Biomedical Signal Processing and Control 70 (2021) 103021

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

A channel-mixing convolutional neural network for motor imagery EEG decoding and feature visualization

Weifeng Ma [a,*], Yifei Gong [a], Gongxue Zhou [a], Yang Liu [a], Lei Zhang [b], Boxian He [a]

[a] School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, 310023, People's Republic of China
[b] School of Biological and Chemical Engineering, Zhejiang University of Science and Technology, Hangzhou, 310023, People's Republic of China

ARTICLE INFO

Keywords:
Brain–computer interfaces (BCIs)
Channel mixing
Deep learning
Interpretability
Motor imagery

ABSTRACT

Convolutional Neural Network (CNN) has achieved great success in decoding EEG signals, decoders based on these architectures make separate feature extraction and classification into an integrated stage, however, a large number of trainable parameters introduced by the model hinder the improvement of EEG decoding performance and challenge the interpretability of decoding process used CNNs. In this paper, we propose an end-to-end shallow and lightweight CNN framework, which allows EEG-Motor Raw dataset as inputs, to boost decoding accuracy by the Channel-Mixing-ConvNet. The first block of network is designed in the way of implicitly stacking temporal–spatial convolution layers for learning temporal and spatial EEG features after EEG channels were mixed, compared to previously independently building a single temporal and spatial convolutional layer, this method combines the feature extraction capabilities of the two layers. The Mixed Channel Process block introducing a depthwise convolution layer is applied for a series of processing such as to decouple and supplement the internal and external mapping relationships existing in the mixed multi-dimensional EEG feature maps. Finally, the classification block is constructed to finish EEG decoding tasks. The lightweight architecture of Channel-Mixing-ConvNet leaves space for the model to exploit its potential performance by stacking other layers. In our experiments, the proposed Channel-Mixing-ConvNet and variants based on different hyper-parameters were evaluated on public EEG-motor datasets BCI-IV 2a and HGD respectively, Channel-Mixing-ConvNet outperformed state-of-the art (SOA) algorithms for EEG decoding. Additionally, via post-hoc interpretation techniques, the results show the learned features are consistent with the neurophysiological principle of the EEG motor imagery, meanwhile, the model also captures the remarkable features associated with channels.

1. Introduction

EEG signals are electric signals generated by cerebral cortical neurons and generally classified into two types: spontaneous and evoked [1]. Brain–computer interfaces (BCIs) based on two types can capture these signals and correctly translate them into executive instructions from the human body, which makes direct interaction between brain and external devices a reality [2,3]. Motor instructions are the target instructions and widely designed to decode in BCI systems. However, these instruction information is distributed across different components of motor imagery electroencephalography (EEG). In order to extract motor imagery information from low SNR (signal-noise-ratio) EEG signals and decode it accurately, a large number of feature extraction and classification algorithms based on machine learning are applied to construct EEG decoding models [4–10]. In particular, Filter bank common spatial pattern (FBCSP), as an extremely popular algorithm, selected effective EEG features by introducing a bank of spatial filters and became the winner of motor imagery datasets in BCI Competition IV [11]. Nevertheless, due to handcrafted features, these traditional feature extraction methods highly relied on priori knowledge, which makes the construction of the whole model for EEG decoding inefficient.

Recently, various proposed high-performance neural networks have gained great success when deep learning is applied in computer vision, speech recognition and other task scenarios. Hence, researchers hope to follow these schemes to design a neural networks specialized for EEG-motor imagery decoding. Actually, EEG data is time series data, owing to various forms in EEG representations, it leads to many branches in the network architecture design [12,13]. An efficient solution is to employ the EEG time–frequency representations as inputs and eventually use the designed CNN or SAE to extract from these spectral features and classify [14–18]. In addition, a CNN model that specifically

* Corresponding author.
E-mail address: mawf@zust.edu.cn (W. Ma).

https://doi.org/10.1016/j.bspc.2021.103021
Received 26 April 2021; Received in revised form 1 July 2021; Accepted 30 July 2021
Available online 11 August 2021

receives topographical representation as inputs was developed and achieved high EEG decoding performance [19]. As these approaches usually only select the specific channels and interested frequency bands to convert EEG raw into images, so too few EEG training data is a challenge for the deep network. In order to make full use of the information in recorded EEG datasets, the method of employing raw multi-channel EEG as inputs is gradually favored and popular [20–27]. Researchers adopt deep learning technique to integrate the traditional EEG decoder architecture, and merge the feature extraction and classification that are separated from each other into a single process. Besides simplifying the structure of the decoder, elaborate model components greatly enhance the ability to extract features from EEG signals. Schirrmeister [21] imitated the feature extraction mode of FBCSP and proposed CNN models of different depths by stacking temporal and spatial convolutional layers while indicated that shallow architectures are more generic for EEG decoding. On this basis, Lawhern and Sakhavi each proposed components with stronger EEG feature information representation capabilities and CNNs [22,23], these efforts have made the decoder no longer limited to specified EEG decoding task and simultaneously gained high decode accuracy across several different BCI paradigms. Remarkably, despite the above methods adopted EEG raw data to feed into the network for training, several existing work has changed and further improved the input representation of raw multi-channel EEG data. Tang et al. generated additional channel data for raw EEG channels with the empirical mode decomposition, and designed an 1-D CNN containing multi-scale convolution kernels for decoding that allows receiving multi-channel input [28]. A mixed-scale CNN based on channel-projection was proposed by Li et al. [29], with the data augmentation method for training, the classification performance on public EEG-motor dataset is better than the CNNs based on the original input representation. Nevertheless, using deep networks to decode, the following problems will inevitably occur: a large number of trainable parameters involved by CNNs hinder the overall interpretability of the model; EEG training data is too small so that leading to models more prone to overfitting. These issues constrain the improvement of the final decoding performance of model, especially the decoding accuracy and robustness.

Research on the interpretability of networks proposed by deep learning method applied in various fields has attracted more and more attention. Regardless of whether the image or time series data is fed to the network, the way convolution kernels learn features in CNNs directly affects the final performance effect [30–34]. For this reason, researchers are committed to exploring EEG decoding models via post-hoc interpretation techniques. These techniques includes introducing kernel visualization, saliency map, ablation tests and other interpretable analysis to specific layers in the model, so as to give better CNN convolutional layers stacking choices and suggestions, simplify structure of blocks, reduce the number of trainable parameters [20,23,35]. Schirrmeister introduced “factor perturbation experiment” to spy on the learning preference of the convolution kernel for EEG features; In addition, directly introducing an interpretable convolutional layer not only makes the CNNs lightweight, but also allows the model to have the ability to learn potential EEG class-discriminative information that general CNN decoders cannot capture [36,37]. Brroa proposed “Temporal sensitivity” analysis based on kernel weight gradient to quantification of the importance of the different temporal kernels inspired by saliency maps, as well as to display the unique feature information extracted by the model.

In this work, we optimized the stacking order (called explicit stacking) commonly used in existing network models, and proposed a channel-mixing convolution neural network (Channel-Mixing-ConvNet), which merges previously separated individual temporal and spatial convolutional layers. Temporal and spatial convolutional layers no longer appear in an explicit manner and the combined convolutional layer (Channel-mixing convolutional layer) competes closely with and even outperforms the state-of-the-art models in its ability to extract EEG time domain, space domain, and frequency domain features. Our proposed CNN is also a lightweight structure, the implicit stacking has been applied to significantly reduce the number of trainable parameters. Furthermore, the proposed architecture was evaluated on sensorimotor rhythms both during motor imagery (MI) and motor execution (ME) using public benchmark datasets, and via post-hoc techniques to compare differences between explicit and implicit feature extraction modes, further proved that the proposed network has good interpretability, and demonstrated the channel features captured by the Channel-Mixing-ConvNet has potential impact on classification.

2. Method

In this section, firstly we define the problem of EEG decoding into the framework of supervised classification learning via CNNs and provide notations useful for the following description, and then we give a detailed description of our proposed network.

2.1. Problem definition and notations

Now given the complete EEG raw data for each subject, we denote one of them by s and we define it as a 2D relational matrix $R^{E\times T}$ with E denoting the number of electrodes and T representing the number of sampled time steps in each trial. Thus, the EEG dataset as set $D^{(S)} = \{(X_1^s, y_1^s), (X_2^s, y_2^s), \dots, (X_N^s, y_N^s)\}$, the set contains N single target trials, X_i^s can be expressed as specific-class single target trial of the sth subject, that is, the EEG data $X_i^s = R^{E\times(T_b:T_e)} \in R^{E\times T}$ recorded between $T_b \sim T_e$(respectively represent the start point and end point of the sample when specific body part moved or imagined during each trial), while y_i^s is the class label of the ith target trial for subject, then $y_i^s \in L = \{l_0, l_1, \dots, l_c\}$, where c represents the total number of labels in the EEG datasets. For example, in the BCI IV 2a dataset, the four labels are respectively imagery of left-hand movement, right-hand movement, foot movement, and tongue movement.

EEG decoding task is to establish the correct mapping relationship between the EEG dataset containing N single target trials $X^s = \{X_1^s, X_2^s, \dots, X_N^s\}$, between the corresponding N labels $y^s = \{y_1^s, y_2^s, \dots, y_N^s\}$, essentially, means to find a mapping function $f(X_i;\theta) : R^{E\times T} \rightarrow L$, parameterized by θ that can be learned from EEG datasets. In fact, the function can be decomposed into two key parts: feature extraction and classification. Therefore, to further decompose f, we use $\phi(X_i : \theta_\phi)$ to abstract the feature extraction process, where parameter θ_ϕ is be studied in this stage. Let us integrate feature extraction and classifier into a single process function $f(X_i^{(S)};\theta^{(S)}) = g(\phi(X_i^{(S)};\theta_\phi^{(S)});\theta_g^{(S)})$. In our proposed method, we used CNN to implement mapping function f, trained the model by loading EEG data and learn the required parameter set $\theta^{(S)} = \{\theta_\phi^{(S)}, \theta_g^{(S)}\}$. Except for trainable parameters generated by the dataset, hyperparameters also introduced by CNN itself, such as the number of layers, number and size of convolutional kernels, type of activation function, which will affect the final decoding accuracy.

2.2. Channel-Mixing-ConvNet

In our paper, we call the method of constructing an EEG decoding CNN model by stacking separated individual temporal and spatial convolutional layers as explicit stacking, which is adopted in the existing architectures based on general shallow framework for EEG decoding proposed by Schirrmeister et al. (Shallow-FbcspNet). In our method, we designed a different way from explicit stacking to build CNN, and called it “implicit stacking”, by combining the previously separated individual temporal and spatial convolutional layers into one, and we changed the input representation of EEG raw used for feeding, at the same time, a suitable network was built. The proposed network architecture was called Channel-Mixing-ConvNet. It is designed with four fundamental blocks, their detailed structure are shown in Fig. 1 and corresponding parameter information is reported in Table 1. The two

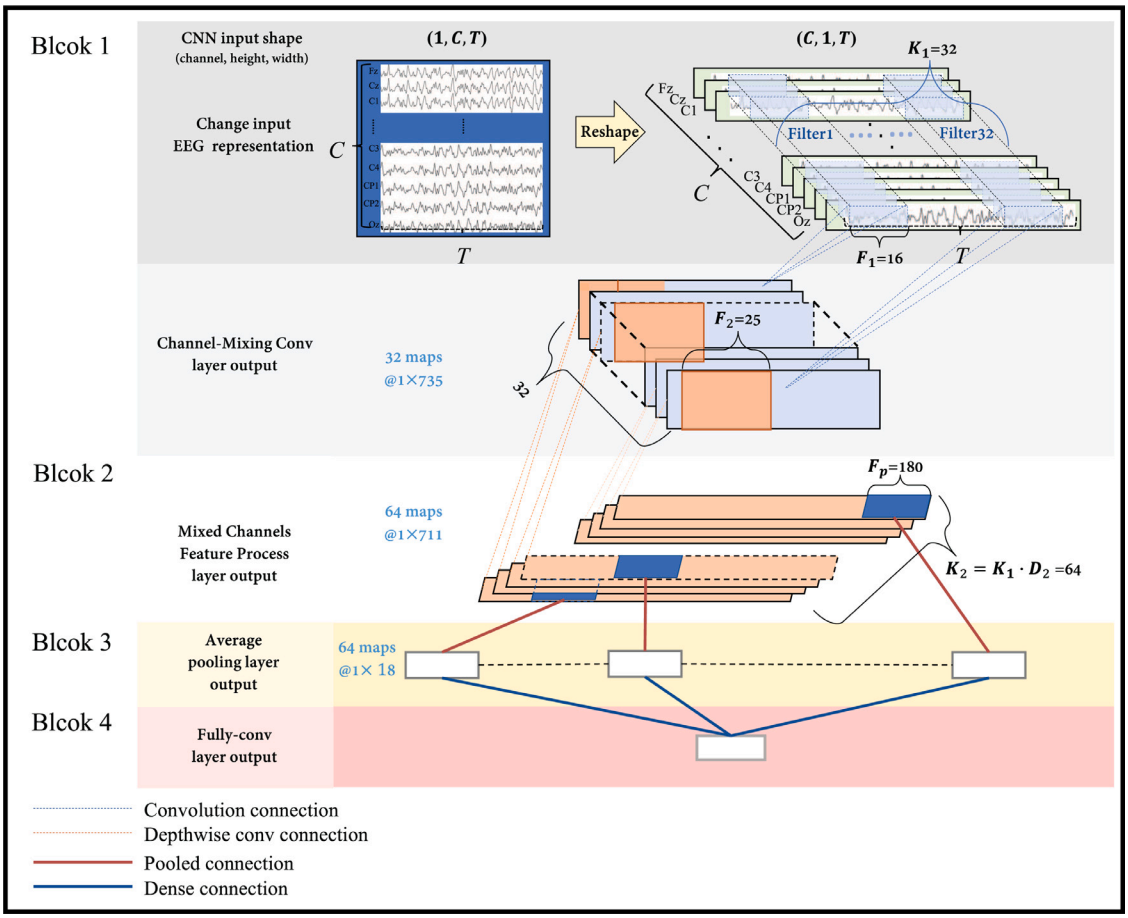

Fig. 1. Channel-mixing Neural Network Architecture. Only major functional blocks are shown, the modules, components and detailed parameters in the network layers are omitted. Additionally, change in EEG input representation is shown.

blocks generated under the implicit stacking mode — Channel Mixing and Mixed Channel Process blocks replaced the original temporal–spatial convolutional layer. Block Channel Mixing is mainly designed to extract related features in time and channel from EEG signals. Block Mixed Channel Process can process and integrate the mixed channels feature maps in depth.

2.3. Block 1: Channel mixing

The differences between EEG signals recorded from different channels are related to the spatial distribution of channels. These differences are also commonly called spatial features of EEG signals and are extracted to represent in spatial convolution of conventional EEG decoding CNNs. The EEG signals in adjacent channels are also more similar in time domain, which reflects the temporal features are localized characteristics of spatial features [24]. Hence, we designed a structure dominated by extraction of channel features in block 1.

Block 1 extracts the global feature of the EEG signal that namely channel feature with the help of the "Channel-mixing Conv2D" convolutional layer. It contains 32 filters implemented by 2D convolution. The size of the convolution kernel F_1 was set to (1,16), and this layer only allows the network to receive a matrix shape of $(C, 1, T)$ as input (see Fig. 1 and Table 1). We introduce a concept similar to RGB channels in images to decode multi-channel EEG data. Compared to the input with the shape of $(1, C, T)$ in conventional EEG decoding CNNs, it allows convolutional kernels to perform multi-channel convolution on EEG signals. This design drive network pay more attention to exploit channel dependency and temporal features while decoding.

Therefore, first, the EEG single trial with the shape of $(C, 1, T)$ is fed into the 32 filters along their C channels(C = 22 or 44 in MI-EEG or ME-EEG, here, with C = 22 as an example). In the following, the 1×16 convolutional kernel is employed to extract temporal and channel representation. Each convolution kernel is across all feature EEG channels, and convolution only across time with a common shape [22], called "the convolution in time". Noticeably, since the height of EEG single trial is 1, the filter bank implemented in a 2-D convolution here is also alternative by an 1-D convolution. Multi-channel convolution will independently convolve the signals from different EEG channels of the same EEG single trial. In fact, since the kernel size of the "Channel-mixing Conv2D" convolutional layer is equal to the number of EEG channels, every signal in channels will produce 22 channel feature maps with shape of $(1, 735)$ additionally but the multi-channel calculation of the convolutional kernel will be involved: for each convolutional kernel, the channel dimension of these feature maps with the shape of $(22, 1, 735)$ will be compressed to one-dimension($(1, 1, 735)$), and 32 channel feature maps containing new signals obtain from 32 filters in the end. The essence of these new signals is to linearly mixing the previous 22 feature maps in each convolution kernel channel, in other words, to linearly combine all the given EEG feature channels. After the layer, the given EEG data removed the original channel form(Fz,Cz, ⋯,Oz) and the channel dimension reduced from 22 to 1, which can be regarded as generating a new single-channel feature signal representation. This is same as the convolution effect produced by the general "channel mixing convolution" [22]. Distinguished from the conventional separate temporal and spatial convolution, "Channel-Mixing Conv2D" simultaneously implicitly completed convolution in time and channel dimension in only a single convolution. As a result, this process actually completed the instantaneous response between the time and channel features and the temporal representations of

Table 1
Detailed parameters of the Channel-Mixing-ConvNet architecture. The model constructed by referring to the following parameters for each network layer, module, and component is denoted as the "basal model". The table describes the functional blocks of the Channel-mixing convolutional neural network, including the corresponding names, modules and component parameters in each network layer; In addition, the output shape, and trainable parameters of each layer, as well as the adopted activation function be enumerated. Here, K and F refer to the number and size of convolutional kernels, S and P refer to the stride size and padding size when performing convolutional kernel operations, and D refers specifically to the multiplier in depthwise convolution; At the same time, several module-level structure: BatchNorm, Exponential Linear Units (ELUs), their hyper-parameters are given in this table.

Block	Layer name	Hyper-parameters	Output shape	Number of parameters	Activation
Block1: Channel mixing	Input		$(C, 1, T)$	0	
	Channel-Mixing Conv2D	$K_1 = 32$ $F_1 = (1, 16)$ $S_1 = (1, 1)$ $P_1 = (0, 0)$	$(K_1, 1, T_1)$	$K_1 \cdot C \cdot F_1$	Linear
Block2: Mixed channel process	DW-Conv2D	$K_2 = K_1 \cdot D_2$ $F_2 = (1, 25)$ $D_2 = (1, 1)$ $S_2 = (1, 1)$ $P_2 = (0, 0)$	$(K_2, 1, T_2)$	$K_2 \cdot F_2[0]$	Linear
Block3: Aggregation	BatchNorm2D	m = 0.99	$(K_2, 1, T_2)$	$2 \cdot K_2$	
	Activation	$\alpha = 1$	$(K_2, 1, T_2)$	0	ELU
	AvgPool2D	$F_P = (1, 180)$ $S_P = (1, 30)$	$(K_2, 1, T_p)$	0	
	Dropout	$p = 0.5$	$(K_2, 1, T_p)$	0	
Block4: Classification	Fallten		$(K_2 \cdot T_p)$	0	
	Fully-Connected	$N_c = 4$	(N_c)	$N_c \cdot K_2 \cdot T_p$	
	Activation		(N_c)	0	Soft-max

each channel are mixed. Hence, the new mixed EEG channel time series contain rich time-channel features that has established a mapping relationship between time and channel dimension.

Indubitably, the implicit stacking makes the depth of the entire network shallower than the explicit stacking, which reduces the large number of trainable parameters introduced by the separated temporal and spatial convolutional layer.

2.4. Block 2: Mixed channel process

The second block is designed to process spatiotemporal features extracted by the block 1 (see Fig. 1 and Table 1). After extracting the channel dependencies and mixed temporal representations, we assume a certain mapping relationship exists in these time-channel(space) features, so the introduced layer will be able to capture and supplement this mapping relationship. The structure of this layer borrows from the approach adopted by many network design schemes in recent years [23, 37–39]: introducing depthwise separable convolution layer. Generally, the depthwise separable convolution is placed in the "spatial convolution" and combines the temporal features from the previous layer, which will bring two advantages: (1) reducing the number of trainable parameters. (2) learning and summarizing each feature map independently, combining the feature map of the previous layer to optimize and merge the output, and it has the ability to decouple the mapping relationships between the internal and cross-feature maps. In this functional block, the size of depthwise convolution kernel F_2 was set to (1,25), and the multiplier D_2 was set to 2, hence a total of $K_2 = K_1 \cdot D_2 = 64$ convolution kernels are involved in performing convolution in time dimension to process EEG time-channel(space) feature maps from block 1. Additionally, kernel maximum norm constraint was adopted. Utilizing of powerful feature extraction capabilities of depthwise convolution, the mapping relationship established in time-channel(space) features extracted by previous layer can be mined and supplemented, more and more different levels temporal and channel features are captured and magnified. A few additional features learned by the convolution kernel may include discriminative information related to motor, which is beneficial to classify.

2.5. Block 3: Aggregation

The main purpose of block 3 is to aggregate the feature maps obtained after processing in block 2 (see Fig. 1 and Table 1). In addition, the structure of function block 3 draws on the layers, components and modules proposed in the current SOTA models. Because the activation function is involved in this block, batch normalization [40] is introduced to optimize the feature maps. Following this module, as well as other CNNs for EEG decoding, Exponential Linear Units (ELUs) were adopted with activation function [41], existing models show that when this activation function is applied, the network has stronger anti-noise ability and more quick training convergence than other activation functions. Furthermore, in order to reduce the parameters to be fitted, an average pooling layer was introduced for down-sampling along the time dimension of the coarser EEG signal feature maps with the shape of $(64, 1, 711)$. The pool size of $F_p = (1, 180)$ and pool stride of $S_p = (1, 30)$ means that the extraction of averaged spatial activations of $\sim$ 720 ms with a stride of $\sim$ 120 ms. As a result, a fine EEG feature representation with shape of $(64, 1, 18)$ is generated. Finally, an added dropout layer can reduce the risk of overfitting [42]. In detailed, the probability of each output p to be set as zero is 0.5.

2.6. Classification

Block 4 is the last block to complete the final classification task. At first, the input of the previous layer is flattened to one 1-D vector, which is a transition from the convolutional layer to the fully-connected layer. Then, these 1-D output values were densely connected with a single fully-connected layer containing $N_C = 4$ neurons.

The entire CNN model is considered a black box function. In fact, for the given EEG data X_i^s of the ith target trial from sth subject, the constructed function $h(X_i^{(S)}; \theta^{(S)}) : \mathbb{R}^{C \times T} \rightarrow \mathbb{R}^{N_C}$ maps X_i^s to one real number per class,where θ is the set of parameters introduced by the entire network. Specifically, the N_C outputs obtained from the fully-connected layer were activated by the soft-max function, which constructs mapping relationship function of the probability that a given input X_i^s belongs to the kth label $l_k \in L = \{l_0, l_1, \ldots, l_{N_C-1}\}$, $p(l_k|X_i^{(S)}, \theta^{(S)}) = \frac{\exp h_k(X_i^{(S)}; \theta^{(S)})}{\sum_{j=0}^{N_C-1} h_j(X_i^{(S)}; \theta^{(S)})}$. Therefore, for a single X_i, soft-max activation will give N_C probability values, and the highest probability corresponding to label was the final classification result of X_i, that is, $y_i^{(S)} = f(X_i^{(S)}, \theta^{(S)}) = \arg\max x_{l_k} p(l_k|X_i^{(S)}, \theta^{(S)})$, then by minimizing the sum of loss function to train the model, which makes the network assign high probabilities to the correct labels as possible, where $loss(y_i^{(S)}, p(l_k|X_i^{(S)}, \theta^{(S)})) = \sum_{k=0}^{N_C-1} -\log(p(l_k|X_i^{(S)}, \theta^{(S)})) \cdot \delta(y_i = l_k)$.

The number of parameters produced by the layers of each block when network is be trained can be seen in Table 1. The Channel-Mixing-ConvNet proposed in this paper introduced a total number of trainable parameters of 32 064 and 17 472, for ME-EEG and MI-EEG signals, respectively.

3. Experiment results

3.1. Dataset

In experiments, the effectiveness of the Channel-Mixing-ConvNet was evaluated on two public benchmark datasets BCI-IV2a and HGD. BCI IV 2a collected by Graz University is a 22-channel dataset for studying MI tasks. This dataset includes four-class motor imagery task (left, right, feet, and tongue) recorded from 9 healthy subjects. The raw EEG data was sampled at 250 Hz and band-pass filtered between 0.5 and 100 Hz. High-gamma dataset (HGD) is a ME-EEG dataset acquired from 14 healthy subjects by Schirrmeister et al. The EEG signals were recorded by 128 electrodes with a sampling rate of 500 Hz. This dataset is suitable for algorithms to test for four-class motor execution decoding tasks (left, right, feet and rest) and to extract information from the high-frequency band of EEG.

We follow the principle of minimal preprocessing of EEG Raw, and completed the preprocessing with the help of the Braincode framework proposed by Schirrmeister et al. The EEG signals of BCI-IV2a and HGD were band-pass filtered between 4 and 38 Hz, 4 and 128 Hz with a 3rd order Butterworth filter, respectively, and each electrode signal was standardized by applying an exponential moving average window with a decay factor of 0.999.

In BCI IV 2a dataset, we extracted 288 target trials X_i^s and selected $X_i^s = R^{E\times(T_b:T_e)}$, where $T_b = 125$, $T_e = 875$, $T = 750$. Considering that the sampling frequency is 250 Hz, this segment of data was extracted by using the same time window [0.5 s, 4 s] on the EEG motor imagery signals. For each subject, here we used EEG data from 288 single trial as training set, further, 20% of the training set were used as the validation set. Additionally, BCI-IV2a dataset also provides EEG data recorded from the same 9 subjects and contained 288 target single trials for the model to evaluate on the test set.

Similarly, in the HGD dataset, the original sampling frequency is 500 Hz, considering that the experiment is needed to keep at the same signal resolution (the sampling frequency of BCI-IV2a is 250 Hz), so the EEG signals were downsampled from 500 to 250 Hz. When extracting trial segment, a window size of [−0.5,4.5] was used and here $T = 1125$. Since the total number of single target trial performed by each subject was different, but roughly about 1000, 160 single target trial were split as test set and the rest was for model training and validating(20% of the training set).

To evaluate the decoding performance of the Channel-Mixing-ConvNet, 5-fold cross-validation was implemented on MI-EEG and ME-EEG dataset. Five models were trained per subject, data from four folds as training set and the held-out fold was for validation. An additional test set was provided to test the classification accuracy of the model. The final result is reported based on averaging.

3.2. Classification performance

We compared our Channel-Mixing-ConvNet (basal model) with state-of-the-art method including Shallow-FbcspNet [21], EEGNet [23], C2CM [22], WaSF-ConvNet [36], Sinc-ShallowNet [37] and CP-MixedNet [29]. Wilcoxon signed-rank test are used for checking significant between SOA methods and machine learning algorithm FBCSP as well as other SOA CNNs including our model. Additionally, the traditional machine learning algorithm FBCSP was also a baseline method for comparison. About selected evaluation metrics to assess the performance of the EEG decoding in this paper, apart from decoding accuracy, kappa value $\kappa = (P_a - P_c)/(1 - P_c)$ was also used, here, P_a denotes the proportion of correct classifications (equals to accuracy), P_c is the proportion of random classification (for example, under four classification tasks $P_c = 0.25$). At the same time, the confusion matrix was calculated to compare the classification performance difference between the our method and the traditional machine learning approach FBCSP in each class. Furthermore, for completeness, we also compared the complexity of the model, including the amount of trainable parameters introduced and the time required for training.

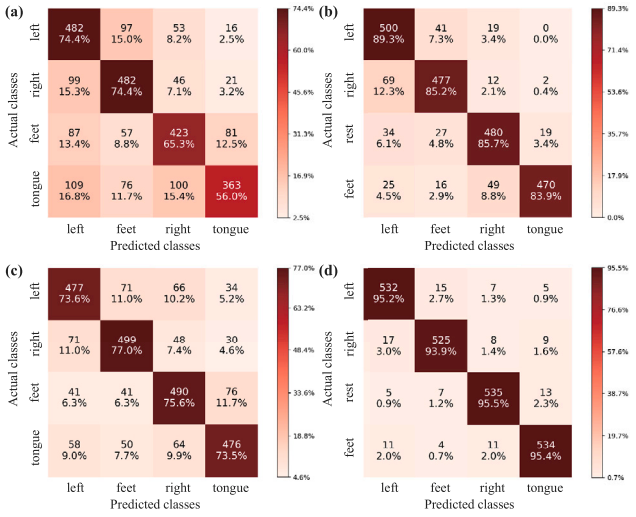

Fig. 2. Confusion matrices of FBCSP+rLDA ((a) and (b)) and Channel-Mixing-ConvNet ((c) and (d)). Matrices (a) and (c) were computed across subjects on MI-EEG dataset; Matrices (b) and (d) were computed across subjects on ME-EEG dataset. All classification results of each class were obtained from test set.

3.3. Overall comparison with baseline models

In this section, the proposed model was compared with traditional machine learning algorithm and the other six aforementioned CNN models in classification performance.

At first, Fig. 2 reports the confusion matrices calculated from the proposed Channel-Mixing-ConvNet and the machine learning algorithm FBCSP+rLDA , with MI-EEG and ME-EEG signals. We summarized the all confusion matrices corresponding to each fold under 5-fold cross-validation and averaged them as the final matrix to measure the classification performance of the model for specific tasks. Specially, the classification accuracy of each class in every confusion matrix was obtained across subjects. After the traditional FBCSP algorithm finished the feature engineering, rLDA was introduced as a classifier. In the MI-EEG classification task, results show that despite FBCSP+rLDA competed closely with the Channel-Mixing-ConvNet on the classification accuracy of Left Hand and Right Hand, but for pattern recognition of Feet and Tongue, this algorithm cannot maintain the previous level. However, Channel-Mixing-ConvNet still has a high decoding accuracy in Feet and Tongue classification tasks (see Fig. 2(a)(c)). In the ME-EEG dataset, the performance of traditional methods on the EEG decoding task was significantly worse than the Channel-Mixing-ConvNet (see Fig. 2(b)(d)).

Secondly, Tables 2 and 3 summarizes the overall decoding accuracies derived (Mean ± Std) by our proposed method and the seven SOA algorithms including FBCSP, Shallow-FbcspNet, EEGNet, C2CM, WaSF-ConvNet, Sinc-ShallowNet, CP-MixedNet from both MI-EEG and ME-EEG datasets, respectively. From these tables, we can see that, Shallow-FbcspNet, C2CM, Sinc-ShallowNet and the proposed Channel-Mixing-ConvNet all have reached 70%+ on decoding accuracy in MI-EEG dataset, respectively were 72.0% ± 13.9, 74.4% ± 14.5, 72.8% ± 12.9, 74.9% ± 14.9. Especially, the Channel - Mixing-ConvNet achieves

Table 2
Overall comparison of decoding accuracy belonging to the test set on MI-EEG signals (P_1 corrected for each CNN vs. FBCSP, P_2 corrected for Channel-Mixing-ConvNet vs. each SOTA CNN).

Methods	Proposed architecture or algorithm	Accuracy (%) ± Std (MI-EEG)	P_1	P_2
Ang (2008) [10]	FBCSP	67.0 ± 13.9		
Schirrmeister(2016) [21]	Shallow-FbcspNet	72.0 ± 13.9	0.028	0.173
Lawhern(2018) [23]	EEGNet	66.0 ± 13.1	0.575	0.007
Sakhavi(2018) [22]	C2CM	74.4 ± 14.5	0.008	0.402
Zhao(2019) [36]	WaSF-ConvNet	68.1 ± 11.6	0.020	0.007
Borra(2020) [37]	Sinc-ShallowNet	72.8 ± 12.9	0.024	0.183
Li(2020) [29]	CP-MixedNet	73.2	0.015	0.213
	CP-MixedNet*	74.6		
Proposed method	**Channel-Mixing-ConvNet**	**74.9 ± 14.9**	0.013	

Where bold fonts indicate the best results and * indicates that the model is trained with data augmentation method.

Table 3
Overall comparison of decoding accuracy belonging to the test set on ME-EEG signals (P_1 corrected for each CNN vs. FBCSP, P_2 corrected for Channel-Mixing-ConvNet vs. each SOTA CNN).

Methods	Proposed architecture or algorithm	Accuracy (%) ± Std (ME-EEG)	P_1	P_2
Ang (2008) [10]	FBCSP	86.0 ± 9.0		
Schirrmeister(2016) [21]	Shallow-FbcspNet	93.9 ± 9.3	0.024	0.021
Lawhern(2018) [23]	EEGNet	88.5 ± 11.0	0.158	0.005
Sakhavi(2018) [22]	C2CM	–		
Zhao(2019) [36]	WaSF-ConvNet	–		
Borra(2020) [37]	Sinc-ShallowNet	91.2 ± 9.1	0.024	0.003
Li(2020) [29]	CP-MixedNet	93.0	0.005	0.015
	CP-MixedNet*	93.7		
Proposed method	**Channel-Mixing-ConvNet**	**95.0 ± 7.3**	< 0.001	

Table 4
Subject-specific decoding accuracy of SOA algorithms for each subject for the BCI IV 2a dataset.

Subject	Shallow-FbcspNet [21]	C2CMC [22]	WaSF-ConvNet [36]	Proposed method
Subject-1	86.65	87.50	71.12	**87.50**
Subject-2	62.29	**65.28**	48.42	52.43
Subject-3	89.86	**90.28**	78.25	88.89
Subject-4	65.61	66.67	61.13	**71.18**
Subject-5	55.19	62.50	67.41	61.11
Subject-6	48.47	45.49	52.25	**52.78**
Subject-7	86.07	89.58	76.64	**90.97**
Subject-8	78.41	83.33	79.47	**85.07**
Subject-9	76.05	79.51	78.12	**84.38**
Mean (%) ± **Std**	72.0 ± 13.9	74.4 ± 14.5	68.1 ± 11.6	**74.9 ± 14.9**

Table 5
Subject-specific decoding accuracy of SOA algorithms for each subject for the HGD dataset.

Subject	Shallow-FbcspNet [21]	Sinc-ConvNet [37]	Proposed method
Subject-1	71.25	83.13	**93.75**
Subject-2	87.50	92.50	**93.13**
Subject-3	96.88	96.88	**100.00**
Subject-4	98.13	98.75	**98.75**
Subject-5	99.38	89.38	**100.00**
Subject-6	93.75	90.63	**96.88**
Subject-7	85.53	86.79	**94.34**
Subject-8	90.63	95.00	**99.38**
Subject-9	98.13	81.88	**98.13**
Subject-10	91.25	92.50	**95.00**
Subject-11	**97.50**	75.00	96.88
Subject-12	94.38	92.50	**96.88**
Subject-13	92.45	87.42	**96.86**
Subject-14	**80.63**	73.75	70.00
Mean (%) ± **Std**	91.2 ± 7.6	88.3 ± 7.3	**95.0 ± 7.3**

Table 6
Comparison of our method with SOA algorithms in terms of κ value for BCI IV 2a.

Subject	Shallow-FbcspNet [21]	C2CM [22]	WaSF-ConvNet [36]	Proposed method
Subject-1	0.822	0.833	0.621	**0.833**
Subject-2	0.497	**0.537**	0.324	0.368
Subject-3	0.865	**0.870**	0.712	0.852
Subject-4	0.541	0.556	0.403	**0.616**
Subject-5	0.403	0.500	**0.586**	0.481
Subject-6	0.313	0.273	0.326	**0.370**
Subject-7	0.814	0.861	0.662	**0.880**
Subject-8	0.712	0.778	0.721	**0.801**
Subject-9	0.681	0.727	0.688	**0.792**
Mean (%)	0.628	0.659	0.560	**0.665**
±**Std**	±0.19	±0.19	±0.15	±0.20

the highest accuracy of up to 74.9% ± 14.9. The decoding performance of C2CM is very close to our proposed model, and the accuracy is only 0.5% lower. Similarly, CP-MixedNet has reached to 73.2%, which is close to the proposed Channel-Mixing-ConvNet. CP-MixedNet* was trained with data augmentation, it achieves the significantly higher accuracy of 93.7% than all the other methods. Although data augmentation improved its decoding accuracy by 1.4% than the original, it is still sightly lower than the proposed Channel-Mixing-ConvNet. In addition, we also evaluated these methods on ME-EEG dataset, among them, Shallow-FbcspNet, Sinc-ShallowNet, CP-MixedNet, CP-MixedNet* still maintain a high performance level across datasets, and their decoding accuracy reaches to 93.9% ± 9.3, 91.2% ± 9.1, 93.0%, 93.7% respectively. However, the Channel-Mixing-ConvNet is 1.3% higher than the CP-MixedNet* which achieves the highest decoding accuracy in six SOA CNNs, reaching to 95.0% ± 7.3. Note that even with the data augmentation, the decoding accuracy of CP-MixedNet* is only improved by 0.7%. Consistent with our expectation, the same as the other SOA CNNs, the accuracy of our proposed model is improved by at least 7% compared with the traditional machine learning algorithm FBCSP on the public datasets.

As well as within-subject decoding accuracy of several SOA models has been shown in Tables 4 and 5. As these above tables shown in, for MI-EEG dataset, compared with Shallow-FbcspNet, C2CM and WaSF-ConvNet, the within-subject decoding accuracy of proposed model has been surpassed by all subjects except subject 2, 3, 5, and the most of the surpassing effects are higher than 5%. Similarly, it is noted that the Channel-Mixing-ConvNet shows stronger decoding ability than Shallow-FbcspNet and Sinc-ConvNet for ME-EEG dataset. In particular, the decoding accuracy of subjects 3 and 5 reaches to 100%, and the

Table 7
Comparison of the total number of trainable parameters.

Datasets	Shallow-ConvNet	EEGNet	Sinc-ShallowNet	CP-MixedNet	Channel-Mixing-ConvNet
BCI IV 2a	40 644	1932	5508	8.36×10^5	17 472
HGD	82 564	2604	13 828	8.36×10^5	32 064

Table 8
Mean training times (hh:mm:ss) over all subjects of different methods.

Datasets	Shallow-ConvNet	Sinc-ShallowNet	CP-MixedNet	Channel-Mixing-ConvNet
BCI IV 2a	00:18:22	00:06:33	00:15:25	00:08:07
HGD	00:39:46	00:15:46	00:24:40	00:19:11

overall result was that the decoding accuracy of subjects 9, 11 and 14 was slightly lower than the other comparison models.

Additionally, in the MI-EEG dataset, for each subject, corresponding kappa value κ was calculated and depicted in Table 6, as a supplement to the performance comparison of motor imagery EEG decoding.

Finally, in order to compare the complexity of the our model with other SOA CNNs, the total number of trainable parameters is reported in Table 7 and the average training time across subjects given in Table 8.

According to the above experimental results, the proposed model Channel-Mixing-ConvNet has the ability to generalize across different EEG datasets, and its decoding ability of motor imagery EEG is close to or even better than the current SOTA algorithm.

3.4. Impact of hyperparameters in Channel-Mixing-ConvNet

In this section, in order to analyze the decoding performance difference between the basal model and variant models, a series of experiments were carried out from the aspects of hyper-parameter evaluation and ablation test for different design motives.

In the hyper-parameter evaluation, we studied the impact of the following hyper-parameters on the model:

(i) Number of Channel-mixing Conv2D K_1 of block 1.
(ii) Size of DW-Conv2D F_2 of block 2.
(iii) Number of DW-Conv2D D_2 of block 2.
(iv) Pooling size F_P and stride S_P of block 3.

Moreover, we validated the impact of single model components by systematically altering the model via ablation test. Therefore, we trained the following variants additionally:

- $K_1 = 16, 22, 64$: Since the number of kernels of the Channel-mixing Conv2D controls the number of new mixed EEG channel signals, and these new signals are coarser intermediate feature representations, we want to test if different number of filters to change the capability of channel feature representation for different EEG datasets containing different number of channels. We expect that heightening the number of the filters may facilitate the extraction of channel feature for EEG dataset including more channels, such as the investigated ME-EEG dataset($C = 44$).
- $D_2 = 4$: We want to test whether a higher D_2 would be better for decoding large EEG data, such as HGD containing a much larger EEG signals than BCI IV 2a.
- $F_2 = (1, 1), (1, 44)$: We want to test the impact of the size of the depthwise convolutional kernel in block 2 on the ability of feature processing. Hence, the 1×1 convolutional kernel is introduced to verify whether it can improve the cross-channel feature; In addition, the kernel with a larger size is also used to construct the variant model and we hoped to expand the receptive fields of the kernel, although this is at the cost of more trainable parameters, to evaluate whether the performance can be improved.
- No DW-Conv2D: Main layer in block Mixed Channel Process was removed to explore impact on the model. Indirect to prove whether the layer can mine more level of features.
- Short-large avg pool size : We want to investigate the impact of a shorter average pooling on the performance. We try to minimize the loss of information in down-sampling to preserve key feature representations, however, which resulted in an increased number of trainable parameters.

A comparison of the decoding performance obtained by training on MI-EEG and ME-EEG datasets of the basal model (parameters are shown in Table 1) and the variants is shown in Fig. 3. In experiments, the overall impact from hyper-parameters changes or alter on our model was quantified by the difference in the decoding accuracy Δ_{acc}, specifically, $\Delta_{acc} = acc_{\text{variant}} - acc_{basal}$. Moreover, the within-subject decoding performance and overall performance of the model were evaluated on MI-EEG and ME-EEG datasets.

It is obvious that there is no significant difference between the average decoding performance of these variants and the basal model, but the decoding accuracy of several subjects is slightly different. Specifically, the decoding accuracy of several subjects in ME-EEG fluctuates by more than 5% compared with the basal model, while the performance in MI-EEG is more stable.

3.5. Interpretation of discriminative information

The interpretation of motor imagery EEG features is generally related to the overall performance of the model. In addition to decoding accuracy can directly reflect the performance of decoding model, whether the most discriminative information can be learned by model from the EEG signals usually shows the sensitivity and capture ability of the model to the motor imagery EEG features. This ability to explain the features provides the basis for generalization of the model. In this paper, several post-hoc interpretation techniques were introduced to analyze features extracted from block Channel Mixing and block Mixed Channel Process, mainly including channel sensitivity analysis and temporal sensitivity analysis.

In terms of channel sensitivity analysis, we visualized 32 learned kernels of the "Channel-mixing Conv2D" convolutional layer to understand the learning behavior of these kernels, which helps us figure out these convolution kernels are more prone to what to learn from EEG signals and especially learning preferences of different channels when performing decoding tasks under ME-EEG and MI-EEG. At the same time, saliency map was introduced to evaluate the comprehensive influence of temporal and channel dimension intersections in EEG feature maps on the classification results.

In temporal sensitivity analysis, with the help of visualization of convolution kernel, we extracted the intermediate output of EEG signals after the "Channel-mixing Conv2D" and the "DW-Conv2D" convolutional layer, and for the sake of time–frequency domain analysis, the intermediate signals are converted to spectrum form. These learned feature analysis can clearly understand the learning preference of the model for different frequency bands including $\theta = (4, 8)$ Hz, $\alpha = (9, 12)$ Hz, $\beta = (13, 30)$ Hz, $\gamma = (31, 125)$ Hz, and the ability to capture the frequency band that contains the most discriminative information for classification across different datasets.

In this section, the introduced interpretable method visualizes the proposed Channel-Mixing-ConvNet feature extraction pattern, and we demonstrate the learning properties of the proposed model on MI-EEG and ME-EEG datasets, as well as its sensitivity to motor imagery information used for decoding in EEG signals.

Fig. 4(a) and (b) illustrates the visualization of each convolution kernel average weight after these convolution kernels in the "Channel-mixing Conv2D" convolutional layer has learned the EEG signals across all subjects in MI-EEG and ME-EEG datasets, respectively. Here, the absolute value of the convolution kernel weight on a certain channel

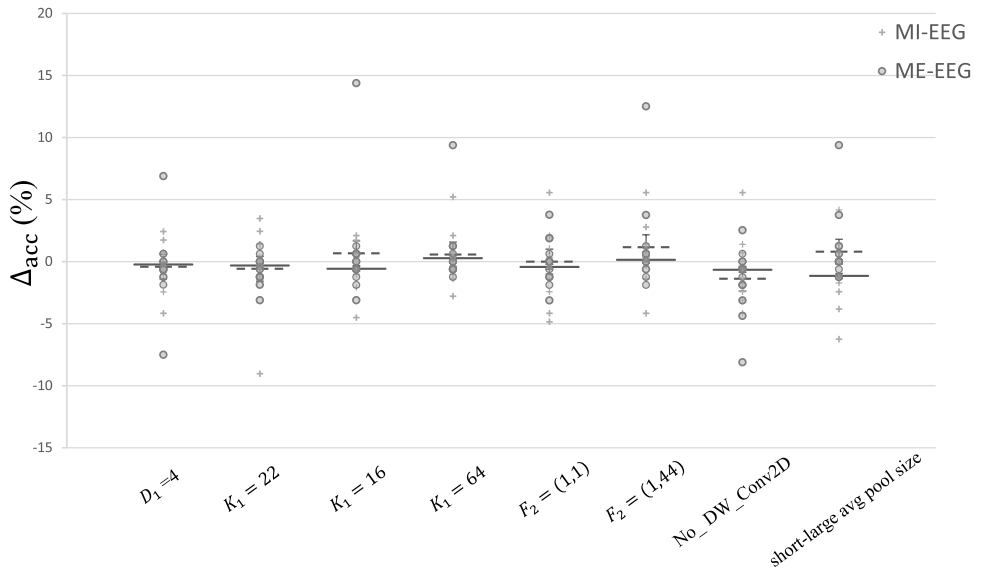

Fig. 3. Results of the impact analyses on the Channel-Mixing-ConvNet between basal model and variants. The real line segment and the dashed line segment represent the difference in average decoding accuracy between the basal model and variants on MI-EEG and ME-EEG datasets, respectively. The within-subject decoding accuracy on MI-EEG and ME-EEG datasets were denoted by symbols(+),(•).

reflects the learning tendency of different channels, so the positive and negative values of the weight were ignored. These sensitive channels are marked in Fig. 5(a) and (b). For the MI-EEG dataset, these sensitive channels are mainly distributed around CZ, C3 and C4 , with a few distributed around Pz (see Fig. 5(a)). In the ME-EEG data set, these channels were largely distributed the area below the Cz and around the CPz, while the remaining channels were scattered around the C3 and C4 (see 5(b)). In order to more directly show the overall learning behavior of the convolution kernel, we averaged the absolute weights of 32 convolution kernels across all subjects and injected them into the EEG topographical map. Figs. 6–8 respectively show the overall channel sensitivity of convolution kernels on MI-EEG and ME-EEG datasets. Average absolute value of kernel weight across all subjects was illustrated in Fig. 6(a) and (b), whether MI-EEG or ME-EEG dataset, since the activation of contralateral primary motor cortex is caused by performing a left hand or right-hand movement or imagination, in detailed, the imagination of the left-hand movement activated the area near the C4, and the imagination of the right-hand movement activated the area near the C3. At the same time, imagination or motor execution of the feet activates the central primary motor cortex, the area around the Cz. In addition, we noted that the convolution kernels were extremely active on both motor imagination and execution datasets for areas below the Cz, often referred to as the primary somatosensory cortex, and activation in this part of the cerebral cortex is associated with the motor imagination of tongue or rest states. Therefore, on the two datasets, the convolution kernel shows a very active state on the C3, C4, Cz and adjacent channels. These investigation results are the same as the results of analysis obtained by Zhao, Brroa et al.

The EEG topographical maps shown in Figs. 7 and 8 also reflects these similar phenomena, and the red region in the figure represents that the convolutional kernels are more active and sensitive to the EEG signal learning at these areas than others. In addition, subject-specific channel sensitivity of the Channel-Mixing-ConvNet is displayed in order of within-subject decoding accuracy from low to high. For different datasets, the model learned the target channels that associate the most relevant activations in motor areas. For example, in MI-EEG, they are C3 and C4 channels. With the target channels captured by the model becoming more and more complete and comprehensive across all subjects, the decoding performance is getting better.

In order to further prove that the features learned by the proposed model from EEG signals contains discriminative information for classification and the interpretability of the model, two subjects were selected from both MI-EEG and ME-EEG datasets to create the within-subject saliency maps. We examined the within-subject saliency maps of two subjects for our model, they were chosen to illustrate saliency maps for high- and low-performing subjects (see Figs. 9 and 10). Positive and negative gradients indicate the size of effect of changing a feature point on the final classification. From the figure, we can see, for MI-EEG dataset (see Fig. 9(a)), in the temporal dimension, the most discriminative information to drive the classification mainly concentrated in after the start of 1 s to 2.5 s in the motor imagery EEG signal, moreover, on the channel dimension, electrode Pz, CP1, C3, C4 and Cz and Fz contains the most critical features used for classification. For ME-EEG dataset (see Fig. 10(a)), the start of 1.6 s to 4 s in the motor imagery EEG signal contains the temporal features that play a decisive role in classification. Meanwhile, CPP2h, CPP1h, CCP5h, CP3, C3, C4 and other adjacent channels contain the channel features that can affect the final classification.

In fact, the time domain and channel cannot exist independently in saliency maps, both complement each other, and EEG features corresponding to the intersections of two dimensions can drive the classification task. For the saliency map of the low-performing subject, has a more scattered appearance across channels and over time, while for the high-performing subject, the points in its saliency maps are presented in a more neat and orderly manner. This phenomenon is usually caused by noise perturbations in the EEG signal. In particular, the saliency map of the low-preforming subject in the ME-EEG (see Fig. 10(b)) have a large amount of noise interference in the channel and temporal dimension of a single sample.

These experiments demonstrate the proposed decoding model is sensitive to the channel features, in addition, in order to further validate the model can capture the frequency domain information for classification, without specifying a particular classification task, the intermediate output of EEG signals after the "Channel-mixing Conv2D" layer were shown in Figs. 11 and 12 as spectrum form. These "new" EEG signals mixed with time-channel features, obtained from the "Channel-mixing Conv2D" convolutionl layer, have a time length of 294 ms on MI-EEG and 444 ms on ME-EEG dataset, respectively. Obviously, we can see from the figures, for MI-EEG dataset, the extracted signals fluctuate frequently in the band $\theta(4, 8)$ Hz, $\alpha(9, 12)$ Hz, $\beta(13, 30)$ Hz which usually

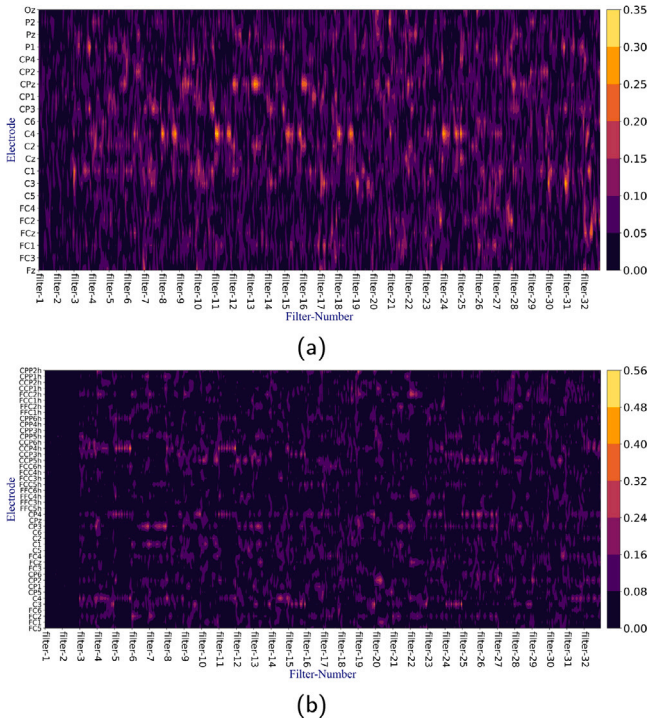

Fig. 4. Visualization results of convolution kernel weights of mixed channel convolutional layers in the Channel-Mixing-ConvNet on different datasets. The weights of each kernel are average weights across all subjects. (a) MI-EEG. (b) ME-EEG.

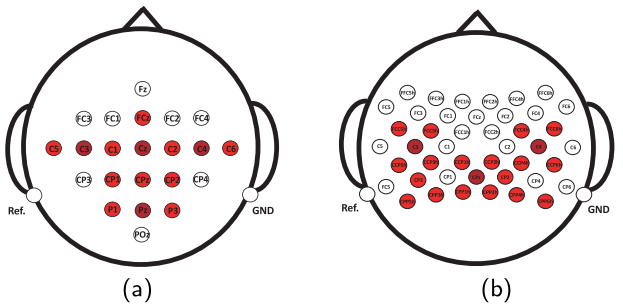

Fig. 5. Main learned sensitive channels for across subjects on different datasets. The darker the color, the more sensitive the region. (a) MI-EEG. (b) ME-EEG.

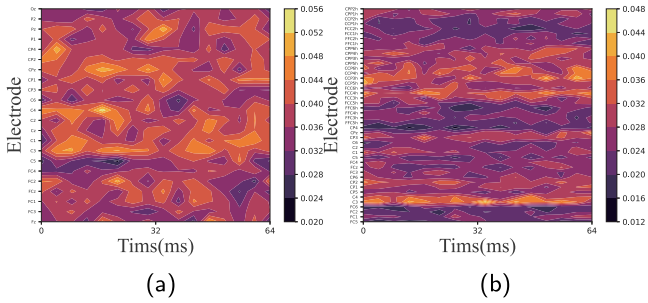

Fig. 6. Learning tendency to different channels of convolution kernels across all subjects on different datasets. (a) MI-EEG. (b) ME-EEG.

contain information related to motor imagery. For the ME-EEG dataset, since the EEG signals recorded in this dataset are all high-frequency γ signals, several convolution kernels are more sensitive to the feature information in the high-frequency band $\gamma(30 \sim 90$ Hz).

4. Discussion

4.1. Classification performance and comparison with state-of-the-art approaches

The performance comparison between Channel-Mixing-ConvNet proposed in this paper and SOA methods shows that our model has a

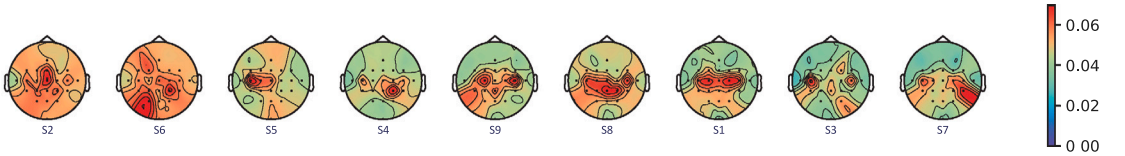

Fig. 7. Overall channel sensitivity of convolution kernels on MI-EEG dataset is displayed on EEG topographical maps. The channel sensitivity learned by the models for each subject is given and they are arranged in order of within-subject decoding accuracy from low to high.

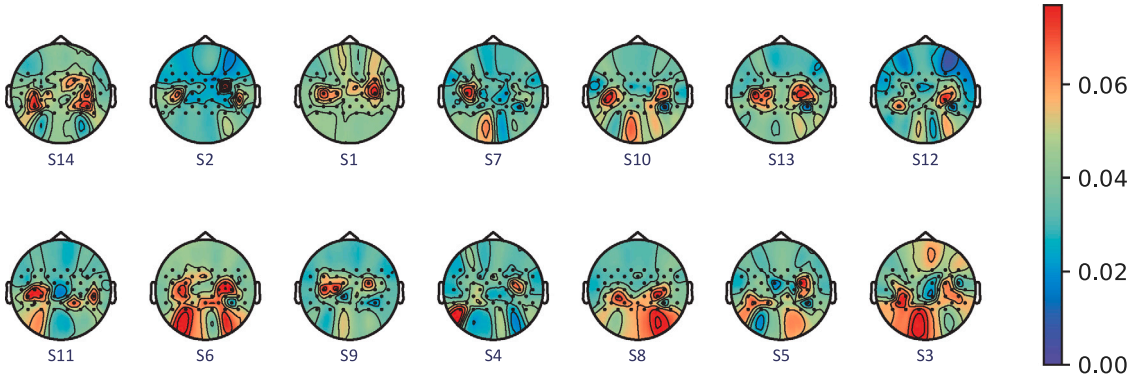

Fig. 8. Overall channel sensitivity of convolution kernels on ME-EEG dataset is displayed on EEG topographical maps. The channel sensitivity learned by the models for each subject is given and they are arranged in order of within-subject decoding accuracy from low to high.

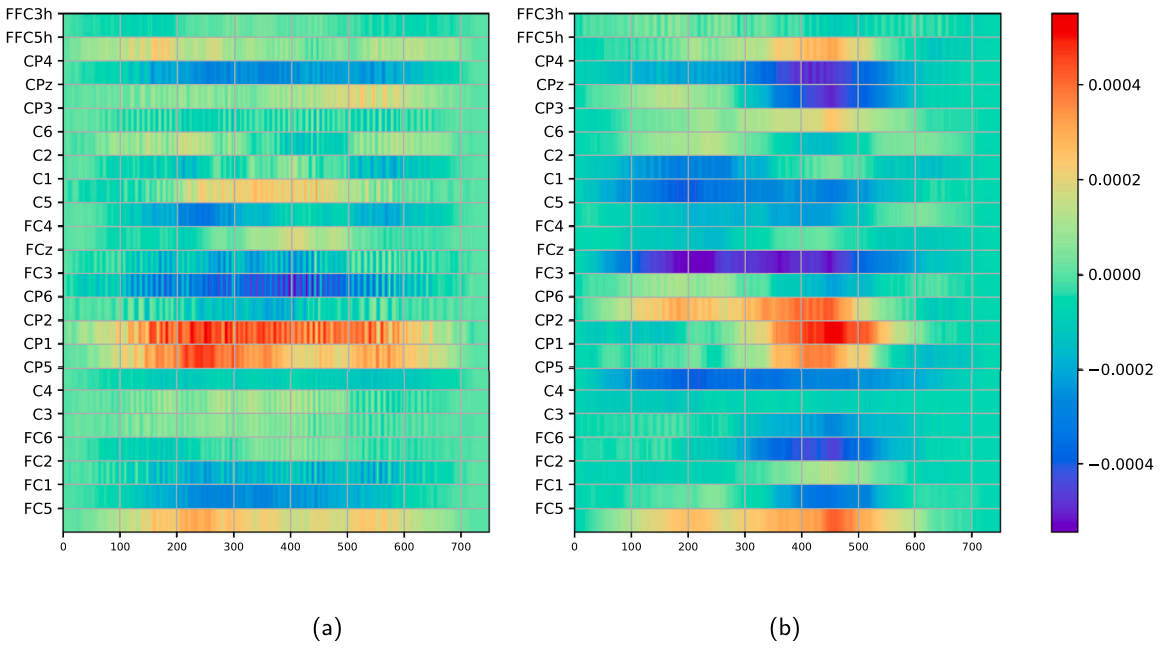

Fig. 9. Within-subject saliency maps of two different subjects, generated by a single example on MI-EEG. The subject on the left achieved high decoding accuracy (87.5%) (a) and the subject on the right achieved low decoding accuracy (61.1%) (b).

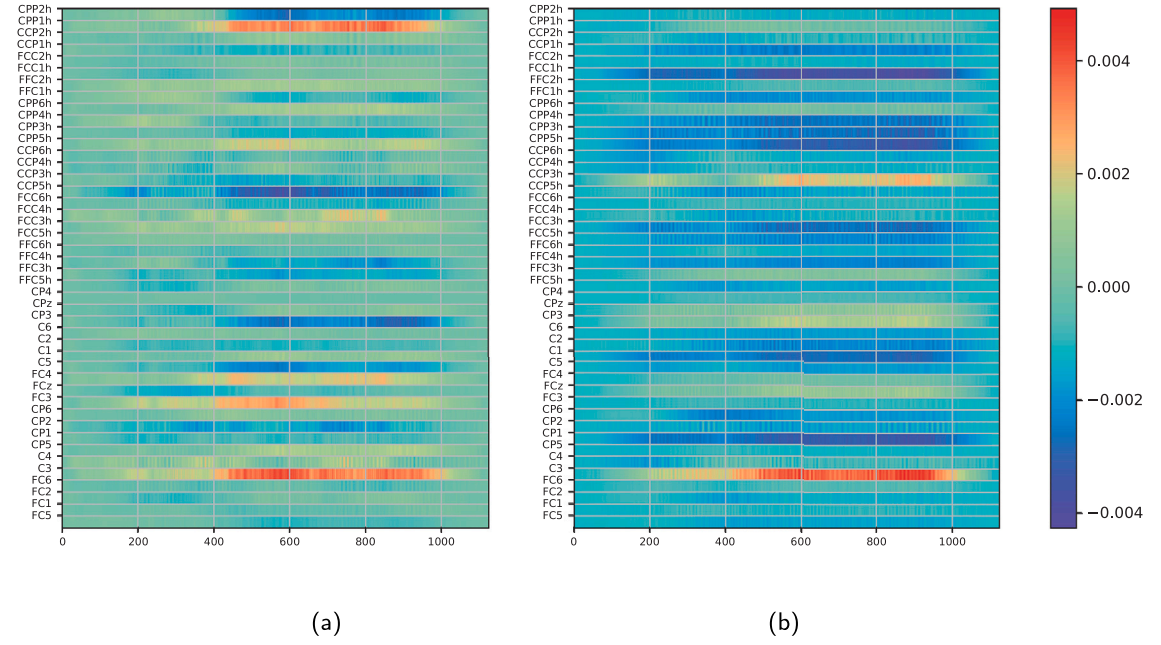

Fig. 10. Within-subject saliency maps of two different subjects, generated by a single example on ME-EEG. The subject on the left achieved high decoding accuracy (100.0%) (a) and the subject on the right achieved low decoding accuracy (70.0%) (b).

strong decoding ability of motor imagery EEG signals. From the results of the confusion matrix, traditional machine learning algorithm FBCSP + rLDA and our proposed method for different classification tasks reflect the difference of the decoding performance. As FBCSP selects features by maximizing the differences between class pattern, for example, in MI-EEG dataset (see Fig. 2(a)(c)), this algorithm can distinguish the class difference between class pattern "Left vs. Right" well, and produced a significantly lower number of misclassifications between "Left vs. Right" than the Channel-Mixing-ConvNet(136 misclassified trials for FBCSP and 178 for Channel-Mixing-ConvNet). However, FBCSP has a obviously weaker ability to identify class pattern "Tongue vs. Feet" than the Channel-Mixing-ConvNet, hence, the classification accuracy of Tongue and Feet respectively is lower, and at the same time, produced a significantly bigger number of misclassifications between "Tongue vs. Feet" than the Channel-Mixing-ConvNet(329 misclassified trials for FBCSP and 190 for the Channel-Mixing-ConvNet).

Compared with SOTA CNN models, the Channel-Mixing-ConvNet also shows stronger advantages in decoding. Among these comparison CNNs, Shallow-FbcspNet, C2CM, Sinc-ShallowNet and the proposed Channel-Mixing-ConvNet are all shallow architecture, Shallow-fbcspNet, Sinc-ShallowNet have a total number of trainable parameters of 82 564, 13 828 in case of ME-EEG signals, and of 40 644, 5508 in case of MI-EEGsignals, respectively, the trainable parameters introduced by the Channel-Mixing-ConvNet(32 064, 17 472) are in the middle of

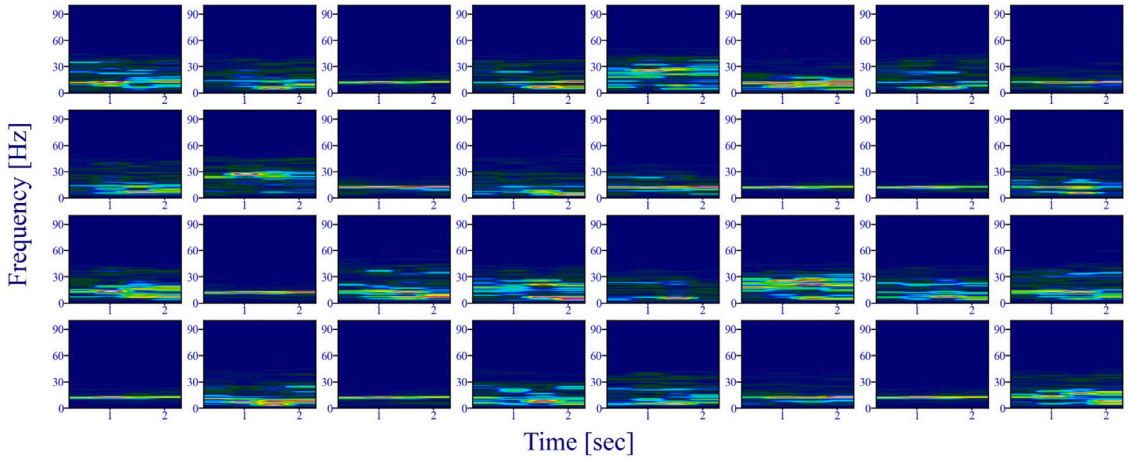

Fig. 11. Spectrum of intermediate signals obtained by 32 filters in the channel-mixing convolutional layer on MI-EEG dataset.

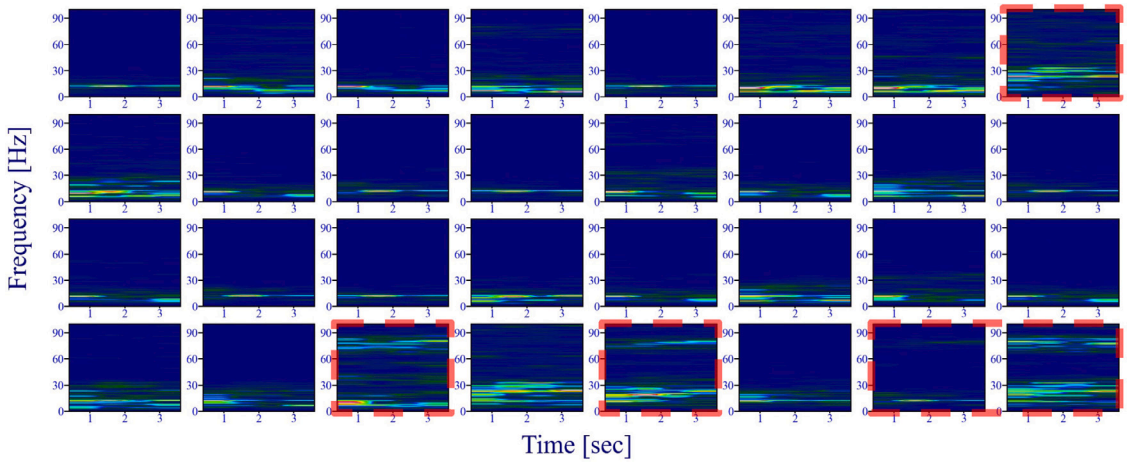

Fig. 12. Spectrum of intermediate signals obtained by 32 filters in the channel-mixing convolutional layer on ME-EEG dataset. Several filters sensitive to the γ(30~90 Hz) band were marked with a red dotted box.

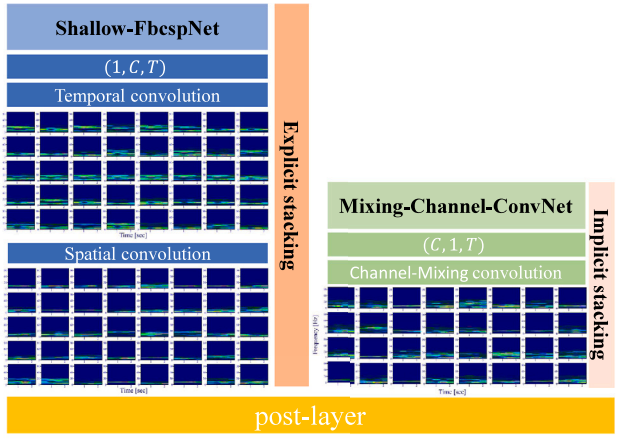

Fig. 13. Spectrum comparison of EEG intermediate signals extracted by CNN constructed in explicit and implicit stacking mode. Left: the time–frequency representation from 2-stage "temporal–spatial" convolutional layers. Right: the time–frequency representation from the "Channel-Mixing" convolutionl layer.

them (see Table 7). Compared with Shallow-FbcspNet, owing to the implicit stacking architecture, the trainable parameters were reduced by more than half, while outperforming these shallow architectures in decoding performance. We perform the training of these networks on NVIDIA Tesla K80 GPU and the average training times over all subjects from the comparison CNNs are listed in Table 8. The results show the Channel-Mixing-ConvNet consumes less time than the other CNNs aforementioned and nearly consumes the same time as the Sinc-ShallowNet. It is indicated that the Channel-Mixing-ConvNet is a lightweight architecture and lower than the shallow models in complexity. Besides, considering real-life BCI system it is a higher priority for quick set-up, high-speed recognition and real-time performance. The training time and size of our proposed model are both in an acceptable range, further, since the Channel-Mixing-ConvNet is simple and has lower complexity, which allows the Channel-Mixing-ConvNet has enough potential to stack other beneficial layers for improve classification accuracy.

We note that the decoding performance of C2CM is very close to the Channel-Mixing-ConvNet, however, this method requires downsampling of the EEG signals for different subjects to adapt to the parameters of the CNN.

Although CP-MixedNet adopts input form similar to mine, it still used an explicit way to stack model depth to extract time-domain features in different levels. Therefore, CP-MixedNet introduces about 8.36×10^5 parameters in the two datasets respectively, which is much larger than the Channel-Mixing-ConvNet. With data augmentation method, the decoding performance in MI-EEG dataset is close to our model. However, in ME-EEG with a more larger amount of EEG data, the

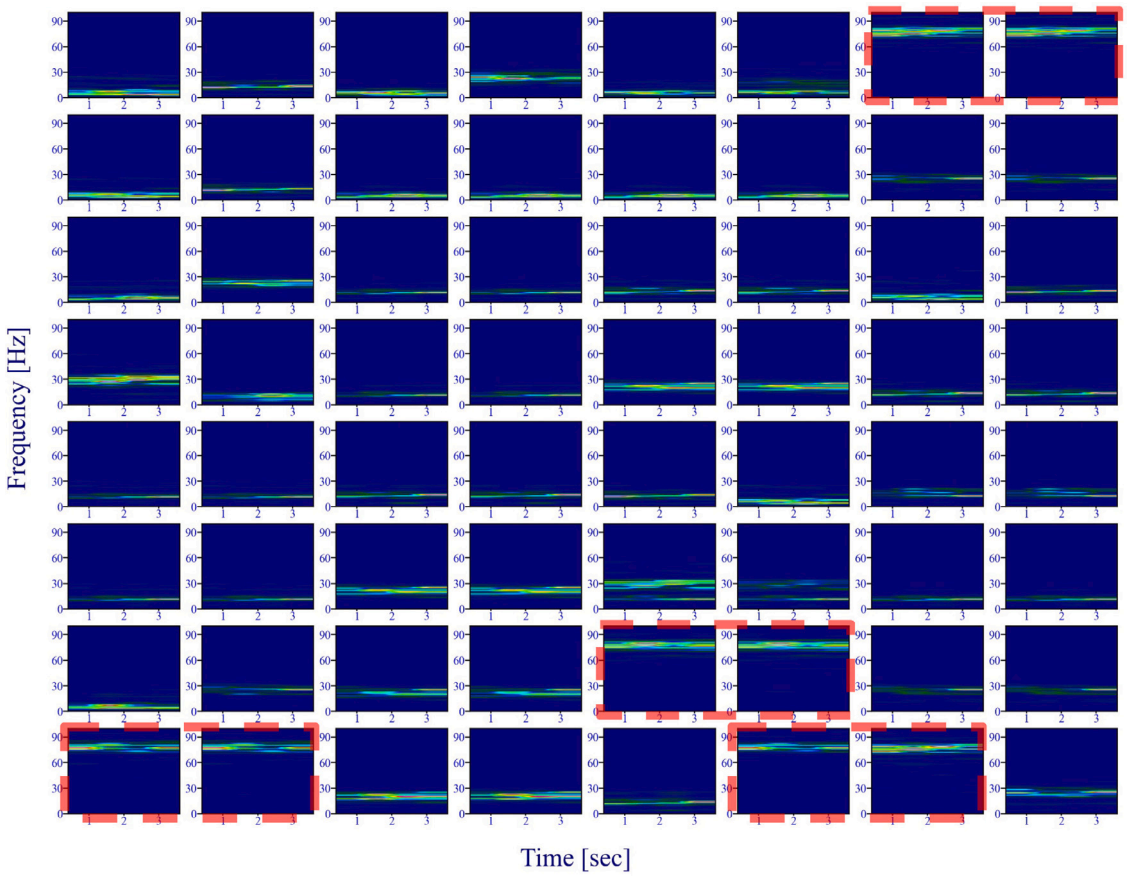

Fig. 14. Visualization of EEG feature maps extracted by 64 filters of depthwise convolution in the Mixed Channel Process block. The feature maps to be processed are from Fig. 12. Filters focusing on high frequency bands were indicated by a red dotted box.

decoding accuracy of CP-MixedNet* is 2% lower than the Channel-Mixing-ConvNet. It indicates that Channel-Mixing-ConvNet has a strong generalization ability across datasets than the CP-MixedNet. In addition, the depth of the Channel-Mixing-ConvNet is more shallower, fewer functional blocks are introduced, but the decoding performance can compete with the deep model CP-MixedNet* with data augmentation, which is enough to prove that the Channel-Mixing-ConvNet has a stronger EEG decoding ability and potential performance to be exploited.

It is worth considering what made several subjects low-performance. In fact, temporal features extracted from the "Channel-Mixing Conv2D" are across multiple time-scales, the length of selected EEG trial as input may become a factor which limits model's decoding performance. In general, motor imagery is pronounced between 0.5 s and 4 s after the cue (as ME-EEG an example), however, this is not necessarily the best target trail for our model to decode. In addition, the trained Channel-Mixing-ConvNet is suitable for denoising data, since we smoothed the EEG signals in the preprocessing but it may reduce the sensitivity of the model to noise, which also limited the robustness of the model especially for subjects with a lot of noise.

Sinc-ShallowNet, WaSF-ConvNet directly introduced an interpretable convolutional layer to improve the performance of the model. Although the proposed Mixed-Channels-ConvNet does not deliberately introduce a similar architecture, it achieved higher decoding accuracy in both MI-EEG and ME-EEG datasets. It is related to the adoption of implicit stacking in Channel-Mixing-ConvNet, which may enable the model to focus on the mapping relation between EEG time-domain and channel features rather than extracting one or the other separately.

4.2. Performance differences from hyperparameters

A series of experiments were carried out to evaluate the decoding performance differences caused by hyperparameters of Channel-Mixing-ConvNet (see Fig. 3). These results of experiments revealed that the impact of hyperparameters on MI-EEG and ME-EEG datasets was in different degree. In particular, the model performance on ME-EEG is more severely affected by the changes of hyper-parameters. Specifically, after altering the number of convolution kernel of the Channel-mixing Conv2D K_1, when $K_1 = 16, 22, 64$, the average decoding accuracy on ME-EEG (denoted by dashed line) has a significant change, while MI-EEG (denoted by dashed line) only has a slight improvement when $K_1 = 64$. This shows that ME-EEG, an EEG dataset with more channels than MI-EEG, is more sensitive to the number of feature maps generated by the "Channel-mixing Conv2D" convolutional layer, and the parameter K_1 controls the number of depthwise convolution kernels K_2 in block Mixed Channel Process. This modification is more beneficial for EEG datasets with a larger number of channels. In addition, when the size of depthwise convolution kernel F_2 increases, the average decoding accuracy of variant increased by 1.16%, while this trend was also observed in MI-EEG, but the improvement was not obvious. For ME-EEG dataset, it is obviously more suitable to introduce more trainable parameters to improve the decoding performance. We also try expend D_2 to enhance the feature process capability of the depthwise convolution layer, but the decoding performance of the variant on each dataset was almost the same as the basal model. It is worth noting that the stride of the average pooling layer has a significant impact on the decoding performance. When a shorter stride pooling layer was created, the average decoding accuracy on MI-EEG dropped by 1.15%, and ME-EEG also achieved a 0.8% performance improvement at the cost of introducing more trainable parameters. Consistent with our

assumptions, after a single removal of the Mixed Channel Process block in the ablation test, the decoding accuracy on ME-EEG and MI-EEG dataset decreased by 1.3% and 0.65%, respectively, which indirectly proves that the Mixed Channel Process block has a gain effect on the decoding ability of the Channel-Mixing-ConvNet.

4.3. Interpretation

The designed "Channel-mixing Conv2D" convolutionl layer is across all EEG feature channels, it primarily drives model to exploit channel dependencies and channel feature representations to improve interpretability of the Channel-Mixing-ConvNet. The 32 filters in this layer can selectively extract the time domain and channel features closely related to the current classification task according to different EEG datasets. In particular, the newly generated mixed signals after the layer still fully contain the channel features required for classification (see Figs. 4–8).

Since our model is dedicated to extracting channel feature representations and depends on critical channels, the decoding accuracy of each subject is positively correlated with the learning quality of the channel features. For the subject with lower decoding performance in MI-EEG dataset, the model cannot accurately learn the key channels C3, C4 that contain the main discriminative information instead mistakenly focus on other irrelevant channels and activations in motor areas. When subject 2 and subject 6 are decoded by the model, the captured target channels are very discrete, which resulted in low decoding accuracy. Obviously, for subject 5, subject 4, the target channels learned by the model are not complete because the key channel C3 or C4 is missing, which hinders event-related desynchronization and synchronization (ERD/ERS) [43] from being revealed by the model under the classification task "Left Hand" and "Right Hand". The similar phenomenon also occurs in ME-EEG, for the low-performance subjects, the channel sensitivity and dependencies learned by the model are significantly weaker than the high-performance subjects. In addition, we also observed that for those high-performance subjects (subject 3,5,6,8,10,11) in ME-EEG dataset, the model not only needs to pay attention to C3 and C4 channels, but also the channels below Cz [44]. It indicates that the Channel-Mixing-ConvNet requires different target channels for decoding task across subjects and datasets. Hence, there are significant differences in the activations in motor areas and sensitivity of the target channels for different EEG datasets when decoding across subjects. This also could not exclude that it is caused by the difference between imagination and execution. Importantly, the saliency maps created under the overall classification task for different within-subject decoding accuracy also prove that Channel-Mixing-ConvNet relies on the key channels summarized above when performing classification and is more targeted for time-channel feature learning when high-performance decoding.(see Figs. 9 and 10).

On the other hand, when the Channel-Mixing-ConvNet was trained on MI-EEG, 32 new signals generated by the Mixed Channel Process block were analyzed by frequency domain sensitivity (see Figs. 11 and 12), frequency band $\alpha(8 \sim 13$ Hz), $\beta(14 \sim 40$ Hz) are the most relevant to these signals and a small number of convolution kernels pay attention to the θ $band(4 \sim 7$ Hz), low-$\gamma(30 \sim 60$ Hz). α and β frequency bands contain the key information used to classify "Left hand" and "Right hand". And in recent studies, the θ frequency band and 6∼12 Hz have also been pointed out that can be used for the classification of motor imagination [45–47]. Low-γ is closely related to classification task of Feet and Tongue (Rest). When training on ME-EEG with a large mount of high-frequency γ signals, in addition to focusing on β and low-γ frequency bands, learned kernels also pay attention to $\alpha \sim \beta$ and $high-\gamma(60 \sim 90$ Hz) of signals (see Fig. 12), which is consistent with the result got by Schirrmeister, Borra et al. where frequency band $\alpha \sim \beta$ and high-γ can enhance decoding ME-EEG [48,49].

As underlined previously, the proposed Channel-Mixing-ConvNet is based on implicitly stacking temporal and spatial filters in the "Channel-mixing Conv2D" convolution layer to design, compared with Shallow-FbcspNet constructed in explicit stacking and WaSF-ConvNet, Sinc-ShallowNet introduced an interpretable convolutional to enhance decoding performance, Channel-Mixing-ConvNet does not decrease in the interpretability of feature learning, but instead captured a lot of time-channel features which drives to classify. In addition, implicit stacking disassembles the previously separated "temporal - spatial" convolution layers and merges them. Therefore, the additional parameters that were originally introduced by building these layers are omitted, and the interpretability of the entire CNN network is improved.

In order to explore the impact of implicit stacking on decoding, we selected Shallow-FbcspNet as the representative model of explicit stacking, and compared the feature maps of EEG signals captured by temporal and spatial convolution layers with the feature maps of signals from the "Channel-mixing Conv2D" convolutional layer. Fig. 13 shows spectrogram of the captured signals, it can been see that the intermediate signals extracted by the temporal convolutionl layer in Shallow-FbcspNet was extremely messy in the frequency domain, and it covers a wide frequency band. It seems that the determination of the most relevant frequency bands including EEG class-discriminative information is still in swing, and the following spatial convolution layer has improved the swing, so that the frequency band was contracted on an interested frequency band related to motor imagery. Although the Channel-Mixing-ConvNet was constructed in an implicit stacking way, the feature maps of signals from the spatial convolutionl layer in Shallow-FbcspNet are more obviously similar with the "Channel-mixing Conv2D" convolutionl layer than the temporal convolutionl layer (see Fig. 13), which indicates that the EEG signals after being convolved this layer were more stable in frequency domain, and their frequency band can shrink to the interested frequency bands related to motor imagery. In fact, the single temporal convolutionl layer can be regarded as a coarse-grained feature extraction. Spatial convolutionl layer introduces channel features to build a mapping between time domain and channel. Similarly, the "Channels-mixing Conv2D" convolutionl layer established the response between time domain and channel via only a single multi-channel convolution. Thus, although the models are constructed differently, they remain the same in terms of the ultimate intermediate representation of features.

Finally, the interpretability of the block Mixed Channel Process was verified. Fig. 14 shows that the visualization result of processing on feature maps of signals from block 1(see Fig. 12) obtained by block Mixed Channel Process in ME-EEG dataset. Compared with the "Channel-mixing Conv2D" convolutional layer, the depthwise convolution in block Mixed Channel Process has twice the number of convolution kernels. As can be seen from the figure, only five learned kernels in the "Channel-mixing Conv2D" convolutionl layer focused on γ band, while after processing by the block 2, the number of learned kernels focusing on this frequency band increased from 5 to 8. It indicates that the block Mixed Channel Process plays a role in refining and supplementing the key features in frequency domain.

5. Conclusion

In summary, we proposed the Channel-Mixing-ConvNet by implicitly stacking temporal–spatial convolutionl layer. The Channel Mixing block mixes the EEG temporal and channel features in a way of generating "new signals", which replaces the conventional separate temporal–spatial convolutionl layers. Compared with the classic EEG decoding CNN, it significantly reduces the amount of trainable parameters and the risk of overfitting; At the same time, the interpretability of the features captured by this layer is not inferior to the directly introduced interpretable convolutional layer. In addition to matching with the neurological phenomenon of motor imagery, the underlying features related to motor imagery decoding were captured. The experiments results on the public datasets proves that our proposed CNN is significantly better than the SOTA algorithm, and the lightweight structure allows stacking other layers to continue to exploit the potential performance of the model.

CRediT authorship contribution statement

Weifeng Ma: Conceptualization, Methodology, Resources, Supervision, Project administration, Writing - review & editing. **Yifei Gong:** Methodology, Software, Investigation, Writing – original draft. **Gongxue Zhou:** Validation, Visualization, Data curation. **Yang Liu:** Supervision, Funding acquisition. **Lei Zhang:** Resources, Supervision, Funding acquisition. **Boxian He:** Data curation, Validation.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.bspc.2021.103021.

Acknowledgments

Our research mainly funded by the Humanities and Social Sciences Foundation of the Ministry of Education of China (20YJA880034), Key Research and Development Program of Zhejiang Province, China (Grants No. 2020C03071), and partly supported by the National Undergraduate Innovation and entrepreneurship Training Program of China (Grants No. 202011057014).

References

[1] R. Jung, W. Berger, Hans bergers entdeckung des elektrenkephalogramms und seine ersten befunde 1924–1931, Arch. Psychiatr. Nervenkrankh. 227 (4) (1979) 279–300.

[2] Y. He, D. Eguren, J.M. Azorín, R.G. Grossman, T.P. Luu, J.L. Contreras-Vidal, Brain–machine interfaces for controlling lower-limb powered robotic systems, J. Neural Eng. 15 (2) (2018) 021004.

[3] A. Rakshit, A. Konar, A.K. Nagar, A hybrid brain-computer interface for closed-loop position control of a robot arm, IEEE/CAA J. Autom. Sin. 7 (5) (2020) 1344–1360.

[4] S. Guan, K. Zhao, S. Yang, Motor imagery EEG classification based on decision tree framework and Riemannian geometry, Comput. Intell. Neurosci. 2019 (2019).

[5] J. Luo, X. Gao, X. Zhu, B. Wang, N. Lu, J. Wang, Motor imagery EEG classification based on ensemble support vector learning, Comput. Methods Programs Biomed. 193 (2020) 105464.

[6] H.-J. Rong, C. Li, R.-J. Bao, B. Chen, Incremental adaptive eeg classification of motor imagery-based bci, in: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–7.

[7] M. Miao, W. Zhang, W. Hu, R. Wang, An adaptive multi-domain feature joint optimization framework based on composite kernels and ant colony optimization for motor imagery EEG classification, Biomed. Signal Process. Control 61 (2020) 101994.

[8] Y. Li, W.-G. Cui, M.-L. Luo, K. Li, L. Wang, High-resolution time–frequency representation of EEG data using multi-scale wavelets, Internat. J. Systems Sci. 48 (12) (2017) 2658–2668.

[9] P. Gaur, R.B. Pachori, H. Wang, G. Prasad, A multi-class EEG-based BCI classification using multivariate empirical mode decomposition based filtering and Riemannian geometry, Expert Syst. Appl. 95 (2018) 201–211.

[10] K.K. Ang, Z.Y. Chin, H. Zhang, C. Guan, Filter bank common spatial pattern (FBCSP) in brain-computer interface, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IEEE, 2008, pp. 2390–2397.

[11] K.K. Ang, Z.Y. Chin, C. Wang, C. Guan, H. Zhang, Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b, Front. Neurosci. 6 (2012) 39.

[12] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T.H. Falk, J. Faubert, Deep learning-based electroencephalography analysis: a systematic review, J. Neural Eng. 16 (5) (2019) 051001.

[13] A. Al-Saegh, S.A. Dawwd, J.M. Abdul-Jabbar, Deep learning for motor imagery EEG-based classification: A review, Biomed. Signal Process. Control 63 (2021) 102172.

[14] Y.R. Tabar, U. Halici, A novel deep learning approach for classification of EEG motor imagery signals, J. Neural Eng. 14 (1) (2016) 016003.

[15] N. Robinson, S.-W. Lee, C. Guan, EEG representation in deep convolutional neural networks for classification of motor imagery, in: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), IEEE, 2019, pp. 1322–1326.

[16] S. Sakhavi, C. Guan, S. Yan, Parallel convolutional-linear neural network for motor imagery classification, in: 2015 23rd European Signal Processing Conference (EUSIPCO), IEEE, 2015, pp. 2736–2740.

[17] Z. Tang, C. Li, S. Sun, Single-trial EEG classification of motor imagery using deep convolutional neural networks, Optik 130 (2017) 11–18.

[18] B. Xu, L. Zhang, A. Song, C. Wu, W. Li, D. Zhang, G. Xu, H. Li, H. Zeng, Wavelet transform time-frequency image and convolutional network-based motor imagery EEG classification, IEEE Access 7 (2018) 6084–6093.

[19] M. Xu, J. Yao, Z. Zhang, R. Li, B. Yang, C. Li, J. Li, J. Zhang, Learning EEG topographical representation for classification via convolutional neural network, Pattern Recognit. 105 (2020) 107390.

[20] H. Cecotti, A. Graser, Convolutional neural networks for P300 detection with application to brain-computer interfaces, IEEE Trans. Pattern Anal. Mach. Intell. 33 (3) (2010) 433–445.

[21] R.T. Schirrmeister, J.T. Springenberg, L.D.J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for EEG decoding and visualization, Hum. Brain Mapp. 38 (11) (2017) 5391–5420.

[22] S. Sakhavi, C. Guan, S. Yan, Learning temporal information for brain-computer interface using convolutional neural networks, IEEE Trans. Neural Netw. Learn. Syst. 29 (11) (2018) 5619–5629.

[23] V.J. Lawhern, A.J. Solon, N.R. Waytowich, S.M. Gordon, C.P. Hung, B.J. Lance, EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces, J. Neural Eng. 15 (5) (2018) 056013.

[24] X. Zhao, H. Zhang, G. Zhu, F. You, S. Kuang, L. Sun, A multi-branch 3D convolutional neural network for EEG-based motor imagery classification, IEEE Trans. Neural Syst. Rehabil. Eng. 27 (10) (2019) 2164–2177.

[25] Y. Li, H. Yang, J. Li, D. Chen, M. Du, EEG-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by grad-CAM, Neurocomputing 415 (2020) 225–233.

[26] L. Wang, W. Huang, Z. Yang, C. Zhang, Temporal-spatial-frequency depth extraction of brain-computer interface based on mental tasks, Biomed. Signal Process. Control 58 (2020) 101845.

[27] E. Dong, K. Zhou, J. Tong, S. Du, A novel hybrid kernel function relevance vector machine for multi-task motor imagery EEG classification, Biomed. Signal Process. Control 60 (2020) 101991.

[28] X. Tang, W. Li, X. Li, W. Ma, X. Dang, Motor imagery EEG recognition based on conditional optimization empirical mode decomposition and multi-scale convolutional neural network, Expert Syst. Appl. 149 (2020) 113285.

[29] Y. Li, X.-R. Zhang, B. Zhang, M.-Y. Lei, W.-G. Cui, Y.-Z. Guo, A channel-projection mixed-scale convolutional neural network for motor imagery EEG decoding, IEEE Trans. Neural Syst. Rehabil. Eng. 27 (6) (2019) 1170–1180.

[30] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, Digit. Signal Process. 73 (2018) 1–15.

[31] M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Towards better understanding of gradient-based attribution methods for deep neural networks, 2017, arXiv preprint arXiv:1711.06104.

[32] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, D.A. Keim, Towards a rigorous evaluation of XAI methods on time series, 2019, arXiv preprint arXiv:1909.07082.

[33] Z. Qin, F. Yu, C. Liu, X. Chen, How convolutional neural network see the world-A survey of convolutional neural network visualization methods, 2018, arXiv preprint arXiv:1804.11191.

[34] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013, arXiv preprint arXiv:1312.6034.

[35] A. Farahat, C. Reichert, C.M. Sweeney-Reed, H. Hinrichs, Convolutional neural networks for decoding of covert attention focus and saliency maps for EEG feature visualization, J. Neural Eng. 16 (6) (2019) 066010.

[36] D. Zhao, F. Tang, B. Si, X. Feng, Learning joint space–time–frequency features for EEG decoding on small labeled data, Neural Netw. 114 (2019) 67–77.

[37] D. Borra, S. Fantozzi, E. Magosso, Interpretable and lightweight convolutional neural network for EEG decoding: application to movement execution and imagination, Neural Netw. 129 (2020) 55–74.

[38] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.

[39] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.

[40] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, PMLR, 2015, pp. 448–456.

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

[42] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

[43] G. Pfurtscheller, C. Brunner, A. Schlögl, F.L. Da Silva, Mu rhythm (de) synchronization and EEG single-trial classification of different motor imagery tasks, NeuroImage 31 (1) (2006) 153–159.
[44] G. Pfurtscheller, D. Flotzinger, C. Neuper, Differentiation between finger, toe and tongue movement in man based on 40 Hz EEG, Electroencephalogr. Clin. Neurophysiol. 90 (6) (1994) 456–460.
[45] Y.-H. Liu, L.-F. Lin, C.-W. Chou, Y. Chang, Y.-T. Hsiao, W.-C. Hsu, Analysis of electroencephalography event-related desynchronisation and synchronisation induced by lower-limb stepping motor imagery, J. Med. Biol. Eng. 39 (1) (2019) 54–69.
[46] E. Weber, M. Doppelmayr, Kinesthetic motor imagery training modulates frontal midline theta during imagination of a dart throw, Int. J. Psychophysiol. 110 (2016) 137–145.
[47] H.-I. Suk, S.-W. Lee, Subject and class specific frequency bands selection for multiclass motor imagery classification, Int. J. Imaging Syst. Technol. 21 (2) (2011) 123–130.
[48] N.E. Crone, D.L. Miglioretti, B. Gordon, R.P. Lesser, Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band, Brain: J. Neurol. 121 (12) (1998) 2301–2315.
[49] M. Mirnaziri, M. Rahimi, S. Alavikakhaki, R. Ebrahimpour, Using combination of μ, β and γ bands in classification of EEG signals, Basic Clin. Neurosci. 4 (1) (2013) 76.