Rethinking Delayed Hemodynamic Responses for fNIRS Classification

Zenghui Wang, Jihong Fang, and Jun Zhang

***Abstract*— Functional near-infrared spectroscopy (fNIRS) is a non-invasive neuroimaging technology for monitoring cerebral hemodynamic responses. Enhancing fNIRS classification can improve the performance of brain–computer interfaces (BCIs). Currently, deep neural networks (DNNs) do not consider the inherent delayed hemodynamic responses of fNIRS signals, which causes many optimization and application problems. Considering the kernel size and receptive field of convolutions, delayed hemodynamic responses as domain knowledge are introduced into fNIRS classification, and a concise and efficient model named fNIRSNet is proposed. We empirically summarize three design guidelines for fNIRSNet. In subject-specific and subject-independent experiments, fNIRSNet outperforms other DNNs on open-access datasets. Specifically, fNIRSNet with only 498 parameters is 6.58% higher than convolutional neural network (CNN) with millions of parameters on mental arithmetic tasks and the floating-point operations (FLOPs) of fNIRSNet are much lower than CNN. Therefore, fNIRSNet is friendly to practical applications and reduces the hardware cost of BCI systems. It may inspire more research on knowledge-driven models for fNIRS BCIs. Code is available at https://github.com/wzhlearning/fNIRSNet.**

***Index Terms*— Functional near-infrared spectroscopy (fNIRS), brain–computer interface (BCI), deep neural network (DNN), delayed hemodynamic response, domain knowledge.**

I. Introduction

FUNCTIONAL near-infrared spectroscopy (fNIRS) is a non-invasive neuroimaging technology that records changes in the concentration of oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) by measuring the absorption of near-infrared light between 650 and 950 nm [1]. Brain–computer interfaces (BCIs) decode signals from patients suffering from movement disorders to establish non-muscle communication with the external environment [2]. Owing to its non-invasiveness, user-friendliness, and portability [3], fNIRS has attracted attention in the BCI community.

Methods of classifying fNIRS signals include traditional machine learning and emerging deep learning. Statistical values (mean, variance, peak, kurtosis, skewness, and slope) are extracted from fNIRS signals to train support vector machine (SVM), linear discriminant analysis (LDA), and k-nearest neighbor (KNN) [4], [5]. Vector-based phase analysis including change in cerebral blood volume (ΔCBV), change in cerebral oxygen exchange (ΔCOE), vector magnitude, and angle is also commonly used to train these classifiers [6], [7]. However, traditional classifiers rely heavily on manual feature engineering and prior knowledge. In recent years, deep learning has become the mainstream of fNIRS classification research. Convolutional neural networks (CNNs), long short-term memory (LSTM), and Transformers have been developed for fNIRS classification [8], [9], [10], [11]. Deep learning is notoriously data-hungry, but limited fNIRS data severely hinders its applications. Unfortunately, the scarcity of fNIRS data is difficult to address in a short time. The high cost of fNIRS equipment may limit the acquisition scale and the burdensome signal acquisition procedures may limit the number of participants. Although some complicated models have been developed, the insufficiency of fNIRS data still limits the improvement of classification performance. More importantly, the domain knowledge of fNIRS signals is not exploited. The changes in HbO and HbR are a slow metabolic process manifested as delayed hemodynamic responses which are also inherent properties of fNIRS signals. Hence, the number of sampling points per unit time is less than high temporal resolution signals such as electroencephalogram (EEG) [12]. The delayed hemodynamic responses occur in both onset and cessation of neuronal activity [13], [14], [15]. Nambu et al. [16] found a 4 s hemodynamic delay when measuring human motor-cortical activation. Shin et al. [4] found that fNIRS classification accuracies reach the maximum after a delay of several seconds. HbO and HbR do not change significantly in the first few seconds of experimental stimulation, while they still have solid hemodynamic responses when the stimulation is over.

Unlike computer vision and natural language processing supported by large-scale data, some general design principles may not be suitable for the fNIRS field, such as deeper architectures and small convolutions. In order to improve classification performance, researchers tend to design more

Manuscript received 1 June 2023; revised 15 October 2023; accepted 31 October 2023. Date of publication 7 November 2023; date of current version 21 November 2023. This work was supported by the National Natural Science Foundation of China under Grant 61872004. *(Corresponding authors: Jun Zhang; Jihong Fang.)*

Zenghui Wang is with the School of Electrical Engineering and Automation, Anhui University, Hefei 230601, China (e-mail: scholarzhwang@163.com).

Jihong Fang is with the Nursing Department, Anhui Provincial Children's Hospital, Hefei 230051, China (e-mail: fangjihong510@126.com).

Jun Zhang is with the School of Artificial Intelligence, Anhui University, Hefei 230601, China (e-mail: junzhang@ahu.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3330911

complex network architectures by increasing the number of kernels and network depth. However, these operations may lead to over-parametrization and overfitting on limited fNIRS data. Finally, researchers have to adopt more regularization methods to solve these tricky problems, such as dropout [17] and flooding [18]. In addition, some studies [19], [20], [21] use small convolutional kernels, e.g., 3 × 3 and 4 × 4, to extract fNIRS signal features. He et al. [22] used smaller kernel sizes of 2 × 1 and 1 × 4 to extract temporal and spatial features, respectively. One-dimensional (1D) CNNs are also popular for processing fNIRS signals and their kernel size is usually set at least three [23], [24]. The biggest issue is that fNIRS signals are fed directly into deep neural networks (DNNs) without considering domain knowledge. A small kernel with a limited receptive field is challenging to extract the features of delayed hemodynamic responses because there is no significant change in HbO and HbR in small neighborhoods. However, stacking more convolutional layers to obtain larger receptive fields may cause overfitting on limited fNIRS data. Therefore, we rethink delayed hemodynamic responses and systematically explore a simple but efficient design philosophy for a deep learning-based fNIRS classification model.

In this study, two core ideas are presented: 1) delayed hemodynamic responses as domain knowledge should be introduced into fNIRS classification models; 2) a simple and efficient model is beneficial to practical applications on limited fNIRS data. We propose a compact fNIRS classification network named fNIRSNet which consists of three convolutional layers and one fully connected (FC) layer without pooling, dropout, and other complicated structures. Three design guidelines are empirically summarized for fNIRSNet: 1) *the size of convolutional kernels is critical for extracting features of delayed hemodynamic responses and decoupling network depth and receptive fields*; 2) *concatenating standard convolutions and depthwise separable convolutions can balance the stability, speed, and efficiency of fNIRSNet*; 3) *activation functions with saturated negative values can alleviate information loss in the first layer*. fNIRSNet achieves superior performance on open-access datasets, while it has extremely few parameters and computational consumption. To the best of our knowledge, fNIRSNet is the least resource-consuming deep learning-based fNIRS classification model. Our study illustrates that a compact model infused with domain knowledge outperforms big models in the fNIRS field. These advantages make fNIRSNet more valuable for applications on mobile and embedded devices. Code is available at https://github.com/wzhlearning/fNIRSNet.

The rest of this article is organized as follows. Section II describes the design ideas of fNIRSNet. Section III introduces open-access datasets, signal preprocessing, and evaluation protocols. In Section IV, comprehensive experiments demonstrate the superiority of fNIRSNet. Discussion is provided in Section V. Finally, Section VI concludes this article.

II. Methods

A. Hemodynamic Response

Neurovascular coupling that links changes in neural activity to the cerebral blood flow (CBF) is the cornerstone of many functional neuroimaging techniques based on hemodynamic responses [25], such as functional magnetic resonance imaging (fMRI) [26] and fNIRS. fMRI can measure blood oxygenation level dependent (BOLD) signals that are modeled as a convolution of the hemodynamic response function and the stimulus function. The hemodynamic response function can be generated by three gamma functions $\Gamma(\cdot)$ [27]:

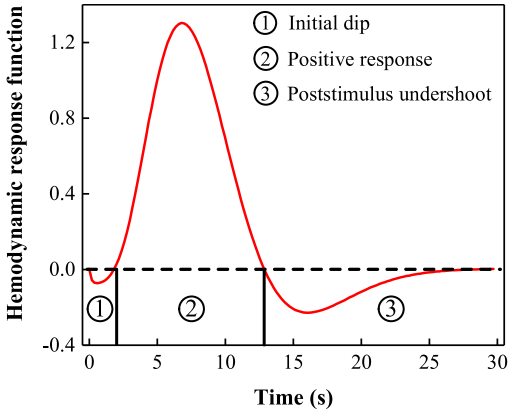


Fig. 1. A canonical hemodynamic response function generated by three gamma functions.

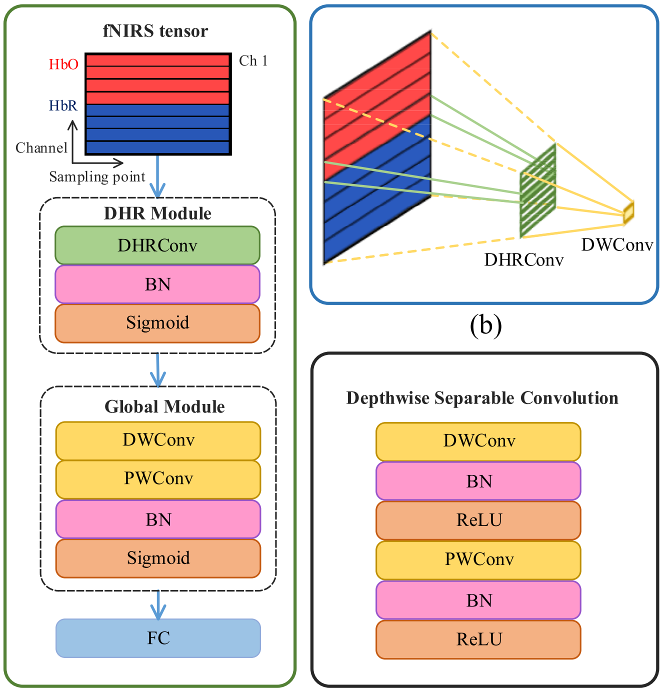
$$h(t) = \sum_{i=1}^{3} \left(A_i \frac{t^{\alpha_i - 1} \beta_i^{\alpha_i} \mathrm{e}^{-\beta_i t}}{\Gamma(\alpha_i)} \right), \tag{1}$$

where A, α, and β control the height and direction, shape, and scale of hemodynamic responses, respectively. Fig. 1 illustrates the canonical hemodynamic response function. It is divided into three phases: initial dip, positive response, and poststimulus undershoot [28]. In the fNIRS field, the initial dip manifests an initial increase/decrease in HbR/HbO, which is associated with neural activity consuming oxygen in nearby local regions. The positive/negative response for HbO/HbR is caused by a large increase in CBF, which is usually manifested as an increase in HbO and a decrease in HbR. The poststimulus period is characterized by an undershoot of HbO and an overshoot of HbR, and the period typically starts between 10 and 20 s after stimulus cessation and lasts up to 60 s [29]. The main reasons for poststimulus undershoot are the continuous increase in the metabolic rate of oxygen and delayed vascular compliance [30].

The delayed hemodynamic response is an inherent property and a major limitation of fNIRS signals. Limited by local receptive fields, small convolutions are challenging to model the long-term dependency on hemodynamic responses. Therefore, we hypothesize that convolutions with fNIRS channel-level receptive fields can extract delayed response features and convolutions with global receptive fields can explore activation patterns of different brain regions.

B. fNIRSNet

1) Notation: The fNIRS tensor is defined to facilitate the following description. In Fig. 2(a), HbO and HbR are arranged to form an fNIRS tensor $X \in \mathbb{R}^{C \times S \times D}$, where C is twice (two chromophores: HbO and HbR) the number of fNIRS channels,



(a) (b) (c)

Fig. 2. (a) Overview architecture of fNIRSNet. (b) Schema of receptive fields. The green and yellow feature maps are the output of DHRConv and DWConv, respectively. The solid lines indicate that the input is directly obtained from the previous layer and the dotted lines indicate the corresponding receptive fields indirectly in the fNIRS tensor. (c) Depthwise separable convolution.

$S = f \times T$ is the number of sampling points, f is the sampling frequency, T is the sampling time, and the depth D is 1.

2) Overview: The architecture of fNIRSNet is illustrated in Fig. 2(a). fNIRSNet consists of a delayed hemodynamic response module (DHR Module) and a global module. Specifically, fNIRSNet contains three convolutional layers with different kernel sizes and a fully connected (FC) layer. Delayed hemodynamic response convolutions (DHRConv) extract the channel-level features of delayed hemodynamic responses. Depthwise separable convolutions consisting of depthwise convolutions (DWConv) and pointwise convolutions (PWConv) are used to reduce model parameters [31]. Batch normalization (BN) accelerates network training and improves classification performance [32]. The sigmoid activation function is used for nonlinear activation and alleviates information loss in the first layer. The feature maps are flattened to 1D vectors and then fed into an FC layer. Finally, a softmax function calculates the conditional probabilities of the K classes. The proposed fNIRSNet does not have pooling, dropout, and other complicated structures. Overall, fNIRSNet is concise and efficient, and comprehensive experiments demonstrate its superiority. Three design guidelines are empirically summarized.

3) Guideline 1: The size of convolutional kernels is critical for extracting features of delayed hemodynamic responses and decoupling network depth and receptive fields. Table I shows the configurations of fNIRSNet. For an fNIRS tensor $X \in \mathbb{R}^{C\times S\times 1}$, the kernel size of DHRConv is $1 \times S$, which means that the width of this kernel equals the number of input sampling points. DHRConv with channel-level receptive fields can directly extract single-channel hemodynamic response features. The kernel size of DWConv is $C \times 1$, which indicates that the height is twice (two chromophores) the number of fNIRS channels. In general, model designers stack many convolutional layers to get larger receptive fields in deeper layers, but this stacking operation brings more computational cost. The $C \times 1$ DWConv can directly obtain global receptive fields without stacking layers. Fig. 2(b) shows the schema of receptive fields. In addition, DWConv can compensate for spatial information because DHRConv alone cannot aggregate spatial information from multi-channel fNIRS, which helps DWConv focus on activation patterns in different brain regions. Therefore, $1 \times S$ DHRConv and $C \times 1$ DWConv decouple the contradiction between network depth and receptive fields. Finally, the 1×1 PWConv projects the output of DWConv into a new channel space.

TABLE I
Configurations of the Proposed Model

Layer	Input size	Size / Filter	Output size
DHRConv	$C \times S \times 1$	$1 \times S$ / F_1	$C \times 1 \times F_1$
DWConv	$C \times 1 \times F_1$	$C \times 1$ / F_2	$1 \times 1 \times F_2$
PWConv	$1 \times 1 \times F_2$	1×1 / F_2	$1 \times 1 \times F_2$
FC	F_2	–	K

4) Guideline 2: Concatenating standard convolutions and depthwise separable convolutions can balance the stability, speed, and efficiency of fNIRSNet. In Fig. 2(c), a depthwise separable convolution is a factorized convolution that factorizes a standard convolution into DWConv and PWConv [31]. DWConv applies a filter to each input channel, and PWConv projects the output of DWConv into a new channel space. Compared with standard convolutions, depthwise separable convolutions reduce computational cost significantly. The computational cost of standard convolutions is defined as

$$K_h \times K_w \times F_{in} \times F_{out} \times M_h \times M_w, \tag{2}$$

where $K_h \times K_w$ is the kernel size, F_{in} is the number of input channels, F_{out} is the number of output channels, and $M_h \times M_w$ is the size of feature map. Depthwise separable convolutions have the computational cost of:

$$K_h \times K_w \times F_{in} \times M_h \times M_w + F_{in} \times F_{out} \times M_h \times M_w. \tag{3}$$

Note that DHRConv is a standard convolution in the first layer because the depthwise separable convolution performs poorly in low-dimensional space (the first layer) [33]. Applying standard convolutions at the first layer has a trade-off between stability and speed. The computational cost of DHRConv is $S \times F_1 \times C$. The computational cost of the global module using standard convolutions and depthwise separable convolutions is $F_1 \times C \times F_2$ and $F_1 \times (C + F_2)$, respectively. Depthwise separable convolutions reduce the computational complexity of the global module. In addition, increasing the length of input signals only increases the computational cost of DHRConv without affecting the global module. We do not add any pooling layer to reduce the number of features. Since the number of convolutional filters F_2 of DWConv equals the number of flattened neurons, an FC layer without dropout is used for classification. Except for the number of filters F_1 and F_2, fNIRSNet has almost no other hyperparameters. Therefore, fNIRSNet is friendly to BCI devices because it has fewer parameters and computational cost.

5) Guideline 3: Activation functions with saturated negative values can alleviate information loss in the first layer. We found that activation functions with saturated negative values (e.g., sigmoid, hyperbolic tangent (tanh), and exponential linear unit (ELU) [34]) work better than other mainstream activation functions (e.g., ReLU and leaky ReLU (LReLU) [35]). ELU, ReLU, and LReLU alleviate vanishing gradients caused by increasing model depth via the identity for positive values. However, we ignore vanishing gradients because fNIRSNet is a shallow model. Since fNIRSNet has very few trainable parameters, inappropriate activation functions lead to information loss in the first convolutional layer (i.e., DHRConv), which is a real concern for our study. The negative value input to ReLU cannot be activated, which causes backpropagation to fail to update weight parameters, called the dead neuron problem. Although LReLU avoids dead neurons by a small and non-zero gradient, it cannot ensure a noise-robust deactivation state [34]. Sigmoid and tanh are bilateral saturation activation functions, while ELU is a one-sided negative saturation that reduces forward propagated variation and information [34]. In Section IV-C, the ablation experiments validate Guideline 3.

C. Label Smoothing

Label smoothing is commonly used to prevent DNNs from over-confidence by the weighted average of hard targets and uniform distribution over labels [36]. A network predicts the probability of each class label $k \in \{1, \ldots, K\}$:

$$p_k = \frac{\exp(z_k)}{\sum_{i=1}^{K} \exp(z_i)}, \tag{4}$$

where z_i is the logit. The cross-entropy loss function is defined as

$$L(y, p) = -\sum_{k=1}^{K} y_k \log(p_k), \tag{5}$$

where $y_k = 1$ for the ground truth, and $y_k = 0$ for the rest. Label smoothing is defined as

$$q_k^{LS} = (1 - \varepsilon) y_k + \varepsilon u_k, \tag{6}$$

where ε is the smoothing parameter and is set to 0.1 by default, and $u_k = 1/K$ is the uniform distribution. Finally, the cross-entropy loss function with label smoothing is written as

$$L\left(q^{LS}, p\right) = -\sum_{k=1}^{K} q_k^{LS} \log(p_k). \tag{7}$$

III. Experiments

A. Open-Access Datasets

Extensive experiments are conducted on open-access datasets, including mental arithmetic (MA[1]) and unilateral finger- and foot-tapping (UFFT[2]). The experimental paradigms and sensor placement are shown in Fig. 3.

1) MA: It consists of 29 healthy subjects (14 males, average age 28.5 ± 3.7 years) [4]. For the MA task, the subjects were instructed to perform subtraction such as "three-digit number minus one-digit number" according to the screen and short beep instructions. For the baseline (BL) task, they were instructed to relax by gazing at a black fixation cross on the screen. Each subject was asked to perform 30 trials for each task. This is a hybrid EEG-fNIRS dataset, but only the fNIRS signals are used for our experiments.

2) UFFT: The dataset contains fNIRS signals of 30 subjects (17 males, 23.4 ± 2.5 years old) for ternary classification tasks [5]. During the task period, the subjects were required to randomly perform three types of overt movements according to instructions on the screen, including right-hand finger-tapping (RHT), left-hand finger-tapping (LHT), and foot-tapping (FT). Each movement was performed randomly for 25 trials. They were instructed to relax during the rest period.

The MA dataset contains MA and BL categories, and the UFFT dataset includes RHT, LHT, and FT categories.

B. Signal Preprocessing

Following the original studies [4], [5], the fNIRS signals of MA and UFFT are downsampled to 10 Hz and 13.3 Hz, respectively. Signal preprocessing usually include the modified Beer–Lambert law [37], filtering, segmentation, and baseline correction. The modified Beer–Lambert law converts optical density ΔOD into concentration changes of HbO and HbR from the absorption of near-infrared light. At time t, it is described as

$$\begin{bmatrix} \Delta HbO \\ \Delta HbR \end{bmatrix} = \frac{\begin{bmatrix} \varepsilon_{HbO}(\lambda_1) & \varepsilon_{HbR}(\lambda_1) \\ \varepsilon_{HbO}(\lambda_2) & \varepsilon_{HbR}(\lambda_2) \end{bmatrix}^{-1} \begin{bmatrix} \Delta OD(t, \lambda_1) \\ \Delta OD(t, \lambda_2) \end{bmatrix}}{d \times l}, \tag{8}$$

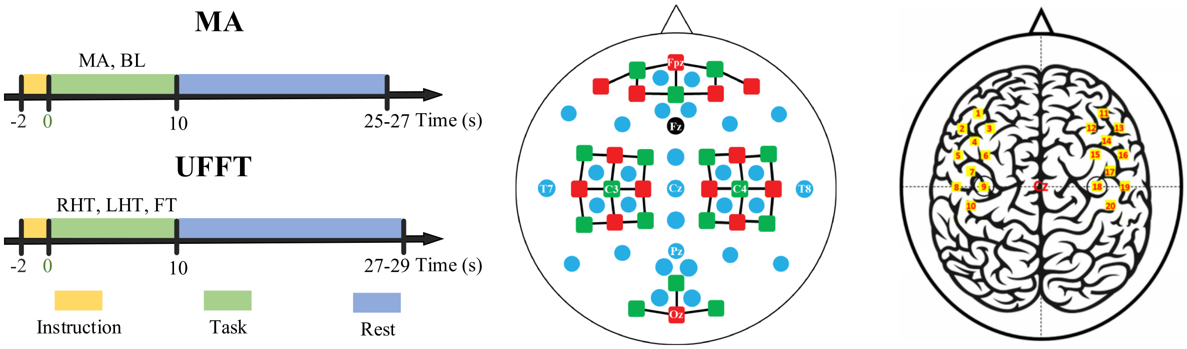
where $\varepsilon_{HbO}(\cdot)$ and $\varepsilon_{HbR}(\cdot)$ are extinction coefficients of HbO and HbR at wavelength λ, d is the differential path-length factor, and l is the distance between source and detector. Raw fNIRS signals contain instrument noise, physiological noise, and motion artifacts [3]. As a result, a band-pass filter with a passband of 0.01–0.1 Hz is used for MA and UFFT. Baseline correction solves the baseline drift problem by subtracting the average value of a reference interval from fNIRS signals. The reference intervals for MA and UFFT are $[-5, -2]$ s and $[-1, 0]$ s, respectively. The fNIRS signals are divided into segments by a sliding window (window size = 3 s, step size = 1 s) [4], [23]. The segmented signal intervals for MA and UFFT are $[-2, 10]$ s and $[0, 10]$ s, respectively. Thus, a trial of MA and UFFT is split into 10 and 8 segments, respectively. Each subject of MA and UFFT includes 600 samples (30 trials × 10 segments × 2 categories and 25 trials × 8 segments × 3 categories). Finally, these segments are normalized by the z-score standardization to accelerate convergence.

C. Evaluation Protocols

Currently, evaluation protocols for fNIRS classification are confusing and some experimental details are inadequately

[1] http://doc.ml.tu-berlin.de/hBCI

[2] https://doi.org/10.6084/m9.figshare.9783755.v1



(a) (b) (c)

Fig. 3. (a) Experimental paradigms for MA and UFFT. A trial consists of an introduction period, a task period, and a rest period. (b) Sensor location layout for MA [4]. The red and green squares are fNIRS sources and detectors, respectively. Solid black lines indicate fNIRS channels. The blue and black (ground) circles are EEG electrodes. (c) fNIRS channel locations for UFFT [5]. Ch 1–10 and Ch 11–20 are located around C3 (Ch 9) and C4 (Ch 18), respectively.

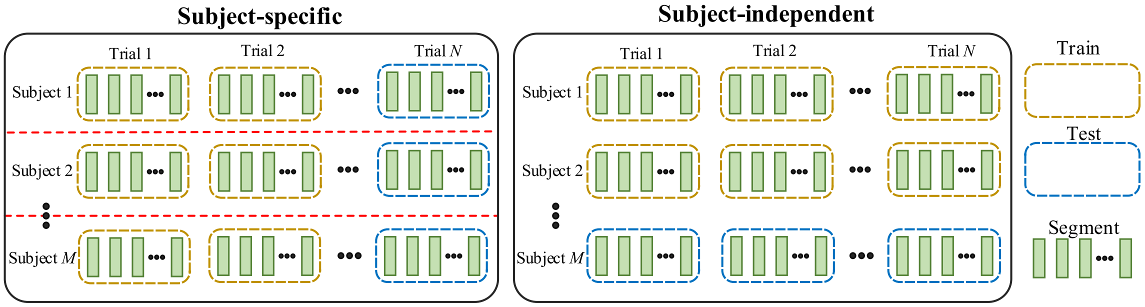


Fig. 4. Schematic diagrams for subject-specific and subject-independent. A dataset contains M subjects and each subject has N trials.

described. In addition, few researchers release their source code for the fNIRS community. It is difficult to reproduce these studies and make fair comparisons. We discuss this issue in Section V. In Fig. 4, we adopt more general and transparent protocols: subject-specific and subject-independent [11], [38].

1) Subject-Specific: DNNs are trained for each subject using a 5-fold cross-validation (KFold-CV) that splits training and test sets according to trials to avoid information leakage. For example, each subject in MA includes 60 trials where 48 trials are used as a training set and 12 trials are used as a test set. Thus, the training set contains 480 samples (48 trials × 10 segments) and the test set includes 120 samples. The final experimental results are the average of all subjects' test sets.

2) Subject-Independent: A leave-one-subject-out cross-validation (LOSO-CV) can rigorously validate inter-individual differences and model generalization. One subject's data is used as the test set and the rest as the training set. The process is repeated until all subject's data has been tested. The reported results are the average of all subjects.

3) Evaluation Metric: Performance metrics include accuracy, precision, recall, F1-score (macro-F1 for UFFT), and Kappa coefficient. F1-score and macro-F1 are defined as

$$\mathrm{F1} = \frac{2 \times P \times R}{P + R}, \tag{9}$$

$$\text{macro-F1} = \frac{2 \times \text{macro-}P \times \text{macro-}R}{\text{macro-}P + \text{macro-}R}, \tag{10}$$

where precision $P = \frac{TP}{TP+FP}$, recall $R = \frac{TP}{TP+FN}$, macro-$P = \frac{1}{n}\sum_{i=1}^{n} P_i$, macro-$R = \frac{1}{n}\sum_{i=1}^{n} R_i$, TP is true positive, FP is false positive, FN is false negative, and n represents the number of pairwise combinations. The final performance metrics are the average of all cross-validation results. Efficiency metrics include model parameters, floating-point operations (FLOPs), inference time, and frames per second (FPS).

D. Experimental Settings

In the subject-specific experiments, F_1 and F_2 of fNIRSNet are 4 and 8, respectively. Considering the increase in training samples, F_1 and F_2 are set to 16 and 32 in the subject-independent experiments, respectively. For the baseline model, the hyperparameters of Transformer-based fNIRS-T[3] [11] are adjusted to fit the size of input fNIRS signals. The kernel sizes of $Conv_S$ and $Conv_C$ are 5 × 10 and 1 × 10, respectively. The Transformer layers of fNIRS-T are set to 4, and the dimension of the linear projection and multi-layer perceptron (MLP) layer is set to 32. Other baseline CNN[4] [8], LSTM[4] [8], and 1D-CNN[5] [23] follow the original references. The CNN contains three convolutional layers, whereas 1D-CNN consists of six 1D convolutional layers, and they both use BN and ReLU. The LSTM has three LSTM layers and each layer has 20 LSTM cells.

All models are optimized by AdamW [39] with an initial learning rate of 0.001. Label smoothing is used to improve

[3] https://github.com/wzhlearning/fNIRS-Transformer

[4] https://github.com/boyanglyu/nback_align

[5] https://github.com/sunzhe839/tensorfusion_EEG_NIRS

TABLE II
EXPERIMENTAL RESULTS FOR SUBJECT-SPECIFIC AND SUBJECT-INDEPENDENT. THE BOLD INDICATES THE BEST RESULT

Experiment	Dataset	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa
Subject-specific	MA	CNN [8]	64.18 ± 13.12	65.37 ± 15.86	61.92 ± 18.92	62.50	0.28
		1D-CNN [23]	64.53 ± 13.03	66.23 ± 14.82	62.63 ± 17.00	63.37	0.29
		LSTM [8]	57.36 ± 11.94	58.21 ± 13.16	57.05 ± 16.25	56.65	0.15
		fNIRS-T [11]	65.17 ± 13.67	66.20 ± 15.15	64.08 ± 17.59	64.28	0.30
		fNIRSNet	**70.76 ± 15.60**	**72.17 ± 17.12**	**69.41 ± 18.74**	**69.95**	**0.42**
	UFFT	CNN [8]	60.57 ± 17.64	61.51 ± 17.68	60.80 ± 17.43	59.77	0.41
		1D-CNN [23]	57.76 ± 17.85	58.46 ± 17.93	57.90 ± 17.75	56.83	0.37
		LSTM [8]	46.14 ± 14.07	46.72 ± 14.07	46.38 ± 14.23	45.37	0.19
		fNIRS-T [11]	61.26 ± 18.54	62.32 ± 18.58	61.41 ± 18.26	60.42	0.42
		fNIRSNet	**68.67 ± 19.85**	**69.97 ± 19.61**	**69.15 ± 19.52**	**68.02**	**0.53**
Subject-independent	MA	CNN [8]	56.05 ± 7.13	45.65 ± 24.17	58.24 ± 33.92	49.82	0.12
		1D-CNN [23]	57.18 ± 6.05	57.23 ± 6.57	60.64 ± 9.34	58.43	0.14
		LSTM [8]	56.01 ± 6.37	56.53 ± 6.67	54.77 ± 9.08	55.28	0.12
		fNIRS-T [11]	56.91 ± 6.14	56.75 ± 6.34	59.92 ± 10.17	57.89	0.14
		fNIRSNet	**65.26 ± 6.65**	**66.14 ± 8.76**	**67.02 ± 10.06**	**65.73**	**0.31**
	UFFT	CNN [8]	57.42 ± 13.36	58.50 ± 13.59	57.42 ± 13.36	56.74	0.36
		1D-CNN [23]	53.36 ± 12.15	54.03 ± 12.03	53.36 ± 12.15	52.83	0.30
		LSTM [8]	57.49 ± 13.43	59.13 ± 12.95	57.49 ± 13.43	56.89	0.36
		fNIRS-T [11]	56.40 ± 13.03	57.30 ± 12.90	56.40 ± 13.03	55.67	0.35
		fNIRSNet	**64.43 ± 15.75**	**66.02 ± 15.67**	**64.43 ± 15.75**	**63.57**	**0.47**

model generalization. For subject-specific, all models are trained with a batch size of 64 for 120 epochs, and the initial learning rate is decayed by a factor of 10 at 60 and 90 epochs. For subject-independent, we apply the cosine learning rate [40] for 30 epochs, and its maximum number of iterations is 30.

IV. Results

A. Comparison With DNNs

Comparison experiments demonstrate that fNIRSNet has excellent advantages. The experimental results (mean ± standard deviation) are shown in Table II. In the subject-specific experiments, fNIRSNet achieves the highest average accuracy on test sets (Wilcoxon signed-rank test, $p < 0.001$). The average accuracy and F1-score of fNIRSNet are 5% higher than fNIRS-T on MA. All metrics of fNIRSNet are significantly higher than the other models (Wilcoxon signed-rank test, $p < 0.001$). The performance of LSTM is lower than other models because individual differences and the scale of fNIRS data prevent LSTM from capturing context information.

The subject-independent experiment can assess model generalization performance because the target subject is not involved in parameter tuning and model training. In Table II, all performance metrics show overall deterioration. However, fNIRSNet still outperforms the other models. Owing to individual differences and limited data, it is important to collect data from target subjects to customize a model. He et al. [22] reported that the accuracy of motor imagery classification decreases by 20% in subject-independent experiments. Although deep models perform better in subject-specific than subject-independent, subject-independent is more suitable for practical applications because it reduces training cost and calibration time significantly.

These results demonstrate the importance of domain knowledge. DHRConv has a channel-level receptive field to extract features of delayed hemodynamic responses, and DWConv with global receptive fields aggregates spatial information.

TABLE III
EFFICIENCY METRICS FOR EACH MODEL

Dataset	Model	Parameters	FLOPs	Inference time (ms)	FPS
MA	CNN [8]	4.54 M	14.97 M	0.43	2333
	1D-CNN [23]	0.93 M	4.03 M	0.64	1574
	LSTM [8]	10.92 K	0.82 M	2.19	456
	fNIRS-T [11]	0.55 M	29.39 M	4.48	223
	fNIRSNet	498	10.46 K	0.07	15459
	fNIRSNet†	2370	42.21 K	0.08	13241
UFFT	CNN [8]	2.78 M	9.96 M	0.38	2644
	1D-CNN [23]	0.29 M	2.63 M	0.71	1408
	LSTM [8]	11.74 K	0.49 M	1.32	756
	fNIRS-T [11]	0.55 M	13.52 M	4.48	223
	fNIRSNet	419	7.46 K	0.06	16897
	fNIRSNet†	2051	30.21 K	0.07	15176

Units K and M refer to $\times 10^3$ and $\times 10^6$, respectively. † denotes subject-independent.

B. Comparison of Efficiency

fNIRSNet is a lightweight model with extremely low parameters and computational cost. Table III reports the efficiency metrics for each model. Inference time and FPS are only used as references due to the metrics relying on hardware platforms. These tests are conducted on NVIDIA GTX 1080 GPU with 8 GB memory. All metrics of fNIRSNet significantly outperform other models. In the subject-specific experiments, fNIRSNet with 498 parameters is 6.58% higher than CNN with 4.54 M parameters on MA, and the inference time of fNIRSNet is 6 times lower than CNN. fNIRS-T has the highest FLOPs because the complexity of the self-attention mechanism is the square of input dimension [41]. Thus, it has a high inference time and low FPS. We also observe a similar situation on UFFT. Therefore, fNIRSNet is friendly for practical applications and could be deployed on embedded devices.

C. Ablation Study

Subject-specific experiments are conducted on the UFFT dataset to ablate the three design guidelines.

1) Guideline 1: In Table IV, the accuracy and F1-score keep improving as the width of DHRConv increases. For a signal

TABLE IV

Subject-Specific Experimental Results of Different Convolutional Kernel Sizes on the UFFT Dataset

DHRConv	DWConv	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa	Parameters	FLOPs (K)
1 × 40	40 × 1	68.67 ± 19.85	69.97 ± 19.61	69.15 ± 19.52	68.02	0.53	419	7.46
1 × 10	40 × 1	66.56 ± 19.12	67.23 ± 19.45	66.72 ± 19.01	65.70	0.50	1019	82.34
1 × 20		66.86 ± 19.71	67.72 ± 19.85	67.06 ± 19.57	66.15	0.50	819	89.38
1 × 30		67.59 ± 19.65	68.62 ± 19.81	67.90 ± 19.57	66.96	0.51	619	64.42
1 × 40	10 × 1	67.85 ± 21.17	69.15 ± 21.03	68.03 ± 21.00	67.09	0.52	1019	11.42
	20 × 1	67.88 ± 20.54	68.94 ± 20.59	68.11 ± 20.36	67.14	0.52	819	10.90
	30 × 1	67.48 ± 20.99	68.20 ± 21.33	67.74 ± 20.69	66.62	0.51	619	9.58

TABLE V

Subject-Specific Experimental Results for Different Types of Convolutions on the UFFT Dataset. STD Means Standard Convolution, and DWS Denotes Depthwise Separable Convolution

1st/2nd Module	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa	Parameters	FLOPs (K)	Inference time (ms)	FPS
STD/DWS	68.67 ± 19.85	69.97 ± 19.61	69.15 ± 19.52	68.02	0.53	419	7.46	0.06	16897
STD/STD	67.22 ± 19.53	68.43 ± 19.72	67.43 ± 19.47	66.35	0.51	1503	8.54	0.06	16157
DWS/DWS	68.26 ± 20.87	69.47 ± 20.80	68.67 ± 20.58	67.42	0.53	304	2.82	0.07	14508
DWS/STD	68.28 ± 19.41	69.68 ± 19.56	68.68 ± 19.29	67.51	0.52	1388	3.90	0.07	13658

TABLE VI

Subject-Specific Experimental Results of Different Activation Functions on the UFFT Dataset

1st/2nd Module	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa
Sigmoid/Sigmoid	68.67 ± 19.85	69.97 ± 19.61	69.15 ± 19.52	68.02	0.53
Tanh/Tanh	67.09 ± 19.88	68.34 ± 19.83	67.34 ± 19.64	66.40	0.51
ReLU/ReLU	61.66 ± 18.81	63.08 ± 18.54	61.93 ± 18.74	60.97	0.42
LReLU/LReLU	61.94 ± 19.03	62.70 ± 19.20	62.06 ± 18.93	61.13	0.43
ELU/ELU	67.13 ± 19.25	68.54 ± 19.26	67.38 ± 19.14	66.42	0.51
Sigmoid/ReLU	67.44 ± 19.37	68.46 ± 19.79	67.52 ± 19.22	66.61	0.51
ReLU/Sigmoid	61.94 ± 19.08	63.21 ± 19.32	62.42 ± 18.86	61.18	0.43

tensor $X \in \mathbb{R}^{40\times40\times1}$ (i.e., 20 channels × 2 chromophores = 40 and 13.3 Hz × 3 s = 40), the 1 × 40 DHRConv can extract the complete features of delayed hemodynamic responses instead of using small convolutions. The 1 × 40 DHRConv also reduces FLOPs significantly. Furthermore, fNIRSNet exhibits poorer classification performance when the height of DWConv is reduced from 40 to 10. Therefore, convolutions with global receptive fields are more beneficial to fNIRS classification than local receptive fields and avoid over-parameterization caused by stacking many convolutional layers to enlarge receptive fields.

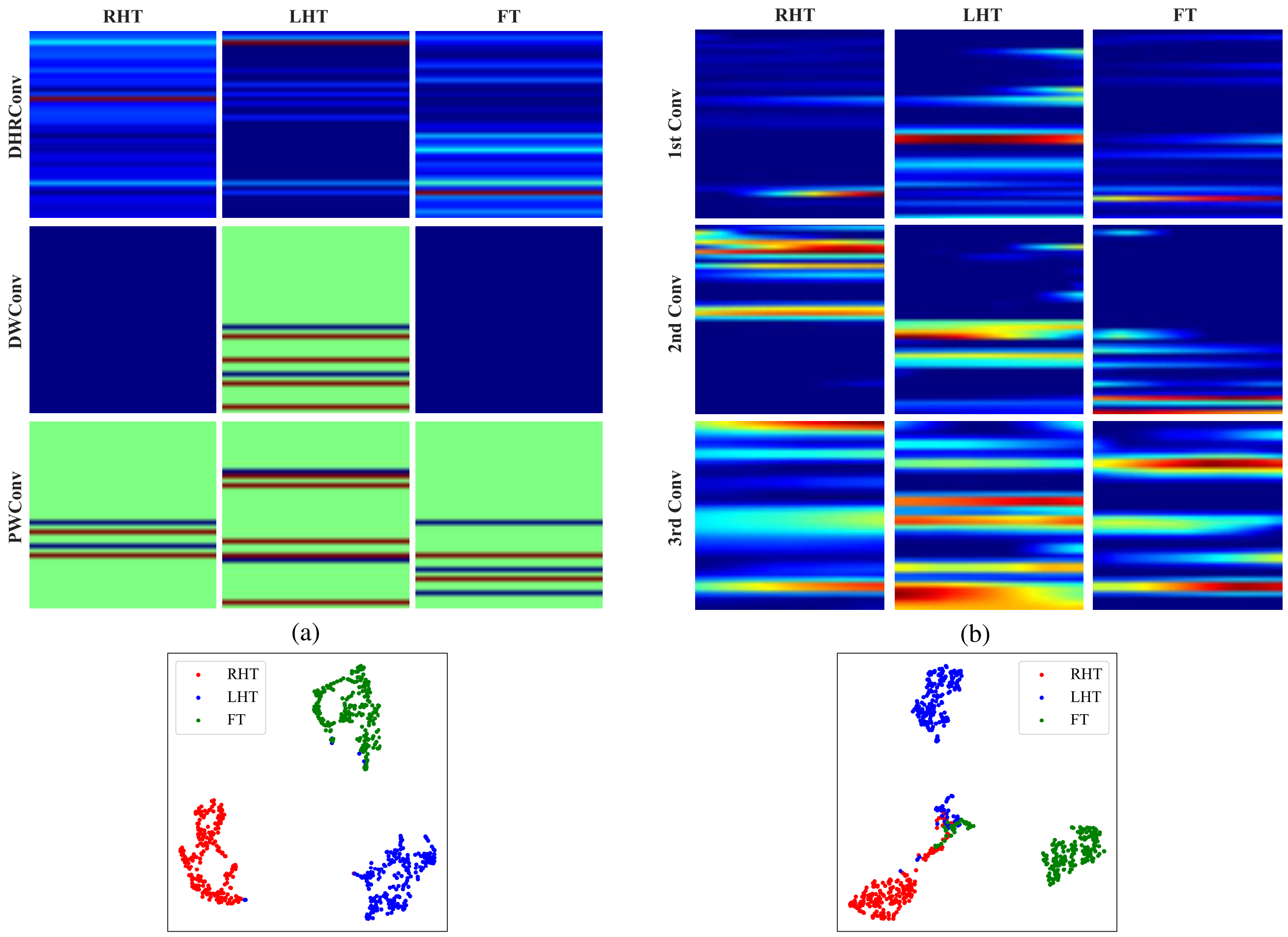
2) Guideline 2: The hybrid pattern that the first module uses standard convolutions and the second uses depthwise separable convolutions mainly to balance the stability, speed, and efficiency of fNIRSNet. The experimental results are reported in Table V. The pure depthwise separable convolutions (i.e., DWS/DWS) have the lowest parameters and FLOPs, while inference time and FPS show deterioration. In practice, the arithmetic intensity (ratio of FLOPs to memory accesses) of depthwise separable convolutions is too low and the hardware usage is inefficient [42]. However, the hybrid pattern (i.e., STD/DWS) yields lower and more stable standard deviations and the highest running efficiency.

3) Guideline 3: The type and position of activation functions affect fNIRSNet performance significantly. The results are summarized in Table VI. Saturation activation functions, such as sigmoid and tanh, outperform the more popular ReLU and LReLU by about 6%. The hybrid activations (i.e., Sigmoid/ReLU and ReLU/Sigmoid) further reveal the effect of activation position on performance. Experimental results using sigmoid and ReLU as the first and second activation functions are significantly higher than those using ReLU and sigmoid. The first convolution layer followed by saturating activation functions is beneficial to preserve information for fNIRSNet. The results of ELU with saturated negative values and LReLU further reveal that this benefit comes from negative saturation that decreases the forward propagated variation and information [34].

D. Visualization

In this subsection, advanced visualization techniques are used to explain how fNIRSNet works on UFFT. Grad-CAM [43] is adopted to study the effect of each convolutional layer. Grad-CAM uses the gradient of the target flowing into convolutional layers to generate a rough heat map to highlight important regions. As shown in Figs. 5(a) and 5(b), the heat map activation pattern of fNIRSNet is different from CNN. The heat map of DHRConv of fNIRSNet covers the entire fNIRS channels. The $1 \times S$ DHRConv has channel-level receptive fields, and the $C \times 1$ DWConv has global receptive fields. The heat map of CNN only covers a part of the fNIRS channels, especially for the first two layers. This phenomenon is related to the local receptive fields of convolutions. As the depth of CNN increases, receptive fields gradually increase. Thus, the heat map of the third layer tends to extend to the whole channel.

Fig. 6 illustrates the grand average of all subjects. Fig. 3(c) shows the fNIRS channel locations. The motor cortex regions in contralateral hemispheres are well-activated when subjects perform finger-tapping tasks [5]. For RHT and LHT, HbO



(a) (b)

(c) (d)

Fig. 5. (a) and (b) are the Grad-CAM visualization of fNIRSNet and CNN for Subject 1, respectively. (c) and (d) represent the t-SNE visualization of fNIRSNet and CNN for Subject 1, sequentially.

and HbR located in the anterior regions of C3 and C4 show significant changes at Ch 5, 6, 15, and 16, while the peak of hemodynamic responses is delayed by nearly 8 s. Unlike RHT and LHT, hemodynamic responses of the FT task are well-activated at 5 s. After the task period (0–10 s), signals still have solid hemodynamic responses. Therefore, convolutions require wider scales or receptive fields to extract delayed features. The kernel width of DHRConv is the number of sampling points, which can fully extract features of delayed hemodynamic responses. Grad-CAM illustrates the degree of contribution to the predicted results by highlighting different fNIRS channels. A redder color indicates a greater contribution. In Fig. 5(a), the Grad-CAM of DHRConv exhibits an automatic channel selection function that removes redundant signals and selects regions of interest to improve robustness. Compared to local receptive fields, DWConv with global receptive fields can extract long-range contextual information and focus on activation patterns in different brain regions. PWConv facilitates information interaction among the channels of DWConv. In Fig. 5(a), DWConv shows highlighting for HbR of the LHT task, while PWConv extends the highlighted distribution to HbO.

The t-SNE [44] is used to visualize features learned by the FC layer in a two-dimensional space. In Fig. 5(c), the t-SNE of fNIRSNet has a distinct feature distribution that intra-clusters are tightly together and inter-clusters are highly separated into a triangular structure. However, Fig. 5(d) shows that the features learned by CNN exhibit non-separability in the middle region. Therefore, fNIRSNet has excellent feature learning capabilities.

E. Pooling and Dropout

The architecture of fNIRSNet does not use pooling or dropout. We are interested in whether these components can further improve performance. A 2 × 1 average pooling layer is inserted between the DHR and global modules. A dropout layer with a dropout rate of 0.5 is added after the global module. In Table VII, average pooling and dropout do not enhance performance. The average pooling reduces model parameters and dropout prevents overfitting. In addition, average pooling has a higher impact on performance deterioration than dropout. They may lead to underfitting and decreased learning capability because fNIRSNet has fewer model parameters.

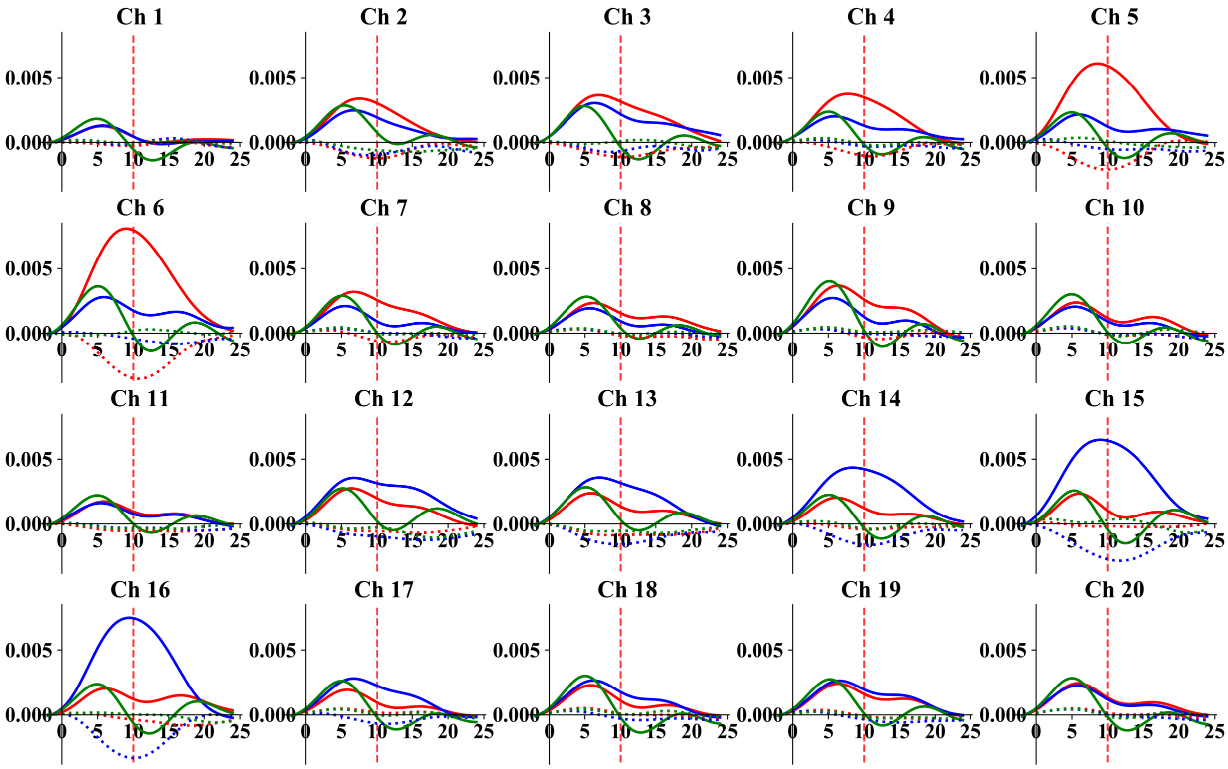


Fig. 6. Grand average of the fNIRS signals over subjects. The X-axis interval is [−2, 24] s and the Y-axis interval is [−0.0038, 0.0085] mM·cm. The red vertical dotted lines indicate the end of the task period (0–10 s). The solid and dotted curves represent HbO and HbR, respectively. The red, blue, and green curves correspond to RHT, LHT, and FT, respectively.

TABLE VII

Subject-Specific Experimental Results on the UFFT Dataset

Pooling	Dropout	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa	Parameters	FLOPs (K)
✗	✗	68.67 ± 19.85	69.97 ± 19.61	69.15 ± 19.52	68.02	0.53	419	7.46
✓	✗	65.67 ± 19.78	66.44 ± 19.91	66.18 ± 19.32	64.73	0.49	339	7.38
✗	✓	67.56 ± 20.30	68.73 ± 20.46	68.26 ± 19.80	66.66	0.52	419	7.46

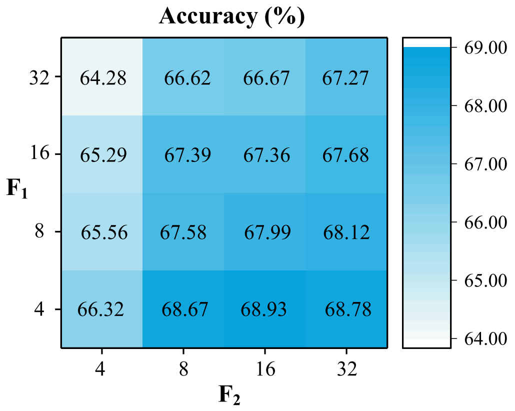


Fig. 7. Parameter sensitivity of fNIRSNet for subject-specific experiments on the UFFT dataset.

F. Parameter Sensitivity

Parameter sensitivity is presented for F_1 and F_2 of fNIRSNet. Fig. 7 shows the average accuracy of the UFFT dataset in subject-specific experiments. We found that increasing F_2 can improve the classification performance of fNIRSNet when F_1 is fixed. Increasing F_2 helps the global module capture the contextual dependencies of fNIRS channels. When F_1 equals 4, the classification performance gradually saturates as the value of F_2 increases to 32. Therefore, we recommend setting the value of F_2 to at least twice that of F_1 if other researchers apply fNIRSNet to their data.

V. Discussion

Currently, evaluation protocols for fNIRS classification are confusing. Deep learning-based fNIRS classification research has become popular in recent years, while some early protocols for open-access datasets are based on traditional machine learning classifiers. We found that these protocols are not suitable for evaluating the performance of DNNs. Although these studies may achieve higher performance, experimental results need to be further investigated. For example, Shin et al. [4] classified signal segments from the same time for each subject on the MA dataset, which does not validate classifier generalization across different time segments. In addition, Shin et al. suggest that their study is not dedicated to benchmark machine learning classifiers [4], while some studies follow the protocols. Sun et al. [23] used DNNs to perform 5-fold cross-validation on 60 segments from the same time segments.

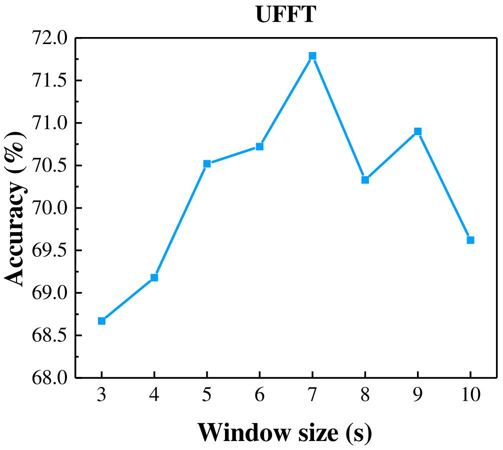


Fig. 8. Subject-specific experiments are conducted using various sliding window sizes on the UFFT dataset.

However, DNNs are difficult to learn more generalized feature representations from a small amount of data. Kwak et al. [45] reported the average and maximum accuracy among 10 segments in a trial. Bak et al. [5] published the UFFT dataset and used leave-one-out cross-validation (LOO-CV) to evaluate classifier performance. However, LOO-CV is rarely used to evaluate DNNs considering the high cost of training. More importantly, the above studies do not conduct subject-independent experiments. Our previous work [11], [46] used KFold-CV and LOSO-CV for trial-wise fNIRS classification. Moreover, we segment signals by sliding windows to increase the number of samples to alleviate overfitting.

The design philosophy of fNIRSNet is different from other models. Our motivation is to introduce domain knowledge into the model design: 1) convolutions with channel-level receptive fields extract the features of slow delayed hemodynamic responses rather than small convolutions with local receptive fields sliding over fNIRS signals; 2) convolutions with global receptive fields help discover the activation patterns of different brain regions. Other deep models that do not introduce domain knowledge struggle to extract more meaningful and discriminative features. Moreover, these over-parameterized models would bring more optimization problems on a limited dataset. fNIRSNet with fewer parameters and FLOPs achieves higher classification performance and reduces BCI hardware resource consumption significantly. In subject-specific experiments, fNIRSNet with only 498 trainable parameters yielded better results than CNNs with millions of parameters on MA. In addition, the model inference time (see Table III) is much lower than other baseline models. Therefore, our study may inspire more knowledge-driven models.

Our study still has limitations. The size of the sliding window limits the long-range dependency of hemodynamic responses, potentially affecting classification performance. This situation is illustrated by subject-specific experiments on UFFT. The step size is set to 1 s and signals for each trial are split into 8 segments to maintain a fixed total number of data samples. As shown in Fig. 8, the average accuracy of fNIRSNet improves from 68.67% to 71.79% when the sliding window increases from 3 s to 7 s, i.e., the size of DHRConv increases from 1 × 40 to 1 × 93 (i.e., 13.3 Hz × 7 s). After that, the average accuracy starts to decrease. However, this blurs the boundary between the task and rest period because some of the signals in the rest period are also considered as a continuation of the task period. For example, the eighth segment of the signal covers a time interval of [7, 17] s when the window size is 10 s, which includes a 7-second rest period. Furthermore, this continuation may interfere with real-time classification for hybrid EEG-fNIRS BCIs. EEG has returned to the resting state, whereas fNIRS still has a delayed response because its lower temporal resolution compared to EEG. The primary aim of this study is to investigate fNIRS classification during the task state. In fact, fNIRS classification studies have rarely discussed this continuation operation, which could be related to specific tasks.

In the future, we will explore more research directions, such as mental health detection [47], [48], brain-related disorder diagnosis [49], Hamilton–Jacobi–Bellman (HJB) equation for training fNIRS models [50], and fusion of EEGNet [51] and fNIRSNet for hybrid EEG-fNIRS BCIs.

VI. Conclusion

In this study, we rethink delayed hemodynamic responses for fNIRS-based BCIs and propose a concise and efficient fNIRSNet for fNIRS classification. We summarize three design guidelines for fNIRSNet. The proposed model with fewer parameters and FLOPs achieves better classification results on open-access datasets. Furthermore, Grad-CAM and t-SNE explain the role of each convolutional layer and feature learning capabilities. fNIRSNet is ideally suitable for real-world applications and reduces the hardware configuration of BCI systems.

References

[1] F. F. Jöbsis, "Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters," *Science*, vol. 198, no. 4323, pp. 1264–1267, Dec. 1977.

[2] J. Mladenovic, J. Frey, S. Pramij, J. Mattout, and F. Lotte, "Towards identifying optimal biased feedback for various user states and traits in motor imagery BCI," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 3, pp. 1101–1110, Mar. 2022.

[3] Z. Liu, J. Shore, M. Wang, F. Yuan, A. Buss, and X. Zhao, "A systematic review on hybrid EEG/fNIRS in brain–computer interface," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102595.

[4] J. Shin et al., "Open access dataset for EEG+NIRS single-trial classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 10, pp. 1735–1745, Oct. 2017.

[5] S. Bak, J. Park, J. Shin, and J. Jeong, "Open-access fNIRS dataset for classification of unilateral finger- and foot-tapping," *Electronics*, vol. 8, no. 12, p. 1486, Dec. 2019.

[6] J. Chao et al., "FNIRS evidence for distinguishing patients with major depression and healthy controls," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2211–2221, 2021.

[7] H. Nazeer et al., "Enhancing classification accuracy of fNIRS-BCI using features acquired from vector-based phase analysis," *J. Neural Eng.*, vol. 17, no. 5, Oct. 2020, Art. no. 056025.

[8] B. Lyu et al., "Domain adaptation for robust workload level alignment between sessions and subjects using fNIRS," *J. Biomed. Opt.*, vol. 26, no. 2, Jan. 2021, Art. no. 022908.

[9] T. Ma et al., "CNN-based classification of fNIRS signals in motor imagery BCI system," *J. Neural Eng.*, vol. 18, no. 5, Apr. 2021, Art. no. 056019.

[10] U. Asgher et al., "Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain–computer interface," *Frontiers Neurosci.*, vol. 14, p. 584, Jun. 2020.

[11] Z. Wang, J. Zhang, X. Zhang, P. Chen, and B. Wang, "Transformer model for functional near-infrared spectroscopy classification," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 6, pp. 2559–2569, Jun. 2022.

[12] K. Liu, M. Yang, Z. Yu, G. Wang, and W. Wu, "FBMSNet: A filter-bank multi-scale convolutional neural network for EEG-based motor imagery decoding," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 2, pp. 436–445, Feb. 2023.

[13] G. Jasdzewski, G. Strangman, J. Wagner, K. K. Kwong, R. A. Poldrack, and D. A. Boas, "Differences in the hemodynamic response to event-related motor and visual paradigms as measured by near-infrared spectroscopy," *NeuroImage*, vol. 20, no. 1, pp. 479–488, Sep. 2003.

[14] M. Li, A. T. Newton, A. W. Anderson, Z. Ding, and J. C. Gore, "Characterization of the hemodynamic response function in white matter tracts for event-related fMRI," *Nature Commun.*, vol. 10, no. 1, p. 1140, Mar. 2019.

[15] A. Zafar and K.-S. Hong, "Reduction of onset delay in functional near-infrared spectroscopy: Prediction of HbO/HbR signals," *Frontiers Neurorobotics*, vol. 14, p. 10, Feb. 2020.

[16] I. Nambu, R. Osu, M.-A. Sato, S. Ando, M. Kawato, and E. Naito, "Single-trial reconstruction of finger-pinch forces from human motor-cortical activation measured by near-infrared spectroscopy (NIRS)," *NeuroImage*, vol. 47, no. 2, pp. 628–637, Aug. 2009.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014.

[18] T. Ishida, I. Yamane, T. Sakai, G. Niu, and M. Sugiyama, "Do we need zero training loss after achieving zero training error?" in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 4554–4564.

[19] N. M. Sommer, B. Kakillioglu, T. Grant, S. Velipasalar, and L. Hirshfield, "Classification of fNIRS finger tapping data with multi-labeling and deep learning," *IEEE Sensors J.*, vol. 21, no. 21, pp. 24558–24569, Nov. 2021.

[20] J. Wang, T. Grant, S. Velipasalar, B. Geng, and L. Hirshfield, "Taking a deeper look at the brain: Predicting visual perceptual and working memory load from high-density fNIRS data," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 5, pp. 2308–2319, May 2022.

[21] M. A. Tanveer, M. J. Khan, M. J. Qureshi, N. Naseer, and K.-S. Hong, "Enhanced drowsiness detection using deep learning: An fNIRS study," *IEEE Access*, vol. 7, pp. 137920–137929, 2019.

[22] Q. He, L. Feng, G. Jiang, and P. Xie, "Multimodal multitask neural network for motor imagery classification with EEG and fNIRS signals," *IEEE Sensors J.*, vol. 22, no. 21, pp. 20695–20706, Nov. 2022.

[23] Z. Sun, Z. Huang, F. Duan, and Y. Liu, "A novel multimodal approach for hybrid brain–computer interface," *IEEE Access*, vol. 8, pp. 89909–89918, 2020.

[24] T. Trakoolwilaiwan, B. Behboodi, J. Lee, K. Kim, and J.-W. Choi, "Convolutional neural network for high-accuracy functional near-infrared spectroscopy in a brain–computer interface: Three-class classification of rest, right-, and left-hand motor execution," *Neurophotonics*, vol. 5, no. 1, Sep. 2017, Art. no. 011008.

[25] M. E. Raichle, "Behind the scenes of functional brain imaging: A historical and physiological perspective," *Proc. Nat. Acad. Sci. USA*, vol. 95, no. 3, pp. 765–772, Feb. 1998.

[26] S. Ogawa et al., "Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging," *Proc. Nat. Acad. Sci. USA*, vol. 89, no. 13, pp. 5951–5955, Jul. 1992.

[27] Z. Y. Shan et al., "Modeling of the hemodynamic responses in block design fMRI studies," *J. Cerebral Blood Flow Metabolism*, vol. 34, no. 2, pp. 316–324, Feb. 2014.

[28] T. Ernst and J. Hennig, "Observation of a fast response in functional MR," *Magn. Reson. Med.*, vol. 32, no. 1, pp. 146–149, Jul. 1994.

[29] M. L. Schroeter, T. Kupka, T. Mildner, K. Uludağ, and D. Y. von Cramon, "Investigating the post-stimulus undershoot of the BOLD signal—A simultaneous fMRI and fNIRS study," *NeuroImage*, vol. 30, no. 2, pp. 349–358, Apr. 2006.

[30] P. C. M. van Zijl, J. Hua, and H. Lu, "The BOLD post-stimulus undershoot, one of the most debated issues in fMRI," *NeuroImage*, vol. 62, no. 2, pp. 1092–1102, Aug. 2012.

[31] L. SIfre and S. Mallat, "Rigid-motion scattering for texture classification," 2014, *arXiv:1403.1687*.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, vol. 1, Jul. 2015, pp. 448–456.

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[34] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.

[35] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn.*, vol. 30, no. 1. Atlanta, GA, USA, 2013, pp. 1–6.

[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[37] M. Cope and D. T. Delpy, "System for long-term measurement of cerebral blood and tissue oxygenation on newborn infants by near infra-red transillumination," *Med. Biol. Eng. Comput.*, vol. 26, no. 3, pp. 289–294, May 1988.

[38] W. Ko, E. Jeon, and H.-I. Suk, "A novel RL-assisted deep learning framework for task-informative signals selection and classification for spontaneous BCIs," *IEEE Trans. Ind. Informat.*, vol. 18, no. 3, pp. 1873–1882, Mar. 2022.

[39] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–18.

[40] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–16.

[41] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5999–6009.

[42] B. Wu et al., "Shift: A zero FLOP, zero parameter alternative to spatial convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9127–9135.

[43] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[44] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

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

[46] Z. Wang, J. Zhang, Y. Xia, P. Chen, and B. Wang, "A general and scalable vision framework for functional near-infrared spectroscopy classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1982–1991, 2022.

[47] S. K. Khare, S. March, P. D. Barua, V. M. Gadre, and U. R. Acharya, "Application of data fusion for automated detection of children with developmental and mental disorders: A systematic review of the last decade," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101898.

[48] H. Yaacob, F. Hossain, S. Shari, S. K. Khare, C. P. Ooi, and U. R. Acharya, "Application of artificial intelligence techniques for brain–computer interface in mental fatigue detection: A systematic review (2011–2022)," *IEEE Access*, vol. 11, pp. 74736–74758, 2023.

[49] S. Roy, I. Kiral-Kornek, and S. Harrer, "ChronoNet: A deep recurrent neural network for abnormal EEG identification," in *Artificial Intelligence in Medicine*, D. Riaño, S. Wilk, and A. ten Teije, Eds. Cham, Switzerland: Springer, 2019, pp. 47–56.

[50] T. K. Reddy, V. Arora, and L. Behera, "HJB-equation-based optimal learning scheme for neural networks with applications in brain–computer interface," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 2, pp. 159–170, Apr. 2020.

[51] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Jul. 2018, Art. no. 056013.