Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 3178284, 17 pages
https://doi.org/10.1155/2023/3178284

Research Article

Fusion of Deep Features from 2D-DOST of fNIRS Signals for Subject-Independent Classification of Motor Execution Tasks

Pouya Khani,[1] **Vahid Solouk**,[2] **Hashem Kalbkhani**,[1] **and Farid Ahmadi**[2]

[1]*Faculty of Electrical Engineering, Urmia University of Technology, Urmia, Iran*
[2]*Department of IT and Computer Engineering, Urmia University of Technology, Urmia, Iran*

Correspondence should be addressed to Farid Ahmadi; f.ahmadi@uut.ac.ir

Received 12 September 2023; Revised 28 November 2023; Accepted 14 December 2023; Published 20 December 2023

Academic Editor: Mohammad R. Khosravi

Functional near-infrared spectroscopy (fNIRS) is a low-cost and noninvasive method to measure the hemodynamic responses of cortical brain activities and has received great attention in brain-computer interface (BCI) applications. In this paper, we present a method based on deep learning and the time-frequency map (TFM) of fNIRS signals to classify the three motor execution tasks including right-hand tapping, left-hand tapping, and foot tapping. To simultaneously obtain the TFM and consider the correlation among channels, we propose to utilize the two-dimensional discrete orthonormal Stockwell transform (2D-DOST). The TFMs for oxygenated hemoglobin (HbO), reduced hemoglobin (HbR), and two linear combinations of them are obtained and then we propose three fusion schemes for combining their deep information extracted by the convolutional neural network (CNN). Two CNNs, LeNet and MobileNet, are considered and their structures are modified to maximize the accuracy. Due to the lack of enough signals for training CNNs, data augmentation based on the Wasserstein generative adversarial network (WGAN) is performed. Several simulations are performed to assess the performance of the proposed method in three-class and binary scenarios. The results present the efficiency of the proposed method in different scenarios. Also, the proposed method outperforms the recently introduced methods.

1. Introduction

1.1. Motivation. The human brain is the most complex organ in the body, consisting of billions of neurons and unique computing capabilities such as parallel processing and learning. Therefore, researchers have always been interested in analyzing it from an early age. Different areas such as neuroscience, artificial intelligence, cognitive science, and the brain-computer interface (BCI) have been explored to understand the brain better [1]. BCI is a tool that translates thoughts and provides an interface for communicating with the outside world. Recent advances in BCI have led to a better understanding of neural functions and connections in the brain. BCI is an extensive study and requires knowledge of computer engineering, neuroscience, psychology, signal processing, and clinical rehabilitation [2].

The functional near-infrared spectroscopy (fNIRS) is a noninvasive imaging technique that measures changes in blood oxygenation levels in the brain. It uses near-infrared light to penetrate the scalp and skull, allowing for the detection of hemodynamic responses associated with brain activity. This noninvasiveness makes fNIRS a safe and comfortable option for users, as it does not require any surgical procedures or direct contact with the brain. fNIRS has good spatial and temporal resolution. It can provide information about the location and timing of brain activity, allowing for the identification of specific brain regions involved in cognitive processes. This spatial and temporal resolution is crucial for BCI applications, as it enables the accurate decoding and interpretation of brain signals for controlling external devices or communicating with the environment. It can be implemented in various settings, including home environments, clinics, or even during real-

world tasks. This flexibility makes fNIRS a practical choice for BCI applications, as it allows for more natural and ecologically valid experiments and interactions. Also, fNIRS is less susceptible to motion artifacts compared to other imaging techniques. It can tolerate small head movements and is less affected by electrical interferences or muscle artifacts. This robustness to motion artifacts makes fNIRS suitable for real-time applications, where users may engage in natural movements or activities while using the BCI system [3]. These characteristics make fNIRS a promising tool for developing practical and user-friendly BCI systems.

1.2. Related Works. In general, BCI systems include signal acquisition, signal processing, and output units. The recorded signals are low-power with a poor signal-to-noise ratio (SNR), nonstationary, nonlinear, and time-varying. Therefore, to improve the real-time processing of these systems, feature extraction methods should reflect the time-frequency characteristics and spatial features. Temporal frequency analysis is widely used in BCI research. These methods are short-time Fourier transform (STFT), wavelet transform, and Hilbert–Huang transform (HHT). Their results can be expressed as power spectrum density (PSD) and are the most effective in processing nonstationary and nonlinear signals.

Some works considered traditional feature extraction schemes based on statistical methods. In [4], the difference between the two mental tasks of computation and rest state was analyzed based on fNIRS signals. The authors extracted six features from each channel of the fNIRS signal. The results showed that multilayer perceptron (MLP) performs better than support vector machine (SVM) and k-nearest neighbor (kNN). In [5], the fNIRS signals with 22 channels were collected during three mental tasks: number subtraction, word generation, and rest. The MLP model based on superficial features determined the task. Subsequently, the authors controlled the robot remotely via fNIRS signals. In [6], the combination of three-channel fNIRS and 123-channel electroencephalography (EEG) signals was used to classify the left/right brain excitatory signals. Sixteen features were extracted from fNIRS signals, and an MLP with four hidden layers was used for classification. In [7], the concentration changes of oxygenated hemoglobin (HbO) and reduced hemoglobin (HbR) were measured, while volunteers repeated each of the three types of overt movements, including left- and right-hand unilateral complex finger-tapping, and foot-tapping, by considering 20-channel fNIRS signals from 30 volunteers classified by SVM. In [8], the authors aimed to distinguish the four brain activities including mental arithmetic (MA), motor imagery of left hand and right hand, and rest from fNIRS signals. After preprocessing, the six different statistical features are obtained in the time domain and 13 Mel-frequency cepstral coefficient (MFCC) features are obtained in the frequency domain, and then, classification is performed by SVM and kNN. The least absolute shrinkage and selection operator (LASSO) homotopy-based sparse representation was employed in [9] for channel selection. Classification profits from statistical spatial features of concentration of blood oxygenation from fNIRS in walk and rest state tasks. In the presence of complicated and nonstationary signals, the mentioned methods based on statistical features cannot achieve the efficient accuracy.

Time-frequency analysis of fNIRS signals was considered in several works. In [10], the frontal hemodynamic responses were recorded considering 19-channel fNIRS signals from nine patients during mental tasks. The authors used continuous wavelet transform for multiscale decomposition and a soft-threshold algorithm for denoising. They considered the MLP, linear discriminant analysis (LDA), and SVM and compared their performances. The multilevel mental workload classification was performed in [11] by using bivariate functional brain connectivity features in three time-frequency bands. They utilized the public hybrid dataset consisting of EEG-fNIRS to evaluate their proposed method. The mentioned approaches extract the nondeep features from time-frequency components and, as a result, fail to perform the correct classification in complex scenarios [12].

Methods based on neural networks and deep learning were also introduced for utilizing fNIRS signals in BCI applications. In [13], multistage fusion was performed to classify left- or right-hand motor-imagery tasks considering the EEG and fNIRS signals. The results showed that the y-shaped neural network with early stage feature fusion has the best performance compared to the others. In [14] participants were asked to do left- and right-hand motor imagery experiments, and the corresponding fNIRS signals were recorded. The classification is based on a convolutional neural network (CNN). A deep belief network (DBN) based on a restricted Boltzmann machine (RBM) was used in [15] to classify fNIRS signals of flexion and extension imagery involving the left and right arms. The features of HbO concentration were used to train two RBMs. In [16], the authors attempted to classify the gender through four-channel fNIRS signals. The authors used a three-layer denoising autoencoder (DAE) to extract distinct features to accommodate gender recognition by MLP. The authors in [17] extracted the features from five fNIRS signals by employing the convolutional autoencoder (CAE) and echo state network (ESN) autoencoder for driver cognitive load levels. In [18], a framework consisting of machine and deep learning methods classified the fNIRS signals of motor execution for walking and rest tasks. They demonstrated deep learning approaches including the CNN, LSTM, and Bi-LSTM with the results of 88.50%, 84.24%, and 85.13%, respectively, that reached higher accuracy compared to kNN, SVM, and LDA. These methods considered the neural network and deep learning approaches; however, they did not consider the time-frequency analysis to consider nonstationary nature of biological signals.

1.3. Contributions. Most biological signals, especially fNIRS signals, are recorded in several channels. When the signals of each channel are analyzed separately, the correlation between channels is not considered and some information is missing. Hence, in this research, it is proposed to use the two-dimensional discrete orthonormal Stockwell transform (2D-DOST) to obtain the time-frequency map (TFM) fNIRS signals of motor execution tasks by considering the correlation between channels in the time domain. Each TFM can be regarded as a feature set and we can give it directly to a CNN for classification. The fNIRS signals are decomposed into HbO and HbR signals as well as their combinations. Hence, there are some TFMs, and appropriate fusion schemes are required to aggregate their information. In this paper, three fusion schemes are employed: early, joint, and late fusions. In these fusion schemes, feature extraction and classification are performed by two CNNs including MobileNet and LeNet. Since the channel selection considers several channels for obtaining TFM, it is proposed to modify the structure of CNN to achieve efficient accuracy. Also, data augmentation based on the Wasserstein generative adversarial network (WGAN) is performed to increase the generalization of CNN training. The results show the efficiency of the proposed method for feature extraction and fusion of extracted TFMs.

The rest of this paper is organized as follows. Section 2 explains the dataset and preliminaries used in this paper. Section 3 presents the proposed method in detail. Section 4 contains the results, and finally, Section 5 concludes the paper.

2. Dataset and Preliminaries

2.1. Dataset. In this paper, we considered the dataset used in [7] in which a total of 30 volunteers participated to collect the dataset. Each volunteer performed the following tasks 25 times in random order: right-hand finger-tapping (RHT), left-hand finger-tapping (LHT), and foot-tapping (FT). Therefore, we have a three-class classification scenario along with binary scenarios. For person-specific classification, there are only 25 measurements for each class which may be not enough for the training of a CNN. On the other hand, if the data of all subjects are merged, there are 750 recordings from each class. The fNIRS data were recorded by a three-wavelength continuous-time multichannel fNIRS system (LIGHTNIRS, Shimadzu, Kyoto, Japan) consisting of eight light sources (Tx) and eight detectors (Rx). Four Tx and Rx were placed around C3 on the left hemisphere, and the rest were placed around C4 on the right hemisphere. Figure 1 depicts the channel locations of the fNIRS. Ch01–10 and Ch11–20 are located around C3 (Ch09) and C4 (Ch18), respectively. The channels are created by a pair of adjacent light sources (Tx) and detectors (Rx) placed 30 mm away from each other.

The experiment diagram is shown in Figure 2. A single trial comprised an introduction period (−2 to 0 s) and a task period (0 to 10 s), followed by an inter-trial break period (10 to 27–29 s). Among RHT (⟶), LHT (⟵), and FT (↓), a random task type was displayed during the introduction

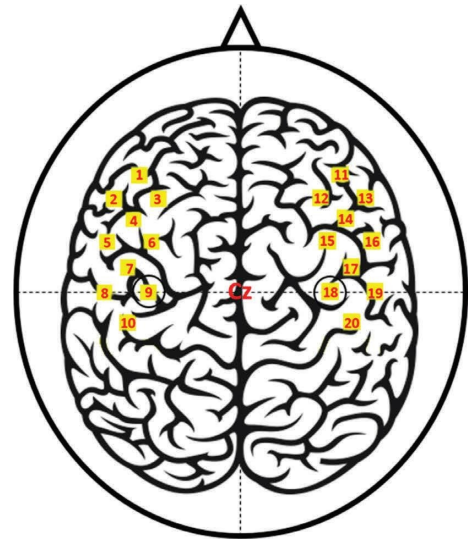

Figure 1: The placement of sources and detectors to record fNIRS signals [7].

period, which the volunteers were required to perform. For RHT/LHT, the volunteers performed unilateral complex finger-tapping. They tapped their thumbs with other fingers one by one in the direction from the index finger to the little finger and repeated it in the reverse order. The tapping continued at a steady rate of two Hz. For FT, the volunteers tapped their foot on the same side of their dominant hand constantly at a one Hz rate. Considering the 20 channels, measuring both oxygenated hemoglobin (HbO) and reduced hemoglobin (HbR), 10 s for task duration, and sampling frequency of 13.33 Hz, the duration of the task contains about 133 samples, and data of each task contain 40×133 matrix.

2.2. 2D-DOST. Stockwell transform (ST) was introduced in [19] and originates from STFT and wavelet transform. It is very efficient in terms of resolution at low frequencies and also has a higher resolution at high frequencies; for this reason, it is possible to access the frequency components in the time-frequency domain. However, it is highly redundant because it requires a lot of time and storage space. Discrete orthonormal ST (DOST), a downsampled version of ST, was proposed to overcome this problem. Because low frequencies have a high period, sampling is performed at a lower rate, and similarly, for high frequencies, high-rate sampling is performed by DOST. Suppose $z(t)$ is a continuous-time signal; its ST is calculated as [20]

$$\mathbf{S}(\tau, f) = \frac{|f|}{2\pi} \int_{-\infty}^{\infty} z(t)\, e^{-\left((t-\tau)^2/2\sigma^2\right)} e^{-j2\pi f\tau} dt, \tag{1}$$

where $j = \sqrt{-1}$, t, and τ are the time variables, f denotes the frequency, and $\sigma = 1/|f|$ is the scale factor. The output of ST is the complex-valued matrix whose rows and columns are related to time and frequency, respectively. On the other hand, assume that the discrete signal $z[k]$, $k = 0, 1, \ldots, N-1$, is obtained from $z(t)$ by sampling. By replacing $\tau \longrightarrow k$ and $f \longrightarrow n/N$, the discrete ST for $z[k]$, $\mathbf{S}[k, n]$, for $n \neq 0$ is calculated as [20]

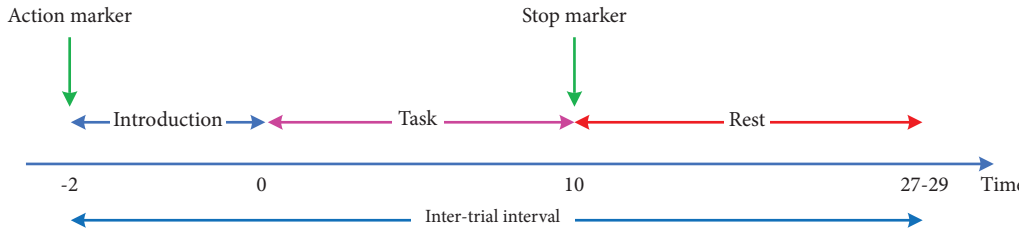

Figure 2: Experiment diagram.

$$\mathbf{S}[k,n] = \sum_{m=0}^{N-1} Z[m+n] e^{\left(2\pi^2 m^2/n^2\right)} e^{(j2\pi mk/N)}, \quad (2)$$

where $Z[n], n = 0, 1, \ldots, N-1$, is the DFT of $z[k]$. For $n = 0$, we have $\mathbf{S}[k,0] = 1/n\sum_{m=0}^{N-1} z[m]$, which equals to DC value of the Fourier transform. There are N^2 ST coefficients for a signal of length N. Computing each coefficient requires the computational complexity of the order (N), and hence total computational complexity is of order (N^3). Let $f(x,y)$ denote a 2D image, and its 2D ST is calculated as [21]

$$\mathbf{S}(u,v,f_u,f_v) = \frac{|f_u||f_v|}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) \cdot e^{\left((u-x)^2(v-y)^2/2\right)} e^{-j2\pi\left(f_u x + f_{vy}\right)} \mathrm{dxdy}, \quad (3)$$

where u and v are shift parameters used to move the Gaussian window on the x and y axes. Also, frequency parameters f_u and f_v are the frequencies related to shift parameters that control the spatial expansion of the window. $\mathbf{S}(u,v,f_u,f_v)$ is a 4D complex-valued matrix. The 2D-DOST of an $N \times N$ image, $f(x,y)$, is defined as follows [20]:

$$\mathbf{S}(u,v,f_u,f_v) = \frac{1}{\sqrt{2^{P_x+P_y-2}}} \sum_{m=-2^{P_x-2}}^{2^{P_x-2}-1} \sum_{n=-2^{P_y-2}}^{2^{P_y-2}-1} \cdot F\left(m+v_x, n+v_y\right) e^{j2\pi\left((mu/2^{P_x-1})+(nv/2^{P_y-1})\right)}, \quad (4)$$

where $v_x = 2^{P_x-1} + 2^{P_x-2}$ and $v_x = 2^{P_y-1} + 2^{P_y-2}$ are the horizontal and vertical frequencies, respectively, and $p_x, p_y = 0, 1, \ldots, \log(N-1)$. Also, $F(m,n)$ is the 2D Fourier transform of the image $f(x,y)$. It should be noted that the dimension of DOST points is equal to that of the input image. By integrating all the values p_x and p_y, a local spatial frequency range consisting of positive and negative frequency components from $(f_u, f_v) = 0$ to $(f_u, f_v) = (N/2, N/2)$ can be constructed. 2D-DOST provides information about frequencies in the bandwidth of $2^{P_x-1} \times 2^{P_y-1}$ frequencies [20].

2.3. CNN. CNN is one of the most efficient machine learning methods for feature extraction and classification of images. Figure 3 presents a typical CNN, and its main layers are explained in the following. Convolution layers scan the pixels using a kernel that passes over the image and create feature maps which are then used to predict the feature class. Due to the large amount of information obtained from the convolution layer, the pooling layer each time retains the important information and reduces redundant information. The fully connected layer acts similar to traditional MLP and predicts the output class using the extracted deep features. In this paper, two CNNs and their modified versions are used for deep feature extraction which is explained in the following.

MobileNet [22] is a class of efficient models used in mobile and embedded vision applications. The number of parameters is significantly reduced because of using separable convolutions in this model when compared to the network with regular convolutions with the same depth. In contrast to the standard convolution combination, in which the combination and filtering are done simultaneously in the same stage, in these networks, by using the ability of deep separation, in one stage, the filter operation is performed and then the combination operation is performed on the other stage. This separation has a strong efficiency in reducing computational complexity. The structure of MobileNet is given in Table 1, where conv. and conv. dw denote the standard and depthwise convolutions, respectively.

The LeNet-5 architecture was introduced in [23]. It is one of the earliest and most basic CNN architectures consisting of seven layers. The first layer consists of an input image with a size of 32×32. It is convolved with six filters of size 5×5 resulting in a dimension of $28 \times 28 \times 6$. The second layer is a pooling operation with a filter size of 2×2 and a stride of two. Hence, the resulting image dimension will be $14 \times 14 \times 6$. Similarly, the third layer also involves a convolution operation with 16 filters of size 5×5 followed by a fourth pooling layer with a similar filter size of 2×2 and stride of two. Thus, the resulting image dimension will be reduced to $5 \times 5 \times 16$. Once the image dimension is reduced, the fifth layer is a convolution with 120 filters with the size of 5×5. The sixth layer is a fully connected layer with 84 units. The final layer is a fully connected layer with ten neurons and a softmax activation function.

3. Proposed Method

Here, we explain the proposed method for fNIRS signal classification. The general overview of the proposed method is given in Figure 4.

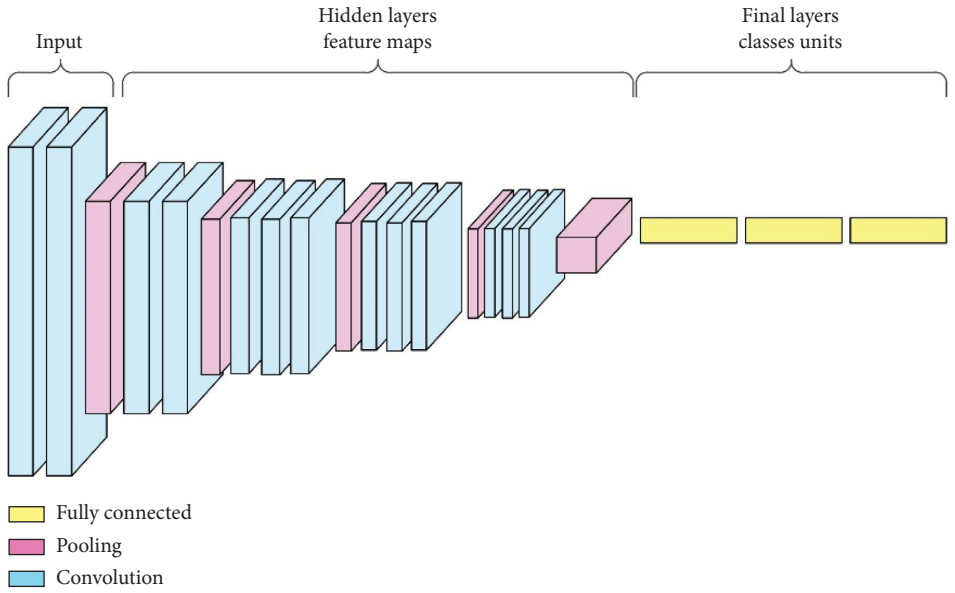

Figure 3: Structure of the typical CNN.

Table 1: The structure of MobileNet.

Layer type/stride	Filter shape	Input size
Conv./2	$3\times3\times3\times32$	$224\times224\times3$
Conv. dw/1	$3\times3\times32$ dw	$112\times112\times32$
Conv./1	$1\times1\times32\times64$	$112\times112\times32$
Conv. dw/2	$3\times3\times64$ dw	$112\times112\times64$
Conv./1	$1\times1\times64\times128$	$56\times56\times64$
Conv. dw/1	$3\times3\times128$ dw	$56\times56\times128$
Conv./1	$1\times1\times128\times128$	$56\times56\times128$
Conv. dw/2	$3\times3\times256$ dw	$56\times56\times128$
Conv./1	$1\times1\times128\times256$	$28\times28\times128$
Conv. dw/1	$3\times3\times256$ dw	$28\times28\times256$
Conv./1	$1\times1\times256\times256$	$28\times28\times256$
Conv. dw/2	$3\times3\times256$ dw	$28\times28\times256$
Conv./1	$1\times1\times256\times512$	$14\times14\times256$
5× Conv. dw/1	$3\times3\times512$ dw	$14\times14\times512$
5× Conv./1	$1\times1\times512\times512$	$14\times14\times512$
Conv. dw/2	$3\times3\times512$ dw	$14\times14\times512$
Conv./1	$1\times1\times512\times1024$	$7\times7\times512$
Conv. dw/2	$3\times3\times1024$ dw	$7\times7\times1024$
Conv./1	$1\times1\times1024\times1024$	$7\times7\times1024$
Average pooling/1	Pool 7×7	$7\times7\times1024$
Fully connected/1	1024×1000	$1\times1\times1024$
Softmax/1	Classifier	1000×1

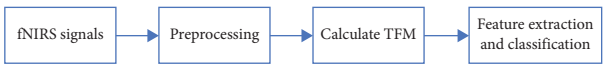

Figure 4: The steps of the proposed method for BCI based on fNIRS signals.

3.1. Preprocessing and Channel Selection. A third-order Butterworth filter with a passband of [0.01, 0.1] Hz was utilized to preprocess the recorded signals. This frequency range is useful for fNIRS signals due to its relevance to the hemodynamic response in the brain and the characteristics of the signals themselves. The hemodynamic response in the brain, which is the basis for fNIRS measurements, has a slow time course. It is primarily driven by changes in cerebral blood flow and oxygenation levels. These changes occur over a longer time scale compared to fast neuronal activity. This frequency range captures the low-frequency oscillations associated with these hemodynamic changes, allowing for the detection and analysis of relevant brain activity. Additionally, fNIRS signals are susceptible to various sources of noise and artifacts, such as physiological processes, motion artifacts, or environmental interference such as DC deviations. This range helps to filter out high-frequency noise and focus on the slower hemodynamic fluctuations of interest. This range is commonly associated with the physiological processes and neural activity related to cognitive functions, making it a suitable range for studying brain responses in fNIRS-based experiments.

Then, the time segmentation from the beginning of work (for example, 0 seconds) was done on the values of HbO and HbR changes. The period includes HbO and HbR changes in three classes (RHT, LFT, FT) that were performed during 25 tests. The signal-to-noise ratio (SNR) is calculated as follows:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{P_s}{P_n}\right), \tag{5}$$

where P_s and P_n represent the power of filtered data (signal estimation) and unfiltered data (noise estimation), respectively. This procedure was done for all channels, and several channels with the highest SNR were used for feature extraction considering the classification scenario.

3.2. TFM Calculation. For each of the HbO and HbR signals, the data obtained from 20 channels include 133 samples. The number of rows and columns of 2D-DOST must be a power of two, and hence four, eight, and sixteen channels with 128

samples can be considered for obtaining the TFM of HbO and HbR signals, and the channels were selected based on the SNR criteria. An example of the HbO and HbR signals considering the top four channels is shown in Figures 5–7. Also, some TFMs considering the top four, eight, and 16 channels for different MI are shown in Figures 8–10. It is observed that TFMs of different MI are different from each other; hence, they can be used for classification.

3.3. Feature Extraction and Classification. Here, we explain the procedures considered for feature extraction and classification after obtaining TFMs. Let $\mathbf{S}_O$ and $\mathbf{S}_R$, respectively, denote the 2D-DOST of HbO and HbR signals. In this paper, we consider four TFMs including $\mathbf{S}_O, \mathbf{S}_R, \mathbf{S}_O + \mathbf{S}_R$, and $\mathbf{S}_O - \mathbf{S}_R$. To use the information of these TFMs for classification, we considered three fusion schemes, including early fusion, joint fusion, and late fusion [20]. The MobileNet and LeNet-5 are considered as base structures and we modify them based on the fusion scheme and the size of TFMs. The additive combination of $\mathbf{S}_O$ and $\mathbf{S}_R$, denoted as $\mathbf{S}_O + \mathbf{S}_R$, captures the additive information from both HbO and HbR signals. This fusion scheme allows for the integration of complementary information from these two sources, potentially enhancing the discriminative power of the features. In a classification scheme, this combined feature can provide a more comprehensive representation of the underlying neural activity by considering both oxygenation and deoxygenation dynamics simultaneously. The differential combination $\mathbf{S}_O - \mathbf{S}_R$ represents the difference between HbO and HbR signals. This differential information can highlight variations in oxygenation patterns that may be critical for distinguishing between different cognitive or motor tasks. In a classification context, this feature can be particularly useful when changes in the balance between oxygenated and reduced hemoglobin are relevant to the classification task. In summary, the rationale for using $\mathbf{S}_O + \mathbf{S}_R$ and $\mathbf{S}_O - \mathbf{S}_R$ lies in their ability to provide a more comprehensive view of the hemodynamic responses by considering both additive and differential aspects of the HbO and HbR signals. These combined features may capture unique patterns that are relevant to the classification scheme, potentially improving the accuracy and discriminative power of the classification model.

3.3.1. Early Fusion. In early fusion, the fusion operation is performed at the feature level. The inputs contain the main features or are extracted as features from different ways [20]. They join together and form the final feature maps before feeding into a machine learning model. Based on the considered channels for computing 2D-DOST, each TFM has the size of $n_{ch} \times 128$. The considered early fusion merges the four TFMs to construct the CNN input as follows:

$$\mathbf{S}_{\text{early}} = \begin{bmatrix} \mathbf{S}_o \\ \mathbf{S}_R \\ \mathbf{S}_o + \mathbf{S}_R \\ \mathbf{S}_o - \mathbf{S}_R \end{bmatrix}. \tag{6}$$

It should be noted that the size of the matrix $\mathbf{S}_{\text{early}}$ is $4 \times n_{ch} \times 128$. The procedure of classification with early fusion is shown in Figure 11. In this procedure, the CNN is trained considering the common training algorithms.

3.3.2. Joint Fusion. The early fusion scheme concatenates TFMs at first and then extracts the deep features using one CNN. The procedure of joint fusion is shown in Figure 12. In contrast to early fusion, this scheme passes each TFM through a CNN and obtains the deep features for each TFM. Let $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$, and $\mathbf{x}_4$ denote the deep feature vectors corresponding to the inputs $\mathbf{S}_O$, $\mathbf{S}_R$, $\mathbf{S}_O + \mathbf{S}_R$, and $\mathbf{S}_O - \mathbf{S}_R$, respectively. These vectors are obtained from the flatten layer of CNNs and the structure of CNNs does not contain the fully connected layers. Since the structure of CNN is the same for all inputs, all feature vectors have the same number of features. These vectors are then concatenated to form the final feature vector $\mathbf{x}_f$ (feature fusion block). The vector $\mathbf{x}_f$ is given to the classifier to predict the output class. The classifier is the traditional MLP, and the structure is the same as fully connected layers of considered CNN. In the training process, the parameters of all CNNs and classifiers are tuned simultaneously.

3.3.3. Late Fusion. The late fusion scheme, which is known as a combination at the decision level, utilizes one CNN to predict the output for each TFM separately as shown in Figure 13. In this scheme, the size of the feature vector given to dense layers is smaller than the joint fusion, but this scheme does not consider the possible correlation among deep features of TFMs. For the final decision, depending on the situation, different methods such as majority vote, averaging, weighted voting, or meta-classification based on model predictions are used. Let the vectors $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3$, and $\mathbf{p}_4$ contain the prediction scores of different classes assigned by each four CNNs, respectively. It should be noted that all vectors have the size of $n_c \times 1$, where n_c is the number of classes. The aggregation scheme calculates the sum of prediction scores to obtain the final score, $\mathbf{p}_{\text{agg}}$, as

$$\mathbf{p}_{\text{agg}}(k) = \sum_{i=1}^{4} \mathbf{p}_i(k), k = 1, \ldots, n_c. \tag{7}$$

Finally, the predicted class $\widehat{y}_{\text{pred}}$ is the one that maximizes $\mathbf{p}_{\text{agg}}$ as

$$\widehat{y}_{\text{pred}} = \text{argmax}\left(\mathbf{p}_{\text{agg}}\right). \tag{8}$$

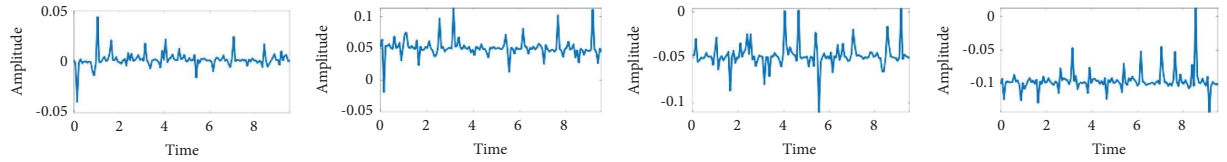

Figure 5: The HbO and HbR channels of RHT.

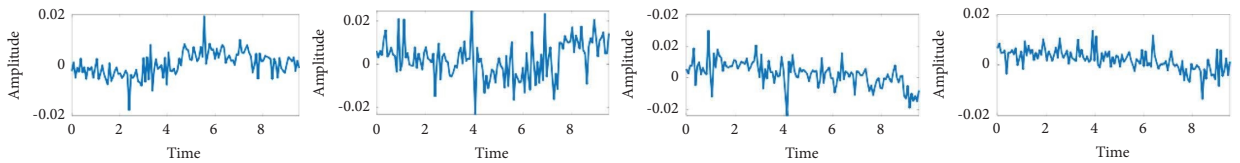

Figure 6: The top four HbO channels of LHT.

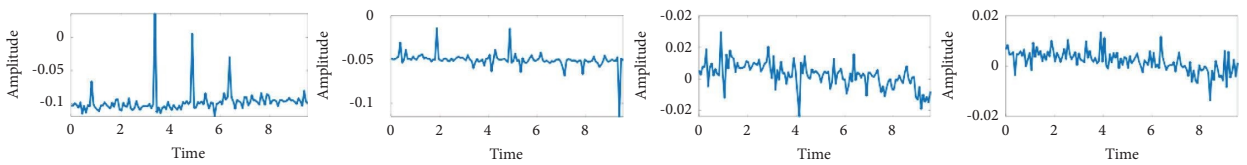

Figure 7: The top four HbO channels of FT.

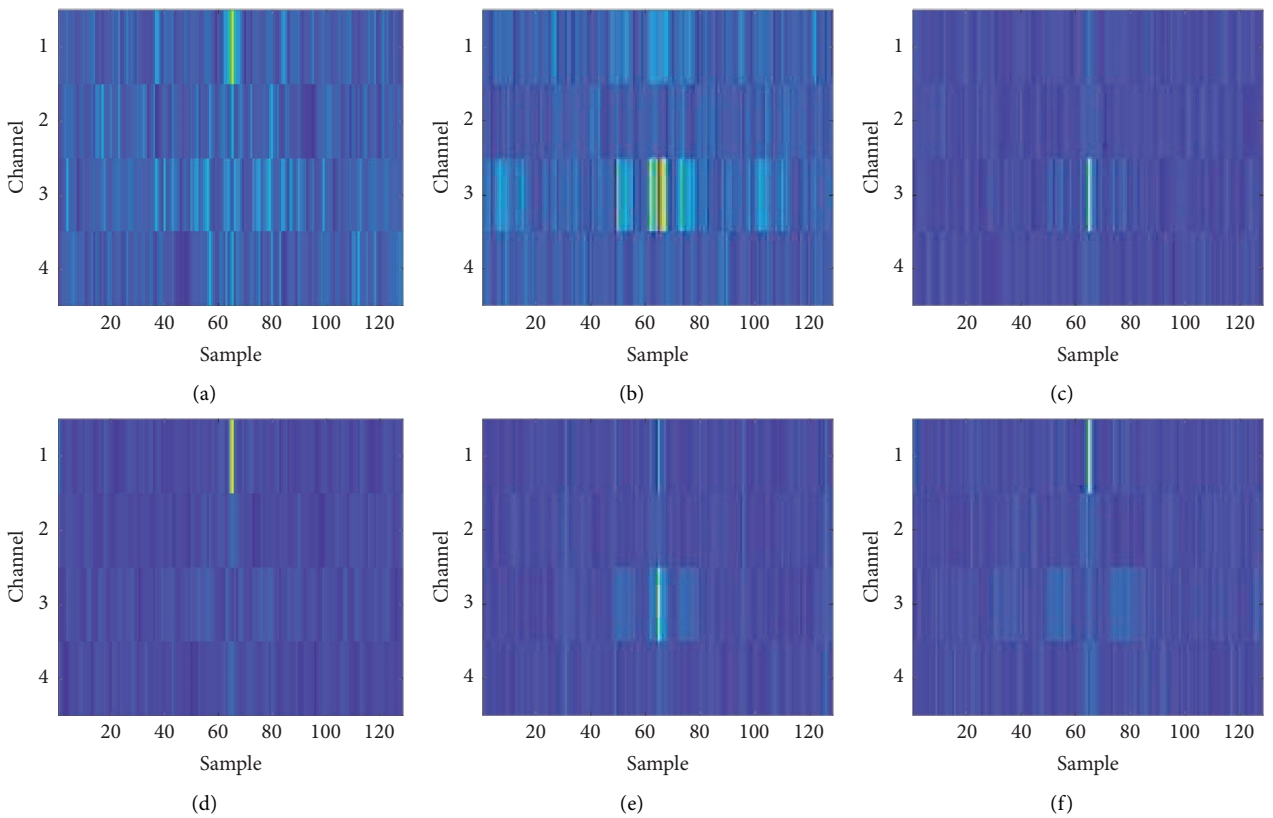

Figure 8: TFMs of different signals considering the top four channels. (a) HbO of FT. (b) HbO of LHT. (c) HbO of RHT. (d) HbR of FT. (e) HbR of LHT. (f) HbO of RHT.

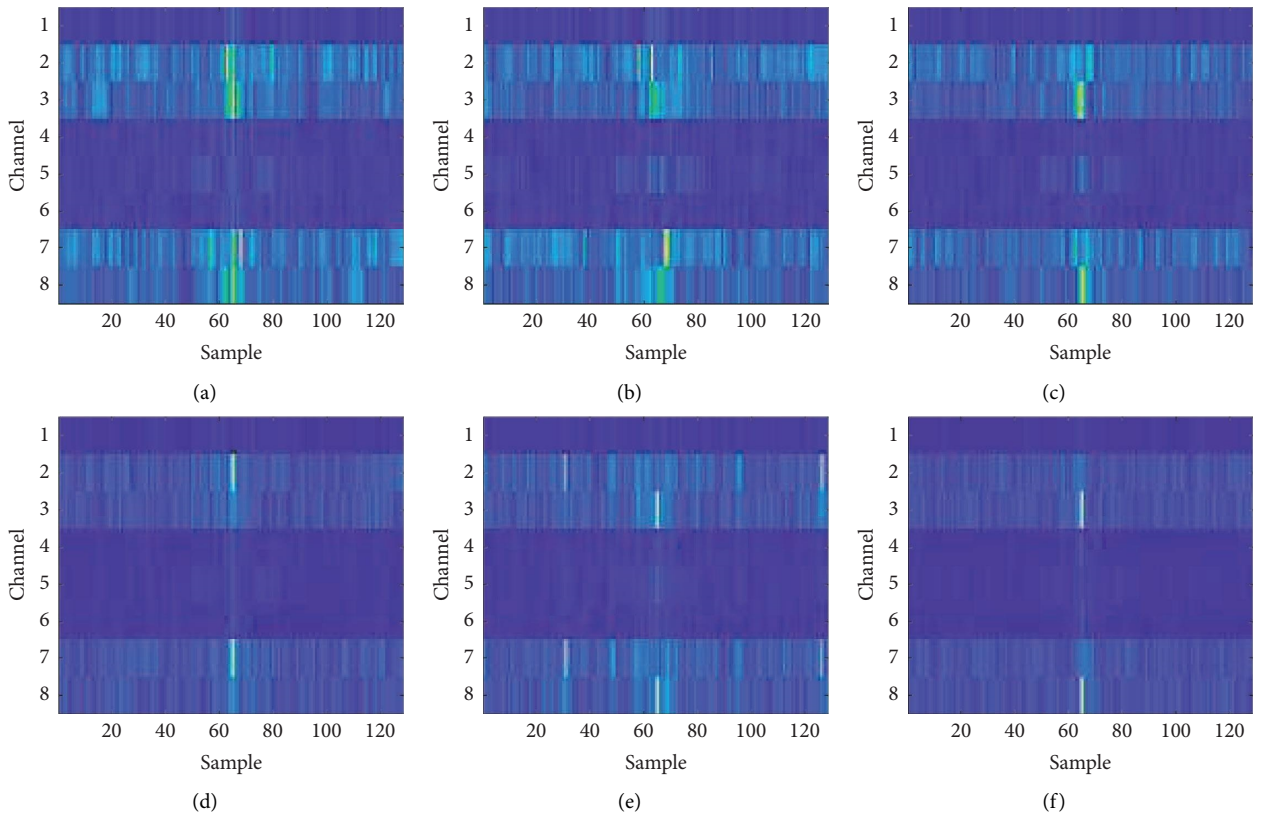

Figure 9: TFMs of different signals considering the top eight channels. (a) HbO of FT. (b) HbO of LHT. (c) HbO of RHT. (d) HbR of FT. (e) HbR of LHT. (f) HbO of RHT.

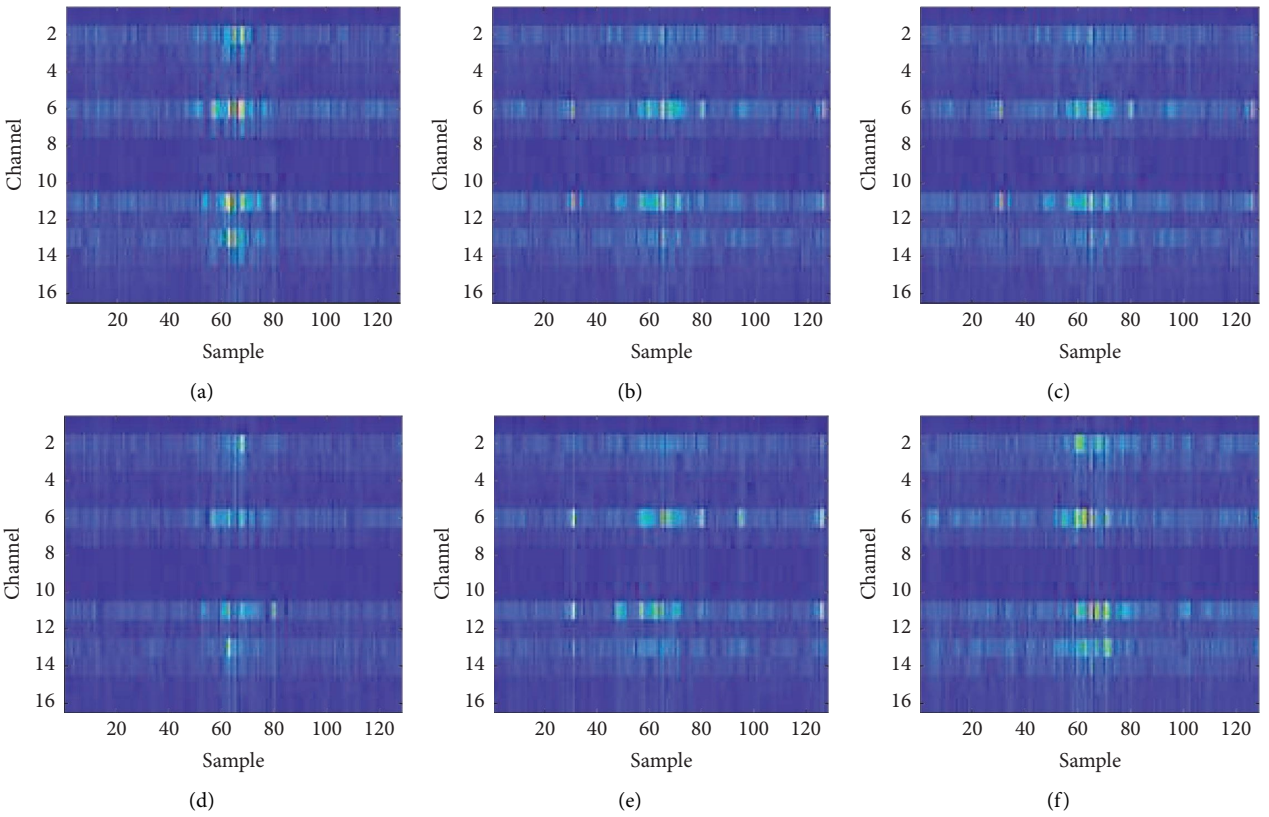

Figure 10: TFMs of different signals considering the top 16 channels. (a) HbO of FT. (b) HbO of LHT. (c) HbO of RHT. (d) HbR of FT. (e) HbR of LHT. (f) HbO of RHT.

Figure 11: The procedure of classification with early fusion.

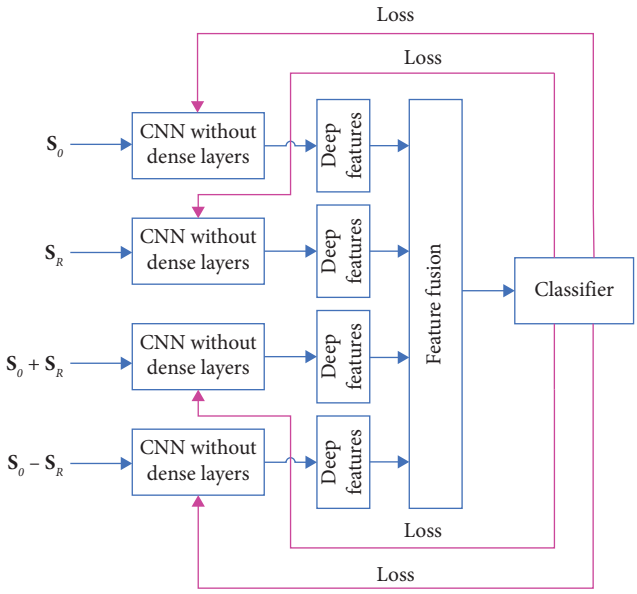

Figure 12: The procedure of joint fusion.

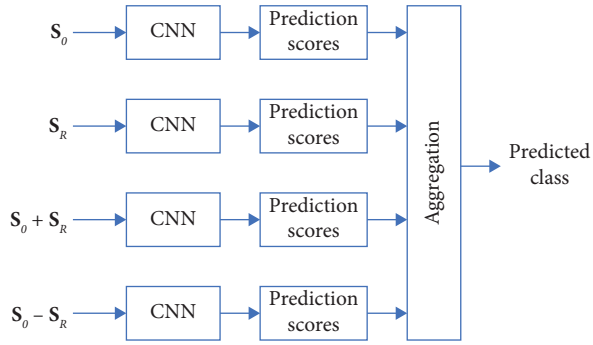

Figure 13: The procedure of late fusion.

4. Results

4.1. Simulation Setup. The three two-class scenarios and one three-class scenario are considered for classification as follows for distinct purposes: (RHT, LHT), (RHT, FT), (LHT, FT), and (RHT, LHT, FT). The three-class scenario enables us to capture intricate nuances in our data, allowing differentiation between multiple states or activities. Simultaneously, binary scenarios address specific research questions with simpler distinctions. This dual approach provides versatility, accommodating a wide range of research objectives and allowing for comparative analysis of classifier performance. Overall, it enriches our research by offering a comprehensive exploration of our dataset, catering to both complex and focused research questions.

The performance of LeNet and MobileNet is obtained for each scenario for the different number of channels and fusion schemes. Also, the structure of the modified CNNs yielding the highest accuracy is presented. In this paper, subject-independent classification is performed. Hence, train and test data were determined by the cross-subject validation protocol. This protocol trains the model with data from 29 subjects, and data from one subject evaluate the test accuracy of the model. This procedure is repeated for each subject as test data and average results are reported. Considering 25 trials for each task per subject, there are 750 signals from each task; hence, there are 725 and 25 signals from each task for training and testing, respectively. It should be mentioned that the data augmentation proposed in [24] is utilized to increase the number of training signals. Table 2 contains the parameters used for training CNN. The learning rate balances the convergence to the optimal solution and stability. Regularization parameter controls the overfitting and encourages generalization. The maximum number of epochs effectively updates the model without overfitting. Batch size balances the training speed by parallel processing and computational complexity. The momentum enhances the convergence and escape from local minima. Learning rate drop factor and drop period are used for fine-tuning the learning rate for effective convergence. SGDM optimizer combines the benefits of stochastic gradient descent with momentum. Cross-entropy loss function is suitable for classification tasks, measuring dissimilarity between predicted and actual class distributions.

The performance of the proposed method is presented in terms of confusion matrix, accuracy (Acc.), sensitivity (Sens.), precision (Prec.), Kappa score (K_p), and F_1-score, which are calculated as follows [25]:

$$\begin{aligned}
\text{Acc.} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\
\text{Sens.} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\
\text{Prec.} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\
K_p &= \frac{\text{Acc} - A_r}{1 - A_r} = \frac{\text{Acc} - 1/n_c}{1 - 1/n_c}, \\
F_1 &= 2\frac{\text{Prec} \times \text{Sens}}{\text{Prec} + \text{Sens}},
\end{aligned} \tag{9}$$

where the true positive (TP) and true negative (TN), respectively, denote the number of correctly classified and rejected fNIRS signals. Also, the false positive (FP) and false negative (FN), respectively, denote the number of incorrectly identified and incorrectly rejected fNIRS signals. Also, $A_r = 1/n_c$ is the random accuracy, where n_c is the number of classes.

4.2. Channel Selection. As mentioned in Section 2, the fNIRS signals were recorded in 20 channels and our criterion for channel selection was SNR. Since the number of rows and

Table 2: Parameters used for training.

Parameter	Value
Learning rate	0.001
Regularization parameter	0.002
Maximum number of epochs	100
Batch size	64
Momentum	0.85
Learning rate drop factor	0.2
Learning rate drop period	20
Optimizer	Stochastic gradient descent with momentum (SGDM)
Loss function	Cross-entropy

columns of input of 2D-DOST should be a power of two, we consider the four, eight, and 16 top channels based on SNR value for ternary and binary classification. To this end, the SNR of channels was sorted in descending order and most repetitive channels among all subjects are considered. Figure 14 demonstrates the repetition of high-SNR channels among subjects. The selected channels are also given in Table 3.

As given in [7], the motor cortex regions in contralateral hemispheres were well activated when the subjects perform finger-tapping and distinct HbR values were observed at channels 5, 6, 15, and 16 located in the anterior areas of C3 and C4. According to Table 3, these channels are among high-SNR ones.

4.3. Accuracy of LeNet. Table 4 presents the accuracy of LeNet and its modified version, accuracy in parentheses, in different scenarios for different number of channels and fusion schemes. As observed, the modified version reaches a higher accuracy than the original structure. Also, the binary scenarios have higher accuracy than the three-class scenario. It is observed that, in general, increasing the number of channels enhances the classification accuracy. Increasing the number of channels provided more information about brain activity, and hence classification accuracy increased. On the other hand, computational complexity increases. From fusion schemes, the joint fusion that extracts deep features from each TFM separately yields the highest accuracy, and the early fusion outperforms the late fusion. Since the joint fusion scheme concatenates the four vectors of deep features, it has a higher complexity compared to other fusion schemes. The three-class scenario reaches the highest accuracy of 90.71%. Also, the scenarios (RHT, LHT), (RHT, FT), and (LHT, FT) have the highest accuracy of 95.72%, 94.88%, and 93.19%, respectively. Also, the standard deviation of classification accuracies obtained in cross-validation is given for modified network. The smaller values of standard deviations depict the generalization of the proposed method.

Table 5 presents the structure of modified LeNet used for feature extraction and classification in joint fusion. The input layer passes the input TFM with the size of $128 \times 16 \times 1$ to the first convolution layer. This structure for classification consists of two convolutions, two average pooling layers, and one flatten layer. Each CNN generates the deep feature vector with the size of 600×1, and considering four CNNs for feature extraction, according to Figure 4, the input of the first fully connected layer is $(4 \times 600) \times 1$. The last fully connected layer acts as the output layer, and its output has the size of $n_c \times 1$.

4.4. Accuracy of MobileNet. The accuracy of MobileNet for different structures is given in Table 6. It is observed that the modified structure yields a higher accuracy than the original structure. The accuracy of the proposed method for the three-class scenario is 93.02%. Also, the accuracy for the scenarios (RHT, LHT), (RHT, FT), and (LHT, FT) is 98.73%, 96.67%, and 95.65%, respectively. These accuracies are obtained in joint fusion with the top 16 channels. As observed, similar to LeNet, the proposed method has a higher accuracy for two-class scenarios compared to the three-class scenario. Comparing Tables 4 and 6, the standard deviations of modified MobileNet are lower than those of the modified LeNet.

The structure of modified MobileNet yielding the highest accuracy in the joint fusion scenario is given in Table 7. The size of the input layer is 128×16. Each Conv. layer is a standard convolutional layer with batch normalization and rectified linear unit (ReLU). Also, the Conv. dw denotes the depthwise separable convolutions with depthwise and pointwise layers followed by normalization and ReLU.

4.5. Confusion Matrix. The confusion matrices of the proposed method with joint fusion and 16 channels for different classification scenarios are given in Tables 8–11. It is observed that in the three-class scenario, the RHT and LHT signals have higher sensitivity than the FT, while in binary scenarios (RHT, FT) and (LHT, FT), the FT has higher sensitivity than the RHT and LHT. Also, in both three-class and binary (RHT, LHT) scenarios, the RHT has higher sensitivity than the LHT. The sensitivity values for all signals in all classes are higher than 94%, except the FT in the three-class scenario.

4.6. Effect of Data Augmentation. The CNNs require more data for training compared to traditional artificial neural networks to avoid issues such as overfitting and underfitting and increase the training accuracy and generalization. As mentioned in this paper, the method based on WGANs proposed in [24] is employed for data augmentation. This network consists of two parts: critic and generator. The former learns the structure of data and the latter generates the artificial data, and both were configured as fully connected feedforward neural network with three layers [24] as given in Table 12, where N_{time} and N_{ch} represent the number of the time samples (=133) and channels (depends on the number of used high-SNR channels), respectively. A bias term was also added to the input and hidden layers. Random

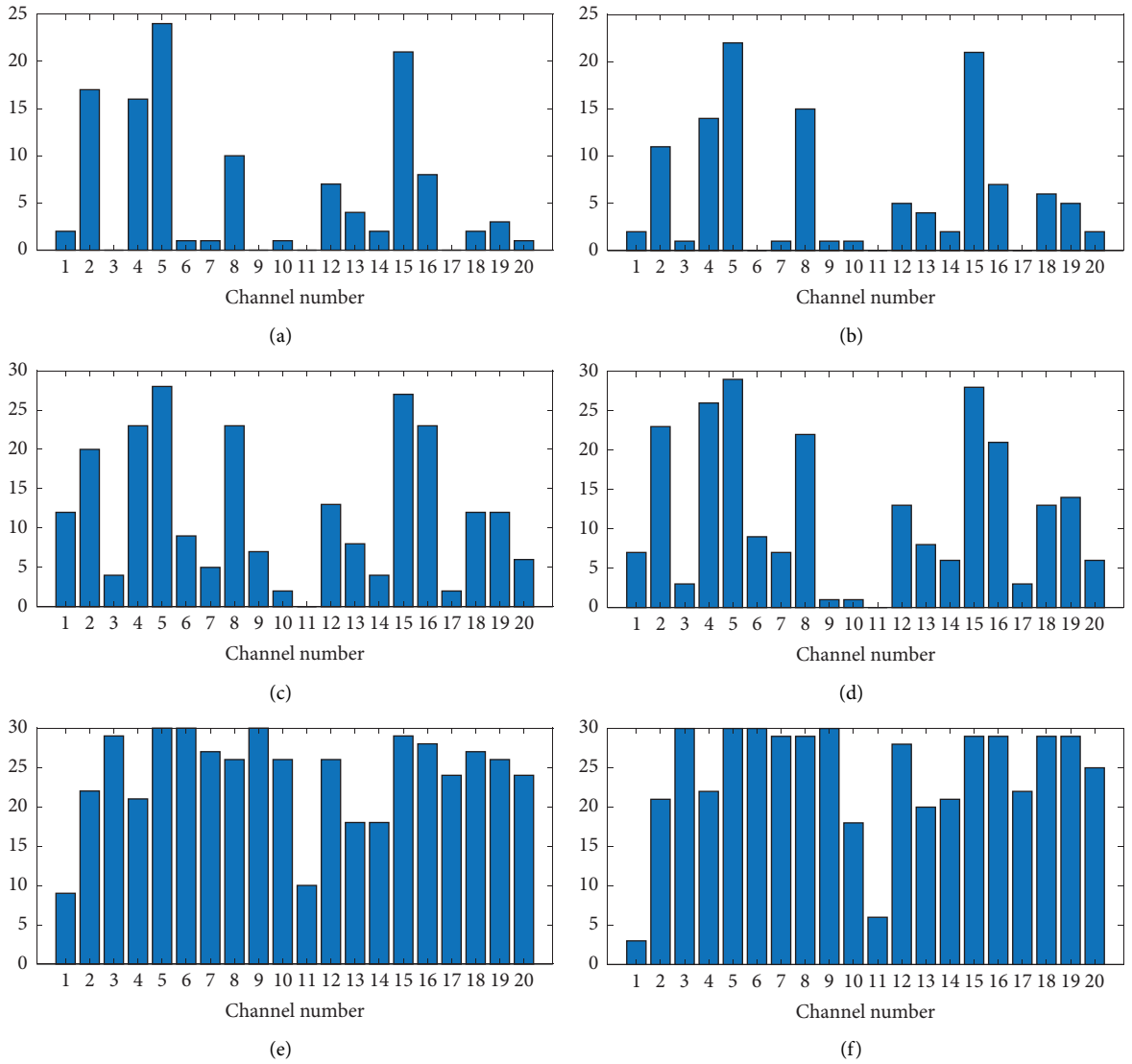

Figure 14: The number of repetition of high-SNR channels among all subjects. (a) HbO signal, top four channels. (b) HbR signal, top four channels. (c) HbO signal, top eight channels. (d) HbR signal, top eight channels. (e) HbO signal, top 16 channels. (f) HbR signal, top 16 channels.

Table 3: The selected channels based on SNR values.

Number of selected channels	Selected channels	
	HbO	HbR
4	5, 15, 2, 4	5, 15, 4, 2
8	5, 15, 4, 8, 16, 2, 12, 1	5, 15, 4, 2, 8, 16, 19, 12
16	5, 6, 9, 3, 15, 16, 7, 18, 8, 10, 12, 19, 17, 20, 2, 4	3, 5, 6, 9, 7, 8, 15, 16, 18, 19, 12, 20, 4, 17, 2, 14

numbers sampled from a uniform distribution in the range [−1, +1] represented input to the generator z that was a vector with N_z dimension (=100) [24].

As mentioned, there are 725 training samples at each cross-validation which are used for training critic network and the accuracy was reported for different number of generated samples. The effect of the number of augmented signals per training signal on the accuracy of different scenarios is shown in Figure 15. It is observed that for the lower number of augmented samples, the accuracy is low, and increasing the number of augmented samples increases the accuracy for all scenarios considering the different number of channels. Hence, using data augmentation is necessary to train the model for the classification of motor fNIRS signals based on deep learning.

4.7. Performance Comparison. Table 13 compares the performance of the proposed method with recently introduced ones on the considered dataset to demonstrate the efficiency

Table 4: The accuracy of LeNet in different scenarios.

Scenario	Number of channels	Fusion scheme		
		Early	Late	Joint
(RHT, LHT, FT)	4	75.27 (77.39 ± 4.16)	71.64 (73.03 ± 4.53)	77.57 (79.49 ± 4.02)
	8	80.26 (81.95 ± 3.41)	77.82 (80.68 ± 3.75)	83.85 (86.14 ± 3.24)
	16	84.74 (86.83 ± 2.26)	83.29 (85.59 ± 2.37)	87.95 (90.71 ± 2.08)
(RHT, LHT)	4	82.49 (84.72 ± 3.03)	77.79 (80.08 ± 3.22)	83.57 (85.36 ± 2.96)
	8	85.09 (88.11 ± 2.51)	85.14 (86.63 ± 2.78)	90.67 (92.09 ± 2.29)
	16	92.16 (93.64 ± 2.08)	89.07 (91.31 ± 2.43)	93.54 (95.72 ± 1.91)
(RHT, FT)	4	80.18 (82.90 ± 3.19)	77.26 (79.77 ± 3.35)	82.86 (85.17 ± 3.11)
	8	85.37 (87.15 ± 2.77)	83.87 (85.35 ± 2.96)	89.93 (91.38 ± 2.47)
	16	90.48 (92.68 ± 2.28)	88.03 (90.10 ± 2.65)	92.26 (94.88 ± 2.06)
(LHT, FT)	4	79.85 (81.82 ± 3.29)	76.58 (78.19 ± 3.48)	82.08 (84.65 ± 3.13)
	8	84.97 (86.32 ± 2.91)	81.17 (83.64 ± 3.15)	87.61 (90.81 ± 2.52)
	16	88.13 (90.18 ± 2.59)	86.58 (88.90 ± 2.85)	91.05 (93.19 ± 2.23)

The accuracy of modified LeNet is given in parentheses.

Table 5: The structure of modified LeNet used in joint fusion.

Operation	Layer	Input	Filter size	Output
Feature extraction	Convolution	$128 \times 16 \times 1$	$5 \times 5 \times 10$	$124 \times 12 \times 10$
	Average pooling	$124 \times 12 \times 10$	Pool 2×2	$62 \times 6 \times 10$
	Convolution	$62 \times 6 \times 10$	$3 \times 3 \times 10$	$60 \times 4 \times 10$
	Average pooling	$60 \times 4 \times 10$	Pool 2×2	$30 \times 2 \times 10$
	Flatten	$30 \times 2 \times 10$	—	600×1
Classification	Fully connected	2400×1	—	128×1
	Fully connected	128×1	—	64×1
	Fully connected	64×1	—	$n_c \times 1$

Table 6: The accuracy of MobileNet in different scenarios.

Scenario	Number of channels	Fusion scheme		
		Early	Late	Joint
(RHT, LHT, FT)	4	79.01 (80.26 ± 3.38)	75.46 (77.29 ± 4.15)	81.08 (83.41 ± 3.84)
	8	82.23 (83.75 ± 2.62)	81.90 (83.03 ± 3.44)	87.74 (89.53 ± 3.09)
	16	88.05 (88.93 ± 2.14)	86.36 (88.35 ± 2.19)	91.12 (93.02 ± 1.95)
(RHT, LHT)	4	85.49 (86.93 ± 2.72)	80.92 (83.49 ± 3.13)	86.02 (88.73 ± 2.72)
	8	88.91 (89.93 ± 2.36)	87.26 (89.07 ± 2.56)	92.82 (94.09 ± 2.14)
	16	94.01 (95.03 ± 1.95)	91.91 (93.41 ± 2.28)	97.15 (98.73 ± 1.41)
(RHT, FT)	4	83.29 (84.83 ± 3.08)	80.26 (81.99 ± 3.11)	85.93 (87.25 ± 3.05)
	8	87.54 (88.28 ± 2.34)	86.05 (87.26 ± 2.77)	92.06 (94.01 ± 2.26)
	16	92.71 (93.82 ± 1.87)	90.98 (92.18 ± 2.49)	95.65 (96.67 ± 1.98)
(LHT, FT)	4	82.17 (83.25 ± 3.12)	79.15 (81.01 ± 3.24)	85.19 (87.25 ± 3.09)
	8	87.90 (89.16 ± 2.68)	84.09 (85.17 ± 3.02)	91.01 (92.13 ± 2.33)
	16	90.88 (92.09 ± 2.16)	89.91 (91.63 ± 2.47)	93.87 (95.66 ± 2.12)

The accuracy of the modified MobileNet is given in parentheses.

of the proposed method for the classification of motor execution fNIRS signals. The results indicate that the proposed method outperforms the recently introduced ones. The average changes of HbO and HbR signals with a length of five seconds were calculated as features in [26] and then classified by Bayesian neural networks. The scenarios of (RHT, FT) and (LHT, FT) were considered, and the maximum accuracy of 86.44% was obtained. The difference of HbO and HbR changes as well as vector size and angle are considered as features of fNIRS signals in [27] and then are classified by LDA. The maximum accuracy of 98.7% and 85.4% was obtained for two- and three-class scenarios, respectively. In [7], the features were average changes of HbO and HbR concentrations and the maximum accuracy of 84.4% and 70.4% was obtained for two- and three-class scenarios, respectively. A method based on the transformer self-attention mechanism was introduced in [28]. To enhance data utilization and network representation, this method leverages spatial and channel representations of fNIRS signals. The results show that the method yields the maximum accuracy of 75.49% for three-class scenario. The authors in [29] designed fNIRSnet considering the inherent

Table 7: The structure of modified MobileNet used in joint fusion.

Layer input		Input	Filter size/stride	Output
Conv.		$128 \times 16 \times 1$	$3 \times 3 \times 32/2 \times 1$	$64 \times 16 \times 32$
Conv. dw		$64 \times 16 \times 32$	$3 \times 3 \times 32/1 \times 1$	$64 \times 16 \times 32$
Conv.		$64 \times 16 \times 32$	$32 \times 1 \times 64/1 \times 1$	$64 \times 16 \times 64$
Conv. dw		$64 \times 16 \times 64$	$3 \times 3 \times 64/2 \times 1$	$32 \times 16 \times 64$
Conv.		$32 \times 16 \times 64$	$32 \times 1 \times 128/1 \times 1$	$32 \times 16 \times 128$
Conv. dw		$32 \times 16 \times 128$	$3 \times 3 \times 128/1 \times 1$	$32 \times 16 \times 128$
Conv.		$32 \times 16 \times 128$	$32 \times 1 \times 128/1 \times 1$	$32 \times 16 \times 128$
Conv. dw		$32 \times 16 \times 128$	$3 \times 3 \times 128/2 \times 1$	$16 \times 16 \times 128$
Conv.		$16 \times 16 \times 128$	$16 \times 1 \times 256/1 \times 1$	$16 \times 16 \times 256$
Conv. dw		$16 \times 16 \times 256$	$3 \times 3 \times 256/1 \times 1$	$16 \times 16 \times 256$
Conv.		$16 \times 16 \times 256$	$16 \times 1 \times 256/1 \times 1$	$16 \times 16 \times 256$
Conv. dw		$16 \times 16 \times 256$	$3 \times 3 \times 256/2 \times 2$	$8 \times 8 \times 256$
Conv.		$8 \times 8 \times 256$	$8 \times 1 \times 512/1 \times 1$	$8 \times 8 \times 512$
Conv. dw		$8 \times 8 \times 512$	$3 \times 3 \times 512/1 \times 1$	$8 \times 8 \times 512$
5×	Conv.	$8 \times 8 \times 512$	$8 \times 1 \times 512/1 \times 1$	$8 \times 8 \times 512$
	Conv. dw	$8 \times 8 \times 512$	$3 \times 3 \times 512/2 \times 2$	$4 \times 4 \times 512$
Conv.		$4 \times 4 \times 512$	$4 \times 1 \times 1024/1 \times 1$	$4 \times 4 \times 1024$
Conv. dw		$4 \times 4 \times 1024$	$3 \times 3 \times 1024/1 \times 1$	$4 \times 4 \times 1024$
Conv.		$4 \times 4 \times 1024$	$4 \times 1 \times 1024/1 \times 1$	$4 \times 4 \times 1024$
Average pooling		$4 \times 4 \times 1024$	Pool 4×4	$1 \times 1 \times 1024$
Flatten		$1 \times 1 \times 1024$	—	1024×1
Fully connected		1024×1000	—	1024×1
Softmax		Classifier	—	$n_c \times 1$

Table 8: Confusion matrix for three-class scenario.

		Predicted class			Sens. (%)	Prec. (%)	F_1-score (%)
		RHT (%)	LHT (%)	FT (%)			
Actual class	RHT	95.07	2.93	2	95.07	93.32	94.19
	LHT	2.27	94.13	3.60	94.13	91.69	92.89
	FT	4.53	5.60	89.87	89.87	94.13	91.95
Acc. = 93.02%, $K_p = 0.8604$							

Table 9: Confusion matrix for (LHT, RHT) scenario.

		Predicted class		Sens. (%)	Prec. (%)	F_1-score (%)
		RHT (%)	LHT (%)			
Actual class	RHT	98.80	1.20	98.80	98.67	98.74
	LHT	1.33	98.67	98.67	98.80	98.74
Acc. = 98.73%, $K_p = 0.9746$						

Table 10: Confusion matrix for (RHT, FT) scenario.

		Predicted class		Sens. (%)	Prec. (%)	F_1-score (%)
		RHT (%)	LHT (%)			
Actual Class	RHT	96.40	3.60	96.40	96.91	96.65
	LHT	3.07	96.93	96.93	96.42	96.67
Acc. = 96.67%, $K_p = 0.9334$						

delayed hemodynamic responses of fNIRS signals [30]. The local interpretable model-agnostic explanation (LIME) algorithm was proposed for the feature selection for fNIRS datasets in [31]. The Gramian angular difference field (GADF) was used to encode multichannel fNIRS signals into multichannel images.

Table 11: Confusion matrix for (LHT, FT) scenario.

		Predicted class		Sens. (%)	Prec. (%)	F_1-score (%)
		RHT (%)	LHT (%)			
Actual class	RHT	95.20	4.80	95.20	96.09	95.64
	LHT	3.87	96.13	96.13	95.24	95.68
Acc. = 95.66%, $K_p = 0.9132$						

Table 12: The structure of critic and generator networks used in this paper [24].

Layer	Network	
	Critic	Generator
Input	$N_{\text{time}} \times N_{ch} + 1$	$N_z + 1$
Hidden	$N_{\text{time}} + 1$	$N_{\text{time}} + 1$
Output	1	$N_{\text{time}} \times N_{ch}$

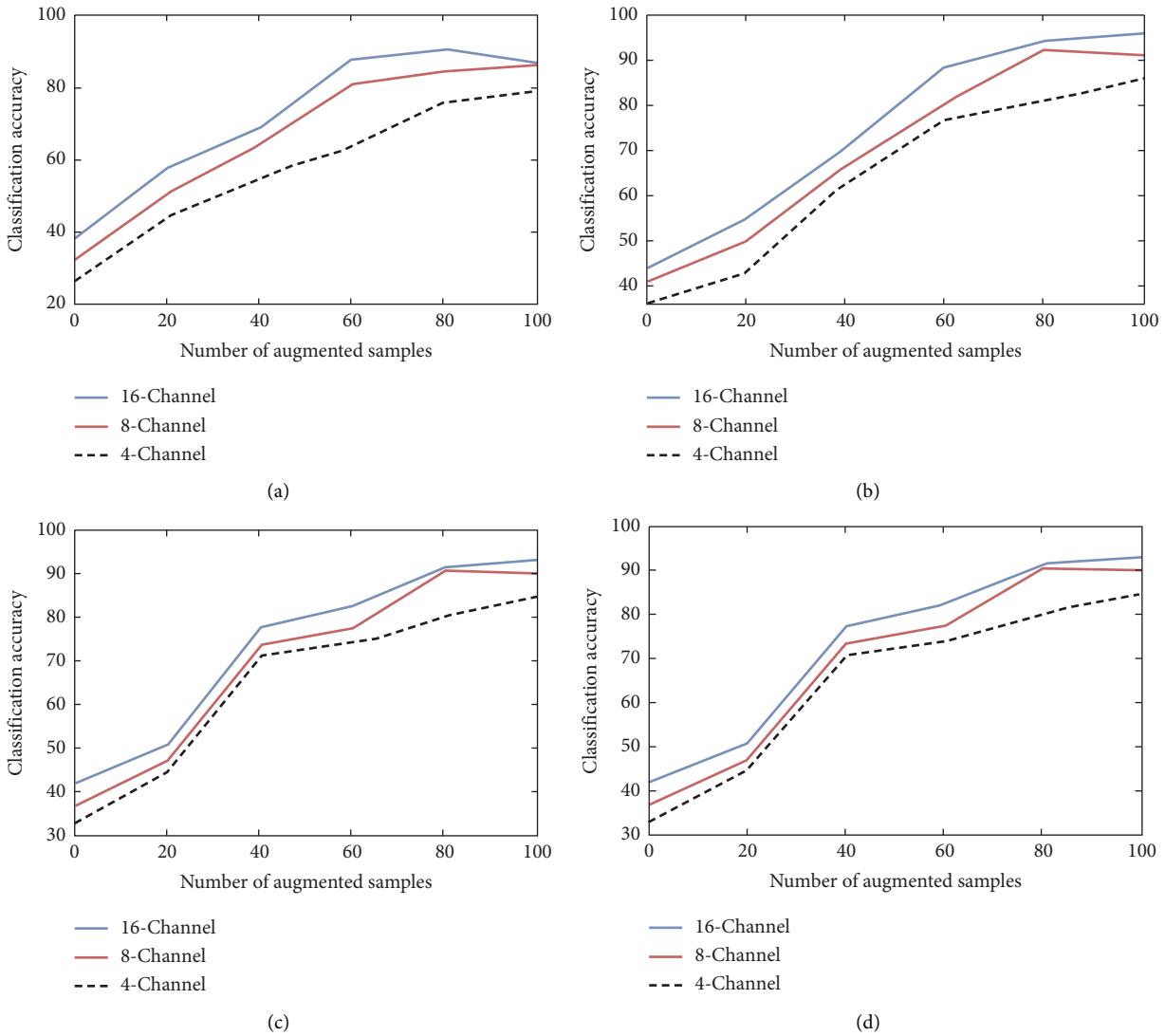

Figure 15: The effect of data augmentation on the accuracy. (a) RHT, LHT, FT. (b) RHT, LHT. (c) RHT, FT. (d) LHT, FT.

Table 13: Performance comparison between the proposed method and others.

Authors	Method	Accuracy
Siddique and Mahmud [26]	Average changes of HbO and HbR, Bayesian neural network	86.44% (2 classes)
Nazeer et al. [27]	Difference between HbO and HbR changes, LDA	98.7% (2 classes) 85.4% (3 classes)
Bak et al. [7]	Average changes of HbO and HbR concentrations, SVM	84.4% (2 classes) 70.4% (3 classes)
Wang et al. [28]	The transformer self-attention mechanism	75.49% (3 classes)
Wang et al. [29]	fNIRSnet	64.43% (3 classes)
Shin [30]	LIME, SVM	86.0% (2 classes)
Wang et al. [31]	GADF	78.22% (3 classes)
Proposed method	2D-DOST, feature fusion, CNN (five-fold cross-validation)	99.07% (2 classes) 93.60% (3 classes)
Proposed method	**2D-DOST, feature fusion, CNN (cross-subject cross-validation)**	**98.73% (2 classes)** **93.04% (3 classes)**

The bold values represent the maximum accuracy.

Also, the results of five-fold cross-validation protocol are given Table 13. As observed, this protocol outperforms the cross-subject cross-validation, while due to the following reasons, the cross-subject protocol is most popular than the k-fold one in BCI applications [32–34].

(1) Generalization to new users: Cross-subject cross-validation ensures that the model is tested on data from individuals who were not part of the training set. This helps assess the system's ability to generalize to new users, which is crucial for biomedical applications where the BCI needs to be applicable to a wide range of individuals.

(2) Real-world variability: Biomedical applications often involve real-world scenarios where users may exhibit individual differences, such as variations in brain anatomy, physiology, or cognitive processes. Cross-subject cross-validation allows for the evaluation of the BCI system's performance in capturing and adapting to these inter-subject variabilities. It provides a more realistic assessment of how the system will perform when deployed in diverse user populations.

(3) Avoiding data leakage: In some cases, k-fold cross-validation may lead to data leakage, where information from the test set inadvertently influences the training process. This can result in overly optimistic performance estimates. Cross-subject cross-validation helps mitigate this issue by ensuring that the training and testing data come from different individuals, reducing the risk of data leakage and providing more reliable performance estimates.

(4) Clinical relevance: Biomedical applications often require BCI systems to be evaluated in a clinical context, where the performance and reliability of the system are critical. Cross-subject cross-validation allows for a more rigorous evaluation of the BCI system's performance across different individuals, which is important for establishing its clinical relevance and potential utility in real-world healthcare settings.

5. Conclusion

In this paper, a new method for the classification of motor execution fNIRS signals was presented. The presented method is based on the joint fusion of TFMs of HbO, HbR, HbO + HbR, and HbO − HbR. The TFMs were obtained by 2D-DOST to simultaneously consider the correlation among samples of different channels as well as the samples of each channel. Joint fusion was considered to merge the deep features extracted from four TFMs using CNN. The open-access dataset with 20-channel fNIRS signals of three motor executions collected from 30 subjects was used for performance evaluation. The performance of LeNet, MobileNet, and their modified version was obtained for different number of top channels and scenarios. The results showed that increasing the number of channels increases the accuracy, and the proposed method reached the maximum accuracy of 98.73% and 93.04% for two-class and three-class scenarios, respectively, when modified MobileNet is used deep feature extraction and classification. Also, performance comparison showed that the proposed method outperforms the recently introduced methods.

Data Availability

Data are available online.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Pouya Khani was responsible for conceptualization, investigation, methodology, software, and original draft preparation. Vahid Solouk was responsible for conceptualization, methodology, validation, and supervision. Hashem Kalbkhani was responsible for conceptualization, software, visualization, and review and editing. Farid Ahmadi was responsible for methodology and review and editing.

References

[1] A. Zafar, K. D. Kallu, M. A. Yaqub et al., "A hybrid GCN and filter-based framework for channel and feature selection: an fNIRS-BCI study," *International Journal of Intelligent Systems*, vol. 2023, Article ID 8812844, 14 pages, 2023.

[2] C. Chen, Y. Wen, S. Cui et al., "A multichannel fNIRS system for prefrontal mental task classification with dual-level excitation and deep forest algorithm," *Journal of Sensors*, vol. 2020, Article ID 1567567, 10 pages, 2020.

[3] M. Ferrari and V. Quaresima, "A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application," *NeuroImage*, vol. 63, no. 2, pp. 921–935, 2012.

[4] N. Naseer, N. K. Qureshi, F. M. Noori, and K.-S. Hong, "Analysis of different classification techniques for two-class functional near-infrared spectroscopy-based brain-computer interface," *Computational Intelligence and Neuroscience*, vol. 2016, Article ID 5480760, 11 pages, 2016.

[5] G. Huve, K. Takahashi, and M. Hashimoto, "Brain activity recognition with a wearable fNIRS using neural networks," in *Proceedings of the 2017 IEEE international conference on mechatronics and automation (ICMA)*, pp. 1573–1578, IEEE, Takamatsu, Japan, August 2017.

[6] A. M. Chiarelli, P. Croce, A. Merla, and F. Zappasodi, "Deep learning for hybrid EEG-fNIRS brain–computer interface: application to motor imagery classification," *Journal of Neural Engineering*, vol. 15, no. 3, Article ID 036028, 2018.

[7] 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, 2019.

[8] M. S. B. A. Ghaffar, U. S. Khan, J. Iqbal et al., "Improving classification performance of four class FNIRS-BCI using Mel Frequency Cepstral Coefficients (MFCC)," *Infrared Physics and Technology*, vol. 112, Article ID 103589, 2021.

[9] A. Gulraiz, N. Naseer, H. Nazeer, M. J. Khan, R. A. Khan, and U. Shahbaz Khan, "LASSO homotopy-based sparse representation classification for fNIRS-BCI," *Sensors*, vol. 22, no. 7, p. 2575, 2022.

[10] B. Abibullaev and J. An, "Classification of frontal cortex haemodynamic responses during cognitive tasks using wavelet transforms and machine learning algorithms," *Medical Engineering and Physics*, vol. 34, no. 10, pp. 1394–1410, 2012.

[11] J. Cao, E. M. Garro, and Y. Zhao, "EEG/FNIRS based workload classification using functional brain connectivity and machine learning," *Sensors*, vol. 22, no. 19, p. 7623, 2022.

[12] Z. Wang, L. Yang, Y. Zhou et al., "Incorporating EEG and fNIRS patterns to evaluate cortical excitability and MI-BCI performance during motor training," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 2872–2882, 2023.

[13] Y. Li, X. Zhang, and D. Ming, "Early-stage fusion of EEG and fNIRS improves classification of motor imagery," *Frontiers in Neuroscience*, vol. 16, Article ID 1062889, 2022.

[14] T. Ma, S. Wang, Y. Xia et al., "CNN-based classification of fNIRS signals in motor imagery BCI system," *Journal of Neural Engineering*, vol. 18, no. 5, Article ID 056019, 2021.

[15] Y. Fu, R. Chen, A. Gong et al., "Recognition of flexion and extension imagery involving the right and left arms based on deep belief network and functional near-infrared spectroscopy," *Journal of Healthcare Engineering*, vol. 2021, Article ID 5533565, 11 pages, 2021.

[16] T. Hiroyasu, K. Hanawa, and U. Yamamoto, "Gender classification of subjects from cerebral blood flow changes using Deep Learning," in *Proceedings of the 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pp. 229–233, IEEE, Orlando, FL, USA, December 2014.

[17] R. Liu, B. Reimer, S. Song, B. Mehler, and E. Solovey, "Unsupervised fNIRS feature extraction with CAE and ESN autoencoder for driver cognitive load classification," *Journal of Neural Engineering*, vol. 18, no. 3, Article ID 036002, 2021.

[18] H. Hamid, N. Naseer, H. Nazeer, M. J. Khan, R. A. Khan, and U. Shahbaz Khan, "Analyzing classification performance of fNIRS-BCI for gait rehabilitation using deep neural networks," *Sensors*, vol. 22, no. 5, p. 1932, 2022.

[19] R. G. Stockwell, L. Mansinha, and R. Lowe, "Localization of the complex spectrum: the S transform," *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 998–1001, 1996.

[20] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *NPJ Digital Medicine*, vol. 3, no. 1, p. 136, 2020.

[21] R. S. Choraś, "Time-frequency analysis of image based on Stockwell transform," in *Image Processing and Communications Challenges 5*, pp. 91–97, Springer, Berlin, Germany, 2014.

[22] A. G. Howard, M. Zhu, B. Chen et al., "Mobilenets: efficient convolutional neural networks for mobile vision applications," 2017, https://arxiv.org/abs/1704.04861.

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[24] T. Nagasawa, T. Sato, I. Nambu, and Y. Wada, "fNIRS-GANs: data augmentation using generative adversarial networks for classifying motor tasks from functional near-infrared spectroscopy," *Journal of Neural Engineering*, vol. 17, no. 1, Article ID 016068, 2020.

[25] H. Kalbkhani and M. G. Shayesteh, "Stockwell transform for epileptic seizure detection from EEG signals," *Biomedical Signal Processing and Control*, vol. 38, pp. 108–118, 2017.

[26] T. Siddique and M. S. Mahmud, "Classification of fNIRS data under uncertainty: a Bayesian neural network approach," in *Proceedings of the 2020 IEEE International Conference on E-health Networking, Application and Services (HEALTHCOM)*, pp. 1–4, IEEE, Shenzhen, China, March 2021.

[27] H. Nazeer, N. Naseer, R. A. Khan et al., "Enhancing classification accuracy of fNIRS-BCI using features acquired from vector-based phase analysis," *Journal of Neural Engineering*, vol. 17, no. 5, Article ID 056025, 2020.

[28] Z. Wang, J. Zhang, X. Zhang, P. Chen, and B. Wang, "Transformer model for functional near-infrared spectroscopy classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 6, pp. 2559–2569, 2022.

[29] Z. Wang, J. Fang, and J. Zhang, "Rethinking delayed hemodynamic responses for fNIRS classification," *IEEE*

Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 4528–4538, 2023.

[30] J. Shin, "Feasibility of local interpretable model-agnostic explanations (LIME) algorithm as an effective and interpretable feature selection method: comparative fNIRS study," *Biomedical Engineering Letters*, vol. 13, no. 4, pp. 689–703, 2023.

[31] Z. Wang, J. Zhang, Y. Xia, P. Chen, and B. Wang, "A general and scalable vision framework for functional near-infrared spectroscopy classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1982–1991, 2022.

[32] K. Roots, Y. Muhammad, and N. Muhammad, "Fusion convolutional neural network for cross-subject EEG motor imagery classification," *Computers*, vol. 9, no. 3, p. 72, 2020.

[33] J. White and S. D. Power, "k-fold cross-validation can significantly over-estimate true classification accuracy in common EEG-based passive BCI experimental designs: an empirical investigation," *Sensors*, vol. 23, no. 13, p. 6077, 2023.

[34] A. Koul, C. Becchio, and A. Cavallo, "Cross-validation approaches for replicability in psychology," *Frontiers in Psychology*, vol. 9, p. 1117, 2018.