEG have adopted MAE-based approaches.(2) However, open-source implementations of these models are unavailable, making reproducibility challenging. (3) Our proposed method also builds on the MAE approach for learning modality-specific domains. To ensure a fair comparison, the baseline MAE was adapted for EEG and fNIRS by applying the same random masking ratio (50%) and patch size configurations, as determined through the ablation study (Section 4.3). However, in contrast to MAE, our method is specifically designed for EEG–fNIRS multimodal learning, leading to fundamental differences in the usage of training dataset and shared feature extraction process. SSCL (Self-Supervised Contrastive Learning) [7] constructs a pre-trained model by generating two augmented samples from the same EEG signal and applying contrastive learning, pulling similar representations closer while pushing dissimilar ones apart. SleepFM [8] is designed for both single- and multimodal physiological signal classification using self-supervised learning. It employs a CNN-based encoder and applies pairwise contrastive learning, where embeddings are trained based on positive and negative pairs. Conformer [22] is a state-of-the-art supervised learning model that combines convolutional layers and transformers to capture both local and global EEG features. Conformer has demonstrated strong performance across various EEG-based tasks, making it a robust baseline for comparison.

Baseline Methods for fNIRS-only Data: For the fNIRS dataset, we selected the same pre-training baseline models used for EEG, as no specialized pre-training models for fNIRS have been studied. However, self-supervised learning models were pre-trained solely on the fNIRS-only dataset. In the case of supervised learning, instead of Conformer [22], which is optimized for EEG, we used fNIRS-T [23], a state-of-the-art model specifically designed for fNIRS to extract spatial- and channel-level features. fNIRS-T [23] has demonstrated strong performance across multiple fNIRS-related tasks, making it a suitable supervised learning baseline.

Baseline Methods for Paired EEG–fNIRS Data: For paired EEG–fNIRS pre-training, SleepFM [8] was used as the baseline model, trained on the paired EEG–fNIRS dataset. Unlike its single-modal counterpart, the multimodal version of SleepFM adopts a Leave-One-Out Contrastive Learning approach, which differs from standard contrastive learning. For supervised multimodal learning, we selected EFNet [18] and BimodalNet [16] as baseline models. EFNet consists of separate encoders for high-frequency EEG and low-frequency fNIRS, extracting modality-specific features before combining them for classification. BimodalNet also utilizes separate encoders for feature extraction but differs from EFNet by incorporating a Gated Recurrent Unit (GRU) to address the long-term dependency issues of CNNs. These models serve as strong baselines for evaluating our proposed multimodal representation learning model.

4. Results and discussion

4.1. Few-shot classification

Table 3 shows that the proposed method achieves classification performance comparable to state-of-the-art supervised learning methods [16,18,22,23], despite using significantly fewer labeled samples

Table 3
Classification accuracy of the proposed and baseline methods on downstream task datasets. Performance is evaluated depending on the number of shots used for downstream task training. One-shot means one sample per class. Red text indicates the best score in each shot.

Methods	EEG-only dataset [48] (Alertness/Sleep)				Methods	EEG-only dataset [48] (Alertness/Non-REM Sleep/REM Sleep)			
Pre-trained models	1shot	2shots	4shots	8shots	Pre-trained models	1shot	8shots	32shots	64shots
MAE [20]	0.59 ± 0.14	0.60 ± 0.14	0.75 ± 0.12	0.70 ± 0.12	MAE [20]	0.37 ± 0.07	0.59 ± 0.14	0.59 ± 0.15	0.64 ± 0.14
SSCL [7]	0.52 ± 0.11	0.55 ± 0.11	0.63 ± 0.12	0.71 ± 0.11	SSCL [7]	0.35 ± 0.09	0.49 ± 0.08	0.51 ± 0.12	0.53 ± 0.09
SleepFM [8]	0.35 ± 0.09	0.45 ± 0.11	0.51 ± 0.11	0.48 ± 0.11	SleepFM [8]	0.25 ± 0.09	0.29 ± 0.10	0.28 ± 0.11	0.43 ± 0.07
EFRM (ours)	0.60 ± 0.14	0.61 ± 0.12	0.73 ± 0.11	0.74 ± 0.11	**EFRM (ours)**	0.35 ± 0.09	0.60 ± 0.13	0.63 ± 0.11	0.65 ± 0.11
Supervised learning	**100shots**	**200shots**	**400shots**	**800shots**	**Supervised learning**	**100shots**	**800shots**	**3200shots**	**6400shots**
Conformer [22]	0.46 ± 0.08	0.54 ± 0.10	0.63 ± 0.07	0.77 ± 0.08	Conformer [22]	0.37 ± 0.08	0.57 ± 0.07	0.71 ± 0.08	0.76 ± 0.06
Methods	fNIRS-only dataset [49] (Mental arithmetic/Resting state)				Methods	Paired EEG and fNIRS dataset [50] (Alertness/Sleep)			
Pre-trained models	2shots	4shots	6shots	8shots	Pre-trained models	1shot	2shots	4shots	6shots
MAE [20]	0.67 ± 0.26	0.67 ± 0.12	0.75 ± 0.10	0.75 ± 0.10	–	–	–	–	–
SSCL [7]	0.70 ± 0.12	0.73 ± 0.13	0.69 ± 0.22	0.69 ± 0.22	–	–	–	–	–
SleepFM [8]	0.75 ± 0.06	0.62 ± 0.09	0.60 ± 0.07	0.58 ± 0.13	SleepFM [8]	0.62 ± 0.03	0.35 ± 0.21	0.65 ± 0.07	0.71 ± 0.10
EFRM (ours)	0.90 ± 0.07	0.81 ± 0.12	0.85 ± 0.11	0.94 ± 0.06	**EFRM (ours)**	0.37 ± 0.22	0.42 ± 0.13	0.80 ± 0.09	0.80 ± 0.10
Supervised learning	**6shots**	**12shots**	**18shots**	**24shots**	**Supervised learning**	**100shots**	**200shots**	**400shots**	**600shots**
fNIRS-T [23]	0.54 ± 0.15	0.58 ± 0.13	0.69 ± 0.10	0.88 ± 0.09	EFNet [18]	0.68 ± 0.11	0.71 ± 0.14	0.85 ± 0.18	0.86 ± 0.10
–	–	–	–	–	BimodalNet [16]	0.52 ± 0.03	0.60 ± 0.09	0.79 ± 0.17	0.85 ± 0.08

Table 4
Statistical significance of performance differences between the proposed method and baseline models, reported as p-values from t-tests. Statistically significant cases ($p < 0.05$) are highlighted in red.

Methods	EEG-only dataset [48] (Alertness/Sleep)				Methods	EEG-only dataset [48] (Alertness/Non-REM Sleep/REM Sleep)			
Pre-trained models	1shot	2shots	4shots	8shots	Pre-trained models	1shot	8shots	32shots	64shots
MAE [20]	0.663	0.542	0.292	0.049	MAE [20]	0.634	0.886	0.419	0.711
SSCL [7]	<0.01	<0.01	<0.01	0.111	SSCL [7]	1.000	0.016	0.011	<0.01
SleepFM [8]	<0.01	<0.01	<0.01	<0.01	SleepFM [8]	<0.01	<0.01	<0.01	<0.01
Methods	fNIRS-only dataset [49] (Mental arithmetic/Resting state)				Methods	Paired EEG and fNIRS dataset [50] (Alertness/Sleep)			
Pre-trained models	2shots	4shots	6shots	8shots	Pre-trained models	1shot	2shots	4shots	6shots
MAE [20]	<0.01	0.033	0.096	0.011	–	–	–	–	–
SSCL [7]	0.018	0.351	0.049	<0.01	–	–	–	–	–
SleepFM [8]	0.033	0.048	<0.01	<0.01	SleepFM [8]	0.579	<0.01	<0.01	0.042

across EEG-only, fNIRS-only, and paired EEG–fNIRS datasets. Additionally, the results indicate that the proposed method outperforms most existing pre-trained models, achieving the highest classification accuracy in multiple cases (highlighted in red). Notably, on the fNIRS-only dataset, our approach significantly outperforms MAE [20], even though both methods share the same encoder architecture. Table 4 further confirms the statistical significance of these improvements, showing that the proposed method demonstrates significant differences in comparison with most existing pre-trained models (highlighted in red). To further analyze the results in Table 3, we conducted additional evaluations using the F1-score. As shown in Fig. 3, the proposed method consistently achieves the highest performance across all datasets, with the most substantial improvement observed in the fNIRS-only dataset. These results demonstrate that the pre-trained model built with the proposed method effectively clustered generalizable multimodal brain signal representations. The clustered brain signals allowed a simple linear layer, trained with a small amount of labeled data, to achieve higher classification performance in comparison with existing methods.

4.2. Analysis of shared domain

In this experiment, we investigate how the shared domain extracted during pre-training contributes to classification performance in downstream tasks. To this end, we evaluate the downstream performance of pre-trained models trained with varying amounts of shared domain in Fig. 5. Specifically, we constructed multiple pre-trained models using a single fNIRS dataset and EEG data with different levels of signal richness. To control the amount of shared domain, we varied the EEG preprocessing by gradually widening the bandpass filter ranges: 0.5–4 Hz, 0.5–13 Hz, and 0.5–50 Hz. These frequency ranges were selected based on confirmed findings from a previous EEG study [51], considering brainwave bands from Delta to Gamma.

In Table 5, mutual information was quantified to assess whether pre-trained models using different EEG frequency bands show measurable differences in the amount of shared domain. Conditional entropy and mutual information were computed for the modality-specific and shared domains using embeddings extracted from paired EEG–fNIRS data through pre-trained models. The locations of each domain in Table 5 are visualized in Fig. 4. The results confirm that narrowing the EEG preprocessing bandwidth leads to a reduction in the amount of shared domain. Fig. 5 shows the downstream classification performance relative to the amount of shared domain. The results demonstrate that an increased shared domain leads to improved classification performance, indicating that the richness of shared representations directly influences model effectiveness. We additionally compare these results to the performance of our model trained solely on fNIRS-only datasets. Notably, the pre-trained model using EEG in the 0.5–4 Hz band performs worse than the fNIRS-only model, suggesting that limited shared representation between EEG and fNIRS hinders effective pre-training. This performance degradation arises when the lack of shared domain

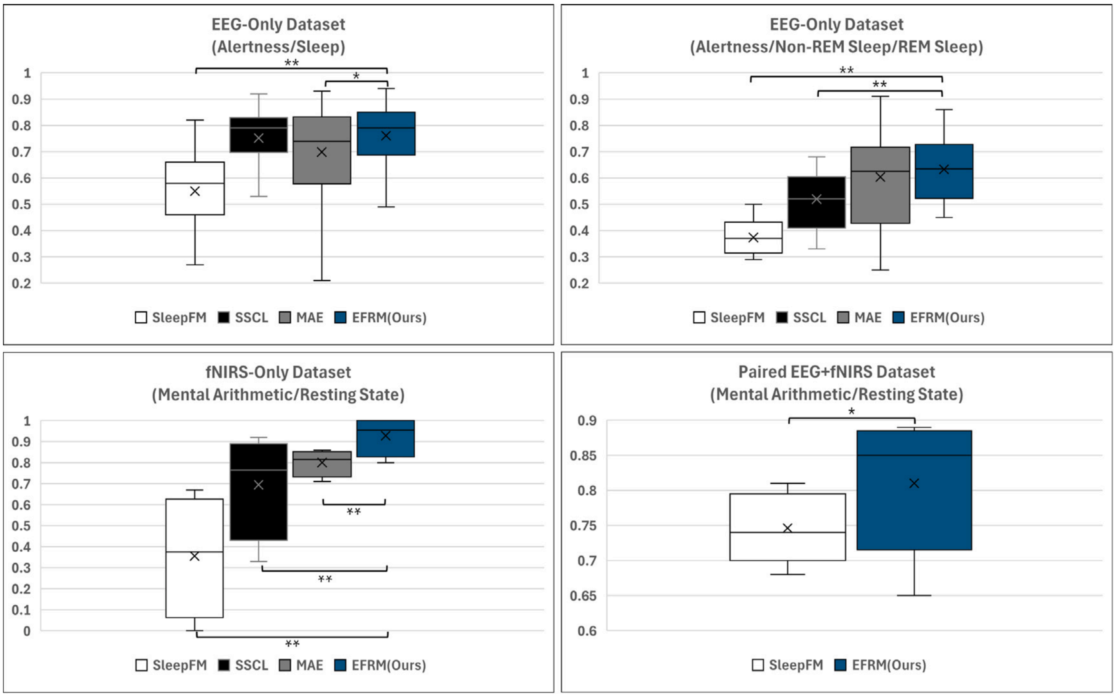

Fig. 3. Boxplots of measured F1 scores for the proposed and baseline Methods. The top, middle, and bottom lines of each box indicate the upper quartile, median, and lower quartile scores, respectively, with whiskers extending to the highest and lowest scores. The × symbol represents the average score. * indicates a significant difference ($p < 0.05$), and ** indicates a highly significant difference ($p < 0.01$). The proposed and baseline methods were optimized based on the maximum number of shot for each dataset (as shown in Table 3).

Table 5
Quantification of shared and modality-specific domains using mutual information and conditional entropy.

EEG band pass filter ranges	I(E;F)	H(E|F)	H(F|E)
Paired fNIRS and EEG (0.5~50 Hz)	0.769	0.091	0.089
Paired fNIRS and EEG (0.5~13 Hz)	0.625	0.109	0.112
Paired fNIRS and EEG (0.5~4 Hz)	0.486	0.259	0.259

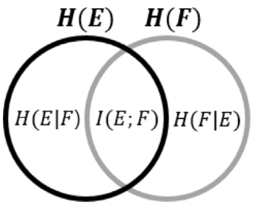

Fig. 4. A visual representation of shared and modality-specific domains across EEG and fNIRS.

between modalities hinders the learning of consistent mutual domain representations.

Fig. 6 shows the impact of shared domain quantity on the learning process during pre-training. The proposed method for shared feature extraction is trained using contrastive learning, where distinguishing between positive and negative samples is essential. When the shared domain is sufficiently rich, as shown on the left in Fig. 6, positive and negative samples are well-separated, enabling consistent and effective contrastive learning. In contrast, as shown on the right in Fig. 6, when the shared domain is limited, similar embeddings are involved in both attraction and repulsion processes, making it difficult for the model to learn stable and discriminative shared representations.

4.3. Ablation study

4.3.1. Pre-training stage: Optimizing encoder patch size

In this experiment, we investigated the optimal patch size for EEG and fNIRS encoders. As shown in Table 6, we first determined the optimal patch size by experimenting with three different patch sizes per modality. To ensure that the models were not influenced by the other modality, we first pre-trained EFRM separately on the EEG and fNIRS domains. Once the optimal patch sizes for each modality were identified, we applied them to the proposed method, which utilizes both modalities, for comparison. For EMFM trained solely on the EEG and fNIRS domains, the approach follows the same methodology as conventional MAE [20].

The results from the single-domain method experiments indicate that both EEG and fNIRS achieve the highest performance with a patch size of 0.25 and 2, respectively. Because EEG is a high-frequency signal, its encoder requires local-level information extraction, whereas fNIRS, with its slow hemodynamic responses, relies on global-level information extraction more heavily. The multimodal EFRM outperforms the single-modality method across most evaluation metrics, demonstrating that pre-training with multimodal data enhances performance even with the same patch sizes.

4.3.2. Pre-training stage: Optimizing ratio of random masking

In self-supervised learning, the masking ratio controls prediction difficulty and promotes learning of high-level representations beyond

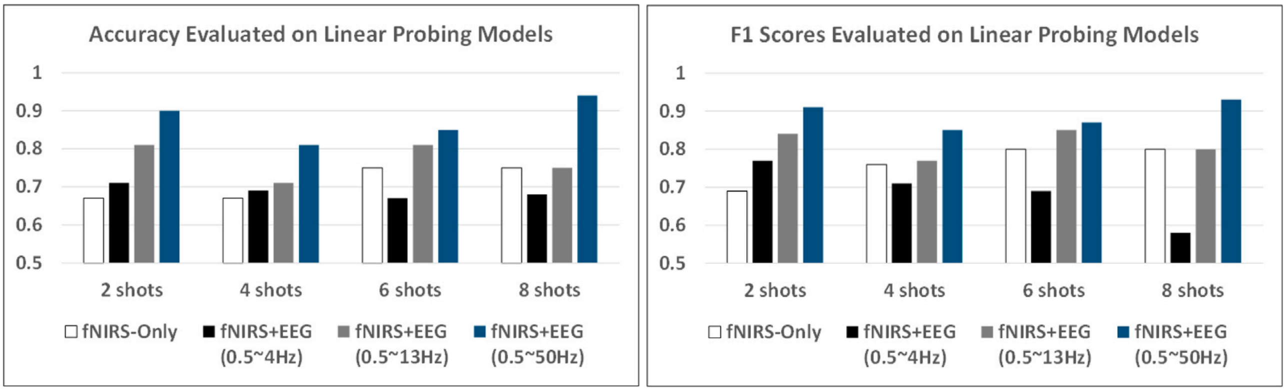

Fig. 5. Measured classification performance based on the amount of shared domain.

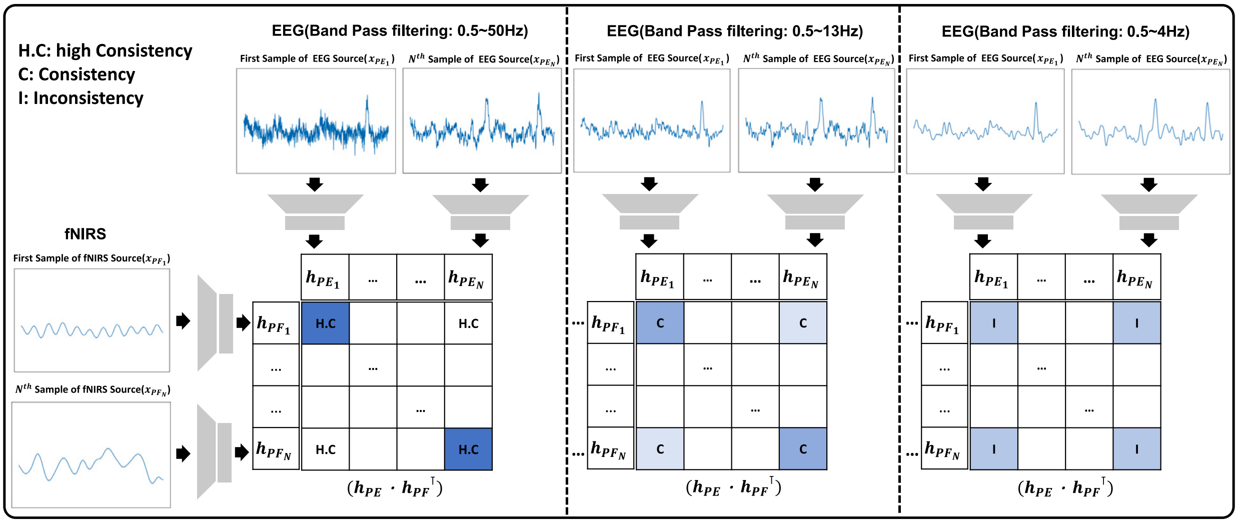
H.C: high Consistency
C: Consistency
I: Inconsistency
EEG(Band Pass filtering: 0.5~50Hz)
EEG(Band Pass filtering: 0.5~13Hz)
EEG(Band Pass filtering: 0.5~4Hz)
First Sample of EEG Source(x_{PE_1})
N^{th} Sample of EEG Source(x_{PE_N})
fNIRS
First Sample of fNIRS Source(x_{PF_1})
N^{th} Sample of fNIRS Source(x_{PF_N})
h_{PE_1}
h_{PE_N}
h_{PF_1}
h_{PF_N}
H.C
C
I
$(h_{PE} \cdot h_{PF}^{T})$

Fig. 6. The impact of shared domain on consistent contrastive learning. When the shared domain is abundant (left), embeddings are clearly separated (blue/white) due to consistent learning. In contrast, when the shared domain is limited (right), inconsistent learning results in ambiguous separation between embeddings (sky blue).

Table 6
Classification accuracy depending on the patch size. Performance is evaluated based on the number of shots used for training. Red text indicates the best score in each shot.

EEG-only dataset (Alertness/Sleep)					fNIRS-only dataset (Mental arithmetic/Resting state)				
Patch sizes	1shot	2shots	4shots	8shots	Patch sizes	2shots	4shots	6shots	8shots
EFRM (EEG domain only) (0.25 s)	0.59 ± 0.14	0.60 ± 0.14	0.75 ± 0.12	0.70 ± 0.12	EFRM (fNIRS domain only) (0.5 s)	0.62 ± 0.12	0.67 ± 0.15	0.69 ± 0.16	0.71 ± 0.23
EFRM (EEG domain only) (0.5 s)	0.47 ± 0.12	0.51 ± 0.11	0.75 ± 0.11	0.68 ± 0.10	EFRM (fNIRS domain only) (1.0 s)	0.62 ± 0.18	0.67 ± 0.16	0.75 ± 0.10	0.75 ± 0.10
EFRM (EEG domain only) (1.0 s)	0.51 ± 0.11	0.50 ± 0.11	0.72 ± 0.10	0.70 ± 0.10	EFRM (fNIRS domain only) (2.0 s)	0.67 ± 0.26	0.67 ± 0.12	0.75 ± 0.10	0.75 ± 0.10
EFRM (proposed) (0.25 s)	0.60 ± 0.14	0.61 ± 0.12	0.73 ± 0.11	0.74 ± 0.11	EFRM (proposed) (2.0 s)	0.90 ± 0.07	0.81 ± 0.12	0.85 ± 0.11	0.94 ± 0.07

low-level patterns. In this experiment, we investigated the optimal random masking ratio for the proposed method during the pre-training stage. As shown in Fig. 7, we trained five pre-trained models with masking ratios ranging from 0.1 to 0.9 and evaluated their classification performance on downstream datasets. The results show that a 50% masking ratio consistently yielded the highest performance across most datasets. Notably, this optimal ratio is lower than the commonly used 75% in image datasets [20], because brain signals are sparser and less redundant than images, making high masking ratios hinder accurate reconstruction. Table 7 presents the amount of computational cost, measured in multiply-accumulate operations (MACs), according to different masking ratios. The results demonstrate that higher masking ratios lead to reduced computational cost. Although Fig. 7 shows comparable performance between the 30% and 50% masking ratios in some cases, the 50% ratio is identified as optimal when both accuracy and computational cost are considered.

4.3.3. Transfer learning stage: Evaluating fine-tuning performance

In this experiment, we also evaluate downstream task performance using fine-tuning as an alternative transfer learning approach. Whereas linear probing freezes layers, fine-tuning enables domain-specific optimization by retraining all layers for the target task but carries a higher risk of overfitting. To mitigate this issue, we trained the model using approximately 20% of the iterations in comparison with linear probing. Despite the risk of overfitting, Fig. 8 demonstrates that the proposed method consistently achieves the highest performance in most evaluations, aligning with the results obtained through linear probing in Fig. 3.

4.3.4. Real-world applicability in brain–computer interfaces

We assess real-world applicability along two dimensions: (i) calibration time and (ii) task performance. We define calibration time as

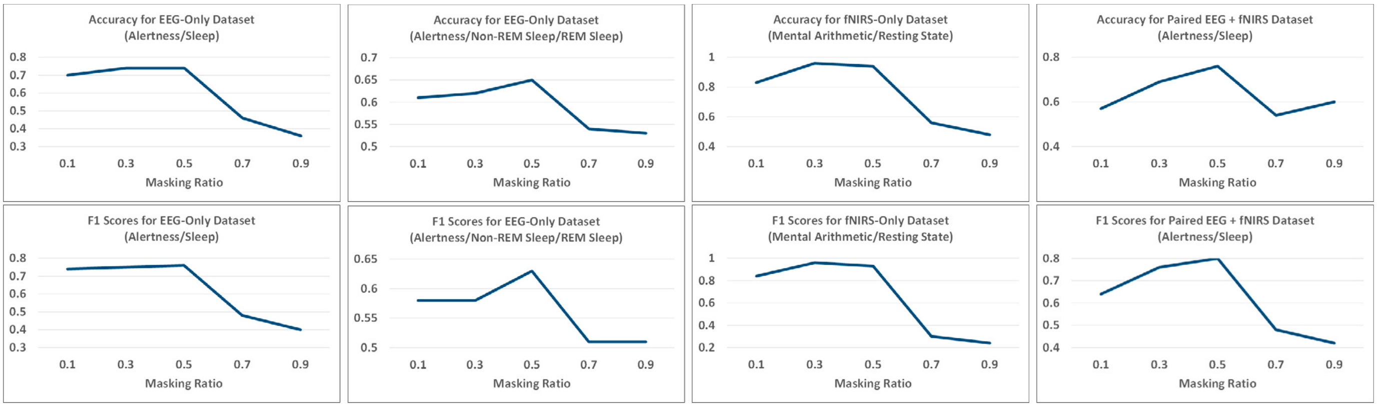

Fig. 7. Classification accuracy and F1 scores for each downstream task under varying random masking ratios. Performance is evaluated across masking ratios ranging from 10% to 90%. The first row shows the classification accuracy for each downstream dataset, while the second row presents the corresponding F1 scores.

Table 7
Computational costs, measured by multiply-accumulate operations (MACs), depends on the masking ratio.

	Masking ratio = 0.1	Masking ratio = 0.3	Masking ratio = 0.5	Masking ratio = 0.7	Masking ratio = 0.9
MAC (×10^9)	960	873	785	697	610

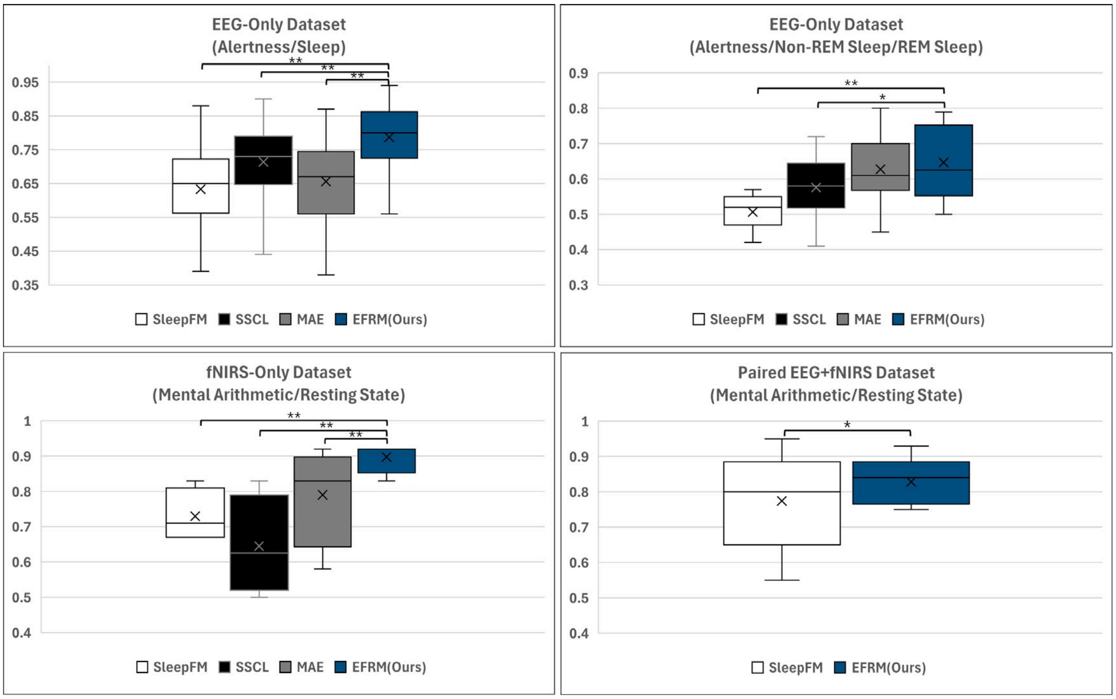

Fig. 8. Boxplots representing measured accuracy of the proposed and baseline methods on multiple downstream task datasets. The top, middle, and bottom lines of each box indicate the upper quartile, median, and lower quartile scores, respectively, with whiskers extending to the highest and lowest scores. The × symbol represents the average score. * indicates a significant difference ($p < 0.05$), and ** indicates a highly significant difference ($p < 0.01$). The proposed and baseline methods were optimized based on the maximum number of shot for each dataset (as shown in Table 3).

the end-to-end pre-use procedure required to adapt the model to user-specific data and reach a target accuracy. In conventional supervised pipelines, calibration typically takes 15–30 min and requires 40–80 trials per class to train a BCI, which limits practical deployment [52]. In our experiments, the proposed EFRM attains stable accuracy with few-shot fine-tuning ($K = 4$–8), achieving binary-classification accuracy comparable to state-of-the-art supervised methods while using 10% as many trials (4–8 vs. 40–80 per class). Multimodal pre-training on EEG and fNIRS mitigates single-modality limitations and yields consistent gains on downstream tasks. Consistent with our results, recent multimodal work shows that aligning EEG with image embeddings improves EEG decoding and classification [53]. Collectively, these properties leverage prior knowledge to reduce user-onboarding time and integrate multimodal data to enhance detection and signal-analysis performance, demonstrating tangible utility across real-world applications such as BCI control and neurorehabilitation.

5. Conclusion

We proposed the first multimodal representation learning model integrating EEG and fNIRS to overcome the modality limitations of existing brain signal representation models. The proposed method utilizes a masked autoencoder-based learning approach to capture modality-specific domain features and contrastive learning to extract shared domain representations. We pre-trained the model using publicly available EEG and fNIRS datasets, enabling its application to both single- and multimodal downstream tasks. Our method enhances generalizability in comparison with traditional supervised approaches that require paired EEG–fNIRS data. We quantitatively evaluated the proposed method, demonstrating higher classification performance than existing self-supervised learning methods and reduced reliance on labeled samples compared to state-of-the-art supervised learning approaches. Additionally, we confirmed that shared information between EEG and fNIRS enables consistent contrastive learning, resulting in a higher-performing multimodal representation model compared to single-modality approaches. Nonetheless, the proposed method exhibits certain limitations. First, there is an imbalance in the number of available public EEG and fNIRS datasets used for pre-training, with the number of EEG recordings outnumbers that of fNIRS by a factor of approximately seven, potentially affecting encoder performance. Second, many of the datasets used in downstream tasks are sleep-related and primarily based on binary or ternary classification, which may limit the generalizability of the model. For future work, we will expand the dataset by acquiring additional EEG and fNIRS recordings to balance the modalities and improve pre-training performance. Furthermore, we will validate the model's applicability across diverse downstream tasks, including multi-class classification.

CRediT authorship contribution statement

Euijin Jung: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jinung An:** Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Ethics statement

Simultaneous EEG and fNIRS data were collected from 11 participants to evaluate the proposed method. The study protocol was reviewed and approved by the Institutional Review Board (IRB) of the Daegu Gyeongbuk Institute of Science and Technology (DGIST) (Approval No. DGIST-171011-HR-035-01). All participants were fully informed of the study's purpose and procedures and provided written informed consent prior to participation. All experimental procedures complied with the ethical guidelines established by the DGIST IRB. For the other datasets used in this study, no additional IRB approval was required, as it comprised fully anonymized, publicly available data.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2023R1A2C2006752), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00302281), and the DGIST R&D Program of the Ministry of Science and ICT (25-IT-02).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.111292.

References

[1] R. Supakar, P. Satvaya, P. Chakrabarti, A deep learning based model using RNN-LSTM for the detection of schizophrenia from EEG data, Comput. Biol. Med. 151 (2022) 106225.

[2] X. Zhang, W. Kou, I. Eric, C. Chang, H. Gao, Y. Fan, Y. Xu, Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device, Comput. Biol. Med. 103 (2018) 71–81.

[3] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwoh, X. Li, C. Guan, An attention-based deep learning approach for sleep stage classification with single-channel EEG, IEEE Trans. Neural Syst. Rehabil. Eng. 29 (2021) 809–818.

[4] M. Diykh, Y. Li, Complex networks approach for EEG signal sleep stages classification, Expert Syst. Appl. 63 (2016) 241–248.

[5] D. Kostas, S. Aroca-Ouellette, F. Rudzicz, BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data, Front. Hum. Neurosci. 15 (2021) 653659.

[6] D. Pulver, P. Angkan, P. Hungler, A. Etemad, EEG-based cognitive load classification using feature masked autoencoding and emotion transfer learning, in: Proceedings of the 25th International Conference on Multimodal Interaction, 2023, pp. 190–197.

[7] X. Jiang, J. Zhao, B. Du, Z. Yuan, Self-supervised contrastive learning for EEG-based sleep staging, in: 2021 International Joint Conference on Neural Networks, IJCNN, IEEE, 2021, pp. 1–8.

[8] R. Thapa, B. He, M.R. Kjaer, H.M. Iv, G. Ganjoo, E. Mignot, J. Zou, SleepFM: Multi-modal representation learning for sleep across brain activity, ECG and respiratory signals, in: Proceedings of the 41st International Conference on Machine Learning, Vol. 235, 2024, pp. 48019–48037.

[9] H.-Y.S. Chien, H. Goh, C.M. Sandino, J.Y. Cheng, MAEEG: Masked auto-encoder for EEG representation learning, in: NeurIPS 2022 Workshop on Learning from Time Series for Health, 2022.

[10] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, A. Gramfort, Uncovering the structure of clinical EEG signals with self-supervised learning, J. Neural Eng. 18 (4) (2021) 046020.

[11] Z. Liu, J. Shore, M. Wang, F. Yuan, A. Buss, X. Zhao, A systematic review on hybrid EEG/fNIRS in brain-computer interface, Biomed. Signal Process. Control. 68 (2021) 102595.

[12] X.-W. Wang, D. Nie, B.-L. Lu, Emotional state classification from EEG data using machine learning approach, Neurocomputing 129 (2014) 94–106.

[13] X. Hu, S. Yuan, F. Xu, Y. Leng, K. Yuan, Q. Yuan, Scalp EEG classification using deep Bi-LSTM network for seizure detection, Comput. Biol. Med. 124 (2020) 103919.

[14] N. Hakimi, A. Jodeiri, M. Mirbagheri, S.K. Setarehdan, Proposing a convolutional neural network for stress assessment by means of derived heart rate from functional near infrared spectroscopy, Comput. Biol. Med. 121 (2020) 103810.

[15] R. Fernandez Rojas, X. Huang, K.-L. Ou, A machine learning approach for the identification of a biomarker of human pain using fNIRS, Sci. Rep. 9 (1) (2019) 5645.

[16] C. Cooney, R. Folli, D. Coyle, A bimodal deep learning architecture for EEG-fNIRS decoding of overt and imagined speech, IEEE Trans. Biomed. Eng. 69 (6) (2021) 1983–1994.

[17] Y. Li, X. Zhang, D. Ming, Early-stage fusion of EEG and fNIRS improves classification of motor imagery, Front. Neurosci. 16 (2023) 1062889.

[18] A. Arif, Y. Wang, R. Yin, X. Zhang, A. Helmy, EF-Net: Mental state recognition by analyzing multimodal EEG-fNIRS via CNN, Sensors 24 (6) (2024) 1889.

[19] M. Liu, B. Yang, L. Meng, Y. Zhang, S. Gao, P. Zan, X. Xia, STA-Net: Spatial–temporal alignment network for hybrid EEG-fNIRS decoding, Inf. Fusion (2025) 103023.

[20] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.

[21] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PmLR, 2021, pp. 8748–8763.

[22] Y. Song, Q. Zheng, B. Liu, X. Gao, EEG conformer: Convolutional transformer for EEG decoding and visualization, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2022) 710–719.

[23] Z. Wang, J. Zhang, X. Zhang, P. Chen, B. Wang, Transformer model for functional near-infrared spectroscopy classification, IEEE J. Biomed. Health Inform. 26 (6) (2022) 2559–2569.

[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[25] S.-w. Yang, H.-J. Chang, Z. Huang, A.T. Liu, C.-I. Lai, H. Wu, J. Shi, X. Chang, H.-S. Tsai, W.-C. Huang, T.-h. Feng, P.-H. Chi, Y.Y. Lin, Y.-S. Chuang, T.-H. Huang, W.-C. Tseng, K. Lakhotia, S.-W. Li, A. Mohamed, S. Watanabe, H.-y. Lee, A Large-Scale evaluation of speech foundation models, IEEE/ACM Trans. Audio Speech Lang. Process. 32 (2024) 2884–2899, http://dx.doi.org/10.1109/TASLP.2024.3389631.

[26] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K.V. Alwala, A. Joulin, I. Misra, Imagebind: One embedding space to bind them all, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15180–15190.

[27] A. Vaid, J. Jiang, A. Sawant, S. Lerakis, E. Argulian, Y. Ahuja, J. Lampert, A. Charney, H. Greenspan, J. Narula, et al., A foundational vision transformer improves diagnostic performance for electrocardiograms, NPJ Digit. Med. 6 (1) (2023) 108.

[28] B. Yu, L. Cao, J. Jia, C. Fan, Y. Dong, C. Zhu, E-FNet: A EEG-fNIRS dual-stream model for Brain–Computer interfaces, Biomed. Signal Process. Control. 100 (2025) 106943.

[29] Q. He, L. Feng, G. Jiang, P. Xie, Multimodal multitask neural network for motor imagery classification with EEG and fNIRS signals, IEEE Sens. J. 22 (21) (2022) 20695–20706.

[30] Y. Kwak, W.-J. Song, S.-E. Kim, FGANet: fNIRS-guided attention network for hybrid EEG-fNIRS brain-computer interfaces, IEEE Trans. Neural Syst. Rehabil. Eng. 30 (2022) 329–339.

[31] S.D. Mayhew, S.G. Dirckx, R.K. Niazy, G.D. Iannetti, R.G. Wise, EEG signatures of auditory activity correlate with simultaneously recorded fMRI responses in humans, Neuroimage 49 (1) (2010) 849–864.

[32] J. Lin, J. Lu, Z. Shu, N. Yu, J. Han, An EEG-fNIRS neurovascular coupling analysis method to investigate cognitive-motor interference, Comput. Biol. Med. 160 (2023) 106968.

[33] R. Blanco, M.G. Preti, C. Koba, D.V.D. Ville, A. Crimi, Comparing structure–function relationships in brain networks using EEG and fNIRS, Sci. Rep. 14 (1) (2024) 28976.

[34] L.-C. Chen, P. Sandmann, J.D. Thorne, C.S. Herrmann, S. Debener, Association of concurrent fNIRS and EEG signatures in response to auditory and visual stimuli, Brain Topogr. 28 (2015) 710–725.

[35] Y. Li, Y. Wang, B. Lei, S. Wang, Scdm: Unified representation learning for eeg-to-fnirs cross-modal generation in mi-bcis, IEEE Trans. Med. Imaging (2025).

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

[37] I. Obeid, J. Picone, The temple university hospital EEG data corpus, Front. Neurosci. 10 (2016) 196.

[38] G. Schalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, J.R. Wolpaw, BCI2000: a general-purpose brain-computer interface (BCI) system, IEEE Trans. Biomed. Eng. 51 (6) (2004) 1034–1043.

[39] W.-L. Zheng, B.-L. Lu, Investigating critical frequency bands and channels for EEG-Based emotion recognition with deep neural networks, IEEE Trans. Auton. Ment. Dev. 7 (3) (2015) 162–175, http://dx.doi.org/10.1109/TAMD.2015.2431497.

[40] R.-N. Duan, J.-Y. Zhu, B.-L. Lu, Differential entropy feature for EEG-based emotion classification, in: 6th International IEEE/EMBS Conference on Neural Engineering, NER, IEEE, 2013, pp. 81–84.

[41] J. Defenderfer, A. Kerr-German, M. Hedrick, A.T. Buss, Investigating the role of temporal lobe activation in speech perception accuracy with normal hearing adults: An event-related fNIRS study, Neuropsychologia 106 (2017) 31–41.

[42] J. San Juan, X.-S. Hu, M. Issa, S. Bisconti, I. Kovelman, P. Kileny, G. Basura, Tinnitus alters resting state functional connectivity (RSFC) in human auditory and non-auditory brain regions as measured by functional near-infrared spectroscopy (fNIRS), PLoS One 12 (6) (2017) e0179150.

[43] J. Hudak, D. Rosenbaum, B. Barth, A.J. Fallgatter, A.-C. Ehlis, Functionally disconnected: a look at how study design influences neurofeedback data and mechanisms in attention-deficit/hyperactivity disorder, PLoS One 13 (8) (2018) e0200931.

[44] S. Bak, J. Park, J. Shin, J. Jeong, Open-access fNIRS dataset for classification of unilateral finger-and foot-tapping, Electronics 8 (12) (2019) 1486.

[45] X. Li, M.A. Krol, S. Jahani, D.A. Boas, H. Tager-Flusberg, M.A. Yücel, Brain correlates of motor complexity during observed and executed actions, Sci. Rep. 10 (1) (2020) 10965.

[46] J. Shin, A. von Lühmann, B. Blankertz, D.-W. Kim, J. Jeong, H.-J. Hwang, K.-R. Müller, Open access dataset for EEG+ NIRS single-trial classification, IEEE Trans. Neural Syst. Rehabil. Eng. 25 (10) (2016) 1735–1745.

[47] B. Blankertz, M. Tangermann, C. Vidaurre, S. Fazli, C. Sannelli, S. Haufe, C. Maeder, L.E. Ramsey, I. Sturm, G. Curio, et al., The Berlin brain–computer interface: non-medical uses of BCI technology, Front. Neurosci. 4 (2010) 2050.

[48] B. Kemp, A.H. Zwinderman, B. Tuk, H.A. Kamphuisen, J.J. Oberye, Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG, IEEE Trans. Biomed. Eng. 47 (9) (2000) 1185–1194.

[49] G. Bauernfeind, R. Scherer, G. Pfurtscheller, C. Neuper, Single-trial classification of antagonistic oxyhemoglobin responses during mental arithmetic, Med. Biol. Eng. Comput. 49 (2011) 979–984.

[50] C. Lee, J. An, LSTM-CNN model of drowsiness detection from multiple consciousness states acquired by EEG, Expert Syst. Appl. 213 (2023) 119032.

[51] E.T. Attar, Review of electroencephalography signals approaches for mental stress assessment, Neurosci. J. 27 (4) (2022) 209–215.

[52] F. Lotte, Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces, Proc. IEEE 103 (6) (2015) 871–890.

[53] E. Shi, H. Hu, Q. Yuan, K. Zhao, S. Yu, S. Zhang, BrainAlign: EEG-Vision alignment via Frequency-Aware temporal encoder and differentiable cluster assigner, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2025, pp. 98–108.