scientific reports

Check for updates

OPEN

Compact convolutional transformer for subject-independent motor imagery EEG-based BCIs

Aigerim Keutayeva[1]✉, Nail Fakhrutdinov[2] & Berdakh Abibullaev[3]

Motor imagery electroencephalography (EEG) analysis is crucial for the development of effective brain-computer interfaces (BCIs), yet it presents considerable challenges due to the complexity of the data and inter-subject variability. This paper introduces EEGCCT, an application of compact convolutional transformers designed specifically to improve the analysis of motor imagery tasks in EEG. Unlike traditional approaches, EEGCCT model significantly enhances generalization from limited data, effectively addressing a common limitation in EEG datasets. We validate and test our models using the open-source BCI Competition IV datasets 2a and 2b, employing a Leave-One-Subject-Out (LOSO) strategy to ensure subject-independent performance. Our findings demonstrate that EEGCCT not only outperforms conventional models like EEGNet in standard evaluations but also achieves better performance compared to other advanced models such as Conformer, Hybrid s-CViT, and Hybrid t-CViT, while utilizing fewer parameters and achieving an accuracy of 70.12%. Additionally, the paper presents a comprehensive ablation study that includes targeted data augmentation, hyperparameter optimization, and architectural improvements.

Keywords Brain–computer interface, EEG, Motor imagery, Compact convolutional transformers, Deep learning, Neural signal processing

The development of brain-computer interfaces (BCIs) offers a promising method for enabling users to control devices through their mental activities. BCIs provide communication channels for people with serious conditions such as amyotrophic lateral sclerosis (ALS) or locked-in syndrome[1–3]. Its potential extends to advanced prosthetics[4], immersive gaming[5], virtual reality[6]and broader scientific research[7–11].

A BCI system works through a carefully structured four-step process that converts brain signals into actionable commands. First, brain activity is recorded through the process of signal acquisition. Among the various techniques for acquiring brain signals, electroencephalogram (EEG) recordings are the most commonly utilized[12,13]. Due to its non-invasive nature, excellent temporal resolution, accessibility and cost-effectiveness, EEG plays a central role in many BCIs[14]. However, there is often a lot of noise in these signals, which can obscure real brain signals and make classification difficult. To overcome these challenges, the most meaningful features must be identified using advanced analytical techniques.

Moreover, there are additional difficulties because different participants have different signal configurations[15,16]. In order to enable personalized future use, the majority of BCIs in use today rely on subject-specific (subject-dependent) training, which requires significant amount of user time for system calibration and training[17,18]. This procedure is particularly time-consuming and complicated for people with disabilities[19]. The limited EEG data size - often consisting of only a few hundred samples - also represents a major obstacle to training traditional, data-hungry deep learning models[20].The size of the available data set is often limited by the meticulous nature of EEG data collection protocols[21,22], although some studies with large subject pools have produced positive results[23–25]. Therefore, current research efforts focus on creating either subject-independent or calibration-free systems as well as efficient classification techniques for small data sets[26–28].

Following the acquisition, the raw signals are pre-processed, which often contain noise or artifacts caused by muscle or eye movements. This step uses various methods to improve the quality of the signals by filtering and cleaning them for more accurate analysis[29]. To improve signal quality prior to classification, a number of denoising techniques have recently been developed and implemented to address this problem. Depending on the type of

[1]Institute of Smart Systems and Artificial Intelligence (ISSAI), Nazarbayev University, Astana 010000, Kazakhstan. [2]Department of Computer Science, Nazarbayev University, Astana 010000, Kazakhstan. [3]Department of Robotics Engineering, Nazarbayev University, Astana 010000, Kazakhstan. ✉email: aigerim.keutayeva@alumni.nu.edu.kz

noise and the characteristics of the EEG signals, these methods, which include wavelet transform[30–34], independent component analysis (ICA)[35], and empirical mode decomposition (EMD)[36], each have different advantages.

The subsequent phase, feature extraction, is crucial for identifying the most important attributes within the preprocessed signal. This phase often involves identifying variations in specific frequency bands, particularly the α and βrhythms, which are closely linked to motor imaging (MI). During the mental practice of a movement, the brain regions involved in the actual execution of that movement show analogous activation patterns. MI-based BCIs are designed to detect these patterns in the sensorimotor cortex and convert them into control commands for external devices[37].

The fundamental step in the translation process is classification. At this point, the extracted features are analyzed by machine learning algorithms to classify them into different mental states or intended actions[38]. It is important for decoding distorted brain signals so that BCI control can work[39–41]. Traditional models, especially ConvNet[42]and EEGNet[43], have been the basis for EEG signal analysis, mainly using convolutional neural network (CNN) architectures to focus on local temporal features[44–49]. However, due to the inherent limitations of their kernel size, these models are less able to capture the long-term dependencies that are essential for comprehensive interpretation of EEG signals. To address this problem, recurrent neural network (RNN) and long short-term memory (LSTM) models were introduced to capture temporal features relevant to EEG classification[50–53]. Despite their advances, these architectures have limitations in parallel training. Furthermore, the influence of dependencies, as processed by their hidden states, tends to decrease rapidly over successive time steps[13].

The research by Vaswani et al. has significantly advanced the field of attention-based models, most notably by introducing the Transformer architecture, which inherently perceives global dependencies[55]. This framework has been effectively used in various areas including natural language processing, computer vision, and EEG classification[56–59]. Expanding on this foundation, a recent investigation by Tao et al[60]. implemented a gated Transformer model for multi-class classification of EEG data. However, these models often overlook the importance of learning local features required for accurate EEG decoding. Consequently, additional feature extraction processes, such as activity mapping and spatial filtering, must be incorporated to address this shortcoming. In another study by Song et al[61]., authors used convolution techniques to capture local temporal and spatial features, followed by the adoption of self-attention mechanisms to capture global temporal features in Conformer model. In a separate study, Keutayeva et al[62]. constructed Hybrid s-CViT and Hybrid t-CViT models, where they used convolution to learn local temporal and spatial features and then adopted Vision Transformer using subject-independent evaluation with nested LOSO strategy. Despite their advancements, these models face challenges in two critical areas: subject independence and efficient handling of limited data. The Conformer still depends a lot on local feature extraction, meaning it may miss broader temporal correlations. Similarly, Hybrid s-CViT and Hybrid t-CViT may not translate well to different subjects and may not fundamentally solve the problem of limited data.

This paper introduces two versions of EEGCCT, an adaptation of the Compact Convolutional Transformer (CCT) model for EEG analysis in motor imagery tasks. The EEGCCT model distinguishes itself in several key aspects:

1. **Hybrid Model Structure:** EEGCCT combines the global, long-range perspective provided by Transformers with the local feature extraction capabilities of CNNs.
2. **Subject Independence:** EEGCCT emphasizes its ability to generalize across several subjects. This makes EEGCCT especially well-suited for a variety of BCI applications in which subject-specific training data may be scarce, particularly when assessed through the application of the LOSO approach.
3. **Handling Limited Data:** Enhancement in performance with a smaller parameter size is a major advantage of EEGCCT over models such as Conformer, Hybrid s-CViT, and Hybrid t-CViT.

The objectives of this paper are: to compare the EEGCCT model with existing approaches like EEGNet, Conformer, Hybrid s-CViT, and Hybrid t-CVIT in terms of structure, subject independence, and performance in limited data scenarios; to demonstrate the enhanced generalizability and efficiency of EEGCCT in subject-independent motor imagery EEG analysis using LOSO strategy; to highlight the improvements in performance through data augmentation, hyperparameter optimization, and architectural advancements in the EEGCCT model.

Methodology

Proposed EEGCCT model architecture

The frameworks of EEGCCT-1 and EEGCCT-2 adapted from[54] are depicted in Figs. 1 and 2, respectively.

The EEGCCT-1 model architecture is composed of three key components: a convolution tokenizer module and a transformer module. Within the convolution tokenizer module, raw two-dimensional EEG trials serve as the input, where spatial and temporal convolutional layers are applied along the electrode channels and time dimensions, respectively. An average pooling layer is then used to mitigate noise interference and enhance generalization. The reshaped spatial-temporal representation generated by the convolution module is subsequently fed into the transformer module. This transformer module includes positional embedding, a transformer encoder, sequence pooling (SeqPool), and an MLP block. It further extracts long-term temporal features by evaluating the global correlations between different time points within the feature maps. Finally, the MLP block produces the decoding results.

The EEGCCT-2 model follows a similar structure but introduces specific convolutional layers for initial feature extraction from EEG signals, followed by depth-wise and separable convolutions from EEGNet architecture[43] for enhanced feature representation. The extracted features are then processed by the transformer structure, similar to EEGCCT-1, for classification.

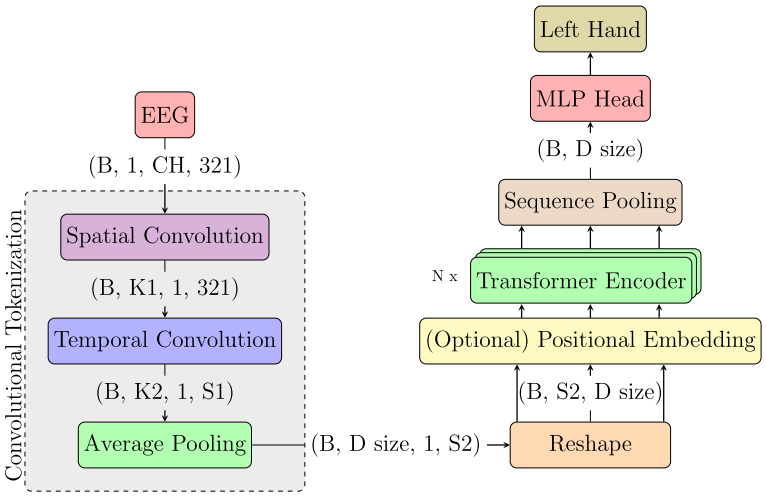

Fig. 1. EEGCCT-1 model architecture from[54] illustrating the process flow from EEG input, through convolution layers and Transformer encoders, to an MLP classification head for imaginary right (left)-hand movement.

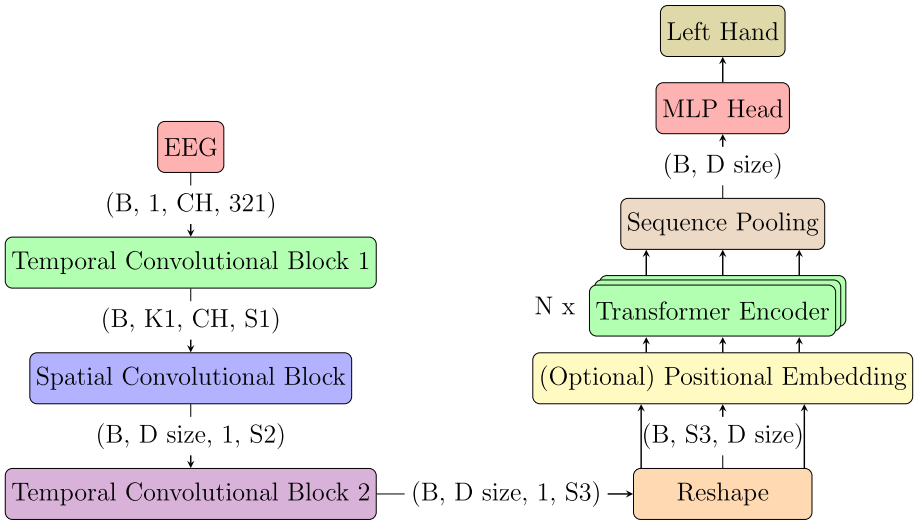

Fig. 2. EEGCCT-2 model architecture, combining methodologies from[43] and[54], processes EEG signals through convolutional blocks and Transformer encoders, concluding with an MLP head for classification. The architecture shows data flow and transformation across each stage.

Convolutional Tokenizer:

The Convolutional Tokenizer introduces an inductive bias into the model by substituting the traditional patch and embedding mechanism in the ViT[62] model with a simplified convolutional block. This block comprises two convolution layers and average pooling, structured as:

$$\mathbf{x}_0 = \text{AvgPool}\,(\,\text{Conv2d}\,(\mathbf{x})). \tag{1}$$

Here, the Conv2d operation employs *d*filters, matching the transformer backbone's embedding dimension. The convolution and average pool operations, potentially overlapping, preserve local spatial information and enhance performance by injecting inductive biases. Unlike models like ViT[62], this design is not strictly bound to input resolutions divisible by a predetermined patch size, offering enhanced flexibility.

The convolutional block aims to encode data into a rich, efficient latent representation, adjustable in downsampling ratio for further resolution reduction. This downsampling reduces the number of tokens, significantly cutting computational demands, albeit possibly at a performance cost. Despite this, the convolutional tokenizer, combined with SeqPool and the transformer encoder, forms the core of compact convolutional transformers, maintaining strong performance even with the potential removal of positional embedding.

SeqPool: The SeqPool is an attention-based technique that aggregates the output sequence of tokens. This process involves transforming the output sequence through a specific mapping operation: $\mathbb{R}^{b\times n\times d} \rightarrow \mathbb{R}^{b\times d}$. Given:

$$\mathbf{x}_L = f(\mathbf{x}_0) \in \mathbb{R}^{b \times n \times d}, \tag{2}$$

where $\mathbf{x}_L$ represents the output of an L-layer transformer encoder f. Here, b denotes the batch size, n signifies the sequence length, and d is the total embedding dimension. The output $\mathbf{x}_L$ is then processed through a linear layer $\mathbf{g}(\mathbf{x}_L) \in \mathbb{R}^{d \times 1}$, followed by the application of softmax activation:

$$\mathbf{x}'_L = \text{Softmax}\,(\mathbf{g}(\mathbf{x}_L)^T) \in \mathbb{R}^{b \times 1 \times n}. \tag{3}$$

This operation generates an importance weighting for each input token, which is subsequently applied as follows:

$$\mathbf{z} = \mathbf{x}'_L \mathbf{x}_L = \text{Softmax}\,(\mathbf{g}(\mathbf{x}_L)^T) \times \mathbf{x}_L \in \mathbb{R}^{b \times 1 \times n}. \tag{4}$$

By flattening, the output $\mathbf{z} \in \mathbb{R}^{b \times d}$ is generated, which can subsequently be passed through a classifier.

SeqPool enables the network to assign weights to the sequential embeddings of the latent space generated by the transformer encoder, facilitating the correlation of information across the input data. This process can be visualized as attending to the sequential data, where importance weights are assigned throughout the sequence only after processing by the encoder. Various versions of this pooling method have been evaluated, including both learnable and static approaches, with learnable pooling demonstrating better performance. Static methods, such as global average pooling, have also been explored by[62]. The learnable weighting is considered more effective because each embedded patch contains varying levels of entropy, allowing the model to assign weights to tokens based on the relevance of their information. Additionally, sequence pooling enables the model to better leverage information from spatially sparse data.

Transformer Encoder: Self-attention is a mechanism that utilizes three primary matrices: query $\mathbf{W}^q$, key $\mathbf{W}^k$, and value $\mathbf{W}^v$, as referenced in[55]. For each input element in an embedded sequence, three corresponding vectors - query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ - are generated through multiplication with these matrices:

$$\mathbf{Q}_i = \mathbf{x}_i \cdot \mathbf{W}^q, \tag{5}$$

$$\mathbf{K}_i = \mathbf{x}_i \cdot \mathbf{W}^k, \tag{6}$$

$$\mathbf{V}_i = \mathbf{x}_i \cdot \mathbf{W}^v \quad . \tag{7}$$

The scaled dot-product attention (SDPA) is computed using the Equation (8).

$$\text{SDPA}\,(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}. \tag{8}$$

By incorporating SDPA layers, which form the multi-head attention (MHA), the model can simultaneously focus on different representational aspects from various positions. The MHA is represented in Equation (9):

$$\text{MHA}\,(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\,\text{head}_1, ..., \text{head}_h], \tag{9}$$

$$\text{head}_i = \text{SDPA}\,(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) \quad . \tag{10}$$

These attention mechanisms can be integrated into various deep learning architectures, such as Transformers and CNNs, to enhance the model's effectiveness in processing EEG data. In our study, we employ EEG-based compact convolutional transformers (EEGCCT), as depicted in Figs. 1 and 2.

Datasets

This study uses open-access EEG datasets, specifically BCI IV 2a and BCI IV 2b, to analyze motor imagery tasks performed by different subjects. The primary aim of this investigation was to decode motor imagery data associated with the visualized movements of the left and right hands. Continuous EEG data were segmented into distinct 4-second intervals corresponding to the onset of mental imagery for both left-hand and right-hand tasks[63]. All datasets were preprocessed to reduce the output classes to two categories: left-hand and right-hand imagery. Additionally, a segmentation and reconstruction (S &R) method[61] was applied in the time domain, complemented by Z-score normalization. Details of each dataset are provided in Table 1.

Dataset	Subjects	EEG Ch.	SR (Hz)	Trials/session	Sessions	Labels
BCI IV 2a[64]	9	22	250	288	2	L/R
BCI IV 2b[65]	9	3	250	120	5	L/R

Table 1. Comparative summary of EEG datasets utilized in the study: This table provides a detailed overview of the BCI IV 2a and 2b datasets, including the designated labels for motor imagery tasks corresponding to left-hand (L) and right-hand (R) movements.

Model selection

This research implemented Leave-One-Subject-Out (LOSO) by training models on seven subjects, validating hyperparameters on one subject, and testing model performance on another, as illustrated in Fig. 3.

Given a collection of candidate BCI models $\mathscr{M} = \{\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_n\}$, each associated with a potential set of hyperparameters, the procedure is as follows:

1. For each model $\mathbf{M}_k \in \mathscr{M}$:

   (a) For each subject $\mathbf{s}_i$ in the dataset $\mathbf{S}$:

(i) Assign samples from 7 subjects to the training set, 1 subject to the validation set for hyperparameter tuning, and the remaining subject to the test set.
(ii) Perform hyperparameter tuning for $\mathbf{M}_k$ using LOSO cross-validation on the training and validation sets.
(iii) With the optimal hyperparameters identified, retrain $\mathbf{M}_k$ on the complete training set.
(iv) Evaluate $\mathbf{M}_k$ the samples exclusively from the test subject $\mathbf{s}_i$.
(v) Record the performance metric for each iteration as $P_{k,i}$.

   (b) Compute the average performance for $\mathbf{M}_k$ across all subjects:

$$P_{\text{LOSO},k} = \frac{1}{|\mathbf{S}|} \sum_{i=1}^{|\mathbf{S}|} P_{k,i} \tag{11}$$

2. Identify the model with the highest average performance:

$$\mathbf{M}^* = \arg \max_{\mathbf{M}_k \in \mathscr{M}} P_{\text{LOSO},k} \tag{12}$$

In BCI settings, it is essential to use LOSO cross-validation to make sure the models do not overfit to participant-specific EEG patterns. Furthermore, this technique ensures that models can generalize effectively across diverse user data. This approach, albeit being computationally demanding, yields unbiased estimates, demonstrating the adaptability of classification models to new and unobserved users. It is designed to minimize information leakage, providing a reliable representation of a model's capacity to interpret subject-to-subject variability.

Hyperparameter selection

In this section, the EEGCCT-1 and EEGCCT-2 architectures' hyperparameter search space is described. To maximize model performance, the right hyperparameter selection is crucial. Our extensive hyperparameter search was conducted with the following key assumptions:

- Systematic variations were made to the average pooling layer's kernel size $(1, k)$ and stride $(1, s)$. The combinations that were examined were as follows: (12, 3), (35, 7), (35, 10), (15, 3), (15, 7), (15, 5), and (15, 10).

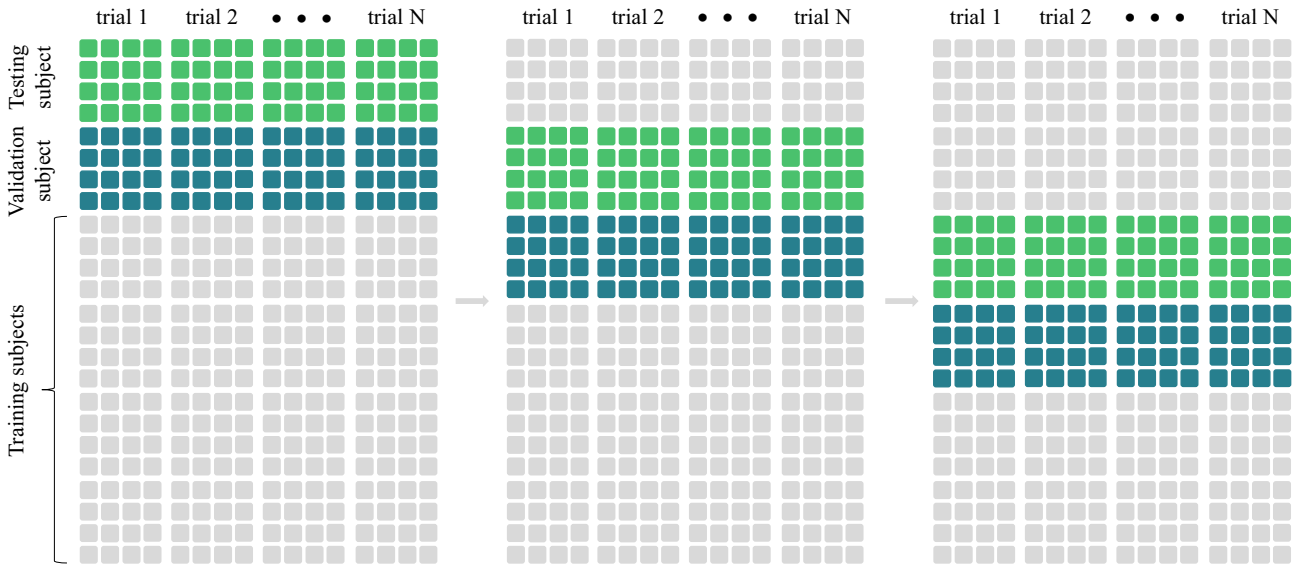

Fig. 3. LOSO model selection strategy used for model training and validation. Each panel represents a different test subject (light green), illustrating the use of all other subjects' data for training (light grey) and the specific subject's data for validation (dark green).

- The size of the transformer encoder layers was varied among three settings: 3, 6, and 9 layers. This variation aimed to balance between model depth and computational efficiency.
- The multi-head self-attention mechanism was configured with either 4 or 8 heads, to evaluate the impact of attention head count on model performance.
- The dimensionality of input embeddings was set to either 32 or 64, to assess the balance between embedding size and computational efficiency.
- The sizes of the Multi-Layer Perceptron (MLP) layers were tested at 64 and 128, to determine the optimal size for balancing model capacity and overfitting risks.
- A dropout rate of either 30% or 50%was applied prior to the fully connected layer to minimize overfitting[66].
- Positional embeddings were experimented with in three formats: learnable, sine, and none, to evaluate their influence on model performance.
- Each model variant was tested with and without data augmentation applied to the EEG data to evaluate the robustness of the models under varied data conditions.
- A batch size of 32 was utilized, and model training was limited to a maximum of 200 epochs, with early stopping criteria applied to prevent overfitting.
- The Adam optimizer was employed with a learning rate of 3×10^{-4}, and first and second-moment estimates set to 0.9 and 0.999, respectively[67].

The computational environment consisted of a Windows 10 workstation with an AMD Ryzen 7 5800H processor, 32 GB of RAM, and an NVIDIA GeForce RTX 3060 Laptop GPU. The PyTorch deep learning framework was used to build the complete workflow.

Our hyperparameter search and model evaluation results demonstrated that certain EEGCCT model configurations performed much better than others, particularly in situations with high subject variability and limited training data. The detailed discussion is covered in Sect. 3.2.

Results

Baseline comparison

This section thoroughly analyzes multiple state-of-the-art deep learning models used on the BCI Competition IV datasets 2a and 2b. Although these datasets have been widely used in previous research, direct comparison and reliable replication of results were limited by variations in experimental setups and techniques.

We used well-established models for this evaluation, such as Conformer[61], EEGNet[43], and the Hybrid s-CViT and t-CViT models[62]. These models were modified to correspond with our experimental setup and guarantee consistency with earlier research. This method enables a comprehensive assessment of the performance of each model independently, enabling a strong evaluation of the efficacy of the models across various physiological profiles.

The outcomes, which show the performance metrics of each model across various subjects, are presented in Tables 2 and 3.

Performance analysis on BCI2a dataset

Our analysis of the BCI2a dataset, specifically the classification accuracies as shown in Table 2, reveals that EEGCCT-1 and EEGCCT-2 are robust performers both with and without data augmentation. Without augmentation, EEGCCT-1 achieved exceptionally high accuracies, particularly reaching 90.62% with Subject 3 and averaging 68.75% across all subjects. EEGCCT-2's best performance was in Subject 4 with an accuracy of 68.06%, and an overall average of 66.55%. Despite having a lower parameter count, EEGNet recorded the highest average accuracy of 69.14%, demonstrating its efficiency. With data augmentation, EEGCCT-2 showed significant improvements, particularly in Subject 2, achieving an average accuracy of 69.14%. Meanwhile, EEGNet and

		Model performance (%)										
Condition	**Model**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**Avg.**	**Param**
No DA	**EEGCCT-1**	67.71	**60.76**	**90.62**	58.33	52.78	**69.10**	58.68	88.54	72.22	68.75	129.1k
	EEGCCT-2	64.93	**55.90**	86.45	**68.06**	52.08	63.54	56.94	84.72	66.32	66.55	85.3k
	EEGNet	80.90	50.69	89.93	66.67	51.74	62.85	62.85	83.33	73.26	69.14	20k
	Conformer	81.94	55.56	70.49	56.60	53.82	67.71	67.36	94.10	72.57	68.90	318.6k
	Hybrid s-CViT	55.90	55.21	50.69	49.65	52.08	52.78	50.00	50.35	57.29	52.66	989.4k
	Hybrid t-CViT	67.36	55.56	80.56	64.24	50.35	64.24	49.31	88.89	71.18	65.74	969.7k
With DA	**EEGCCT-1**	72.57	**63.89**	**86.46**	55.21	52.43	**69.79**	57.64	87.50	65.63	67.90	305.1k
	EEGCCT-2	78.82	**62.85**	77.43	64.58	55.21	68.40	57.99	87.85	69.10	69.14	85.3k
	EEGNet	82.29	54.51	80.56	64.93	50.35	66.67	61.46	79.17	69.79	67.75	20k
	Conformer	71.18	57.29	71.88	62.50	56.60	69.79	58.68	90.97	71.18	67.79	318.6k
	Hybrid s-CViT	52.08	52.43	54.17	54.86	52.78	57.99	58.33	57.29	54.17	54.90	989.4k
	Hybrid t-CViT	68.40	60.42	83.68	73.61	55.21	68.40	55.21	90.97	71.88	69.75	969.7k

Table 2. Comparative analysis of EEGCCT and benchmark models on BCI2a dataset with and without data augmentation (DA). Significant values are given in bold.

		Model performance (%)										
Condition	**Model**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**Avg.**	**Param**
No DA	**EEGCCT-1**	**75.14**	56.47	**58.06**	**82.70**	67.30	67.36	65.14	64.21	72.92	67.70	142.2k
	EEGCCT-2	68.89	55.00	**58.06**	75.81	**71.35**	75.69	**73.19**	**72.37**	**80.69**	**70.12**	65.8k
	EEGNet	70.28	59.12	53.89	75.27	70.54	67.22	71.53	67.89	80.00	68.42	19.3k
	Conformer	67.50	57.79	54.17	77.43	71.08	79.44	71.67	71.58	76.81	69.72	288.2k
	Hybrid s-CViT	65.97	55.15	51.94	68.51	59.05	58.33	56.81	58.03	63.89	59.74	405.7k
	Hybrid t-CViT	66.39	54.41	52.92	81.35	68.65	64.86	68.89	68.03	77.08	66.95	968.5k
With DA	**EEGCCT-1**	72.92	55.88	**56.11**	78.65	64.05	73.19	62.08	69.08	77.64	67.73	308k
	EEGCCT-2	72.08	57.21	**56.67**	76.89	**72.97**	74.03	**75.00**	66.97	80.69	**70.28**	65.8k
	EEGNet	71.11	60.15	55.42	75.54	72.29	77.22	72.08	66.32	80.97	70.12	19.3k
	Conformer	74.03	56.62	55.28	77.42	63.11	63.33	72.78	70.79	80.14	68.17	288.2k
	Hybrid s-CViT	68.47	56.91	50.42	81.08	60.68	61.67	62.22	70.00	68.47	64.44	405.7k
	Hybrid t-CViT	66.39	55.74	52.36	82.70	72.57	63.89	68.89	65.92	72.64	66.79	968.5k

Table 3. Comparative analysis of EEGCCT and benchmark models on BCI2b dataset with and without data augmentation (DA). Significant values are given in bold.

Conformer maintained stable performances, with average accuracies around 68%. The Hybrid t-CViT model's performance increased to an average accuracy of 69.75%, demonstrating the benefits of data augmentation in enhancing model robustness and performance.

Performance analysis on BCI2b dataset

The BCI2b dataset's classification accuracies, detailed in Table 3, show EEGCCT-2 as the leading performer without data augmentation, achieving the highest average accuracy of 70.12%, particularly performing well in Subject 9 with an accuracy of 80.69%. EEGNet and Conformer also demonstrated strong performances, with average accuracies of 68.42% and 69.72%, respectively. The Hybrid models, however, performed less effectively in this configuration. When data augmentation was applied, EEGCCT-2 maintained its leading position with a slight improvement to an average accuracy of 70.28%. EEGNet closely matched this performance with an average accuracy of 70.12%, particularly impressive in Subject 9 with an accuracy of 80.97%. The Conformer model showed modest improvements but remained slightly less effective than other listed models.

Ablation study and parameter sensitivity

This section details an extensive ablation study conducted on the EEGCCT models, EEGCCT-1 and EEGCCT-2, using the BCI Competition IV datasets 2a and 2b. The study focuses on assessing the impact of various architectural elements and hyperparameters, such as pooling kernel sizes, number of layers, attention head size, embedding dimensions, MLP sizes, and dropout rates, on the models' performance in both augmented and non-augmented settings.

1. **Standard Configuration:** Initially, the models were evaluated using their standard configurations to establish a baseline for comparison (details in Table 4).
2. **Kernel Size Variations:** Both EEGCCT-1 and EEGCCT-2 exhibited sensitivity to changes in pooling kernel sizes. Specifically, EEGCCT-1 showed optimal performance with kernel sizes of (15, 7) in dataset 2a under augmented conditions, achieving an average accuracy of 67.58%. EEGCCT-2 preferred different kernel configurations, suggesting variability in how each model processes spatial features. Figure 5 demostrates the pooling kernel size alterations of EEGCCT-1 for BCI2b dataset.
3. **Layer Variations:** Increasing the number of layers in EEGCCT-1 led to improved performance up to three layers on dataset 2a, but additional layers did not enhance accuracy, indicating diminishing returns beyond this point. In contrast, EEGCCT-2 achieved its highest accuracy with just one layer in the same dataset under augmented conditions, averaging 68.95%, which points to significant differences in the models' depth optimization.
4. **Attention Head Size:** Incrementing the size of the attention heads benefited both models, with EEGCCT-1 and EEGCCT-2 achieving their best performances at different head sizes (see Fig. 4).
5. **Embedding Dimension:** A larger embedding dimension of 64 improved performance for EEGCCT-1, and while EEGCCT-2 also showed benefits, the impact was less pronounced. This suggests that while increased embedding dimensions can enhance model capacity, the optimal dimensionality is model and dataset-specific.
6. **MLP Size:** Neither model demonstrated significant performance gains when increasing the MLP size from 64 to 128. This indicates that larger MLP configurations may not contribute to better performance and could potentially lead to overfitting or increased computational complexity without significant benefits.
7. **Dropout Rate:** Varying the dropout rate had a significant impact on performance. EEGCCT-1 achieved its best results with a 50% dropout rate in dataset 2a with augmentation, whereas EEGCCT-2 reached optimal

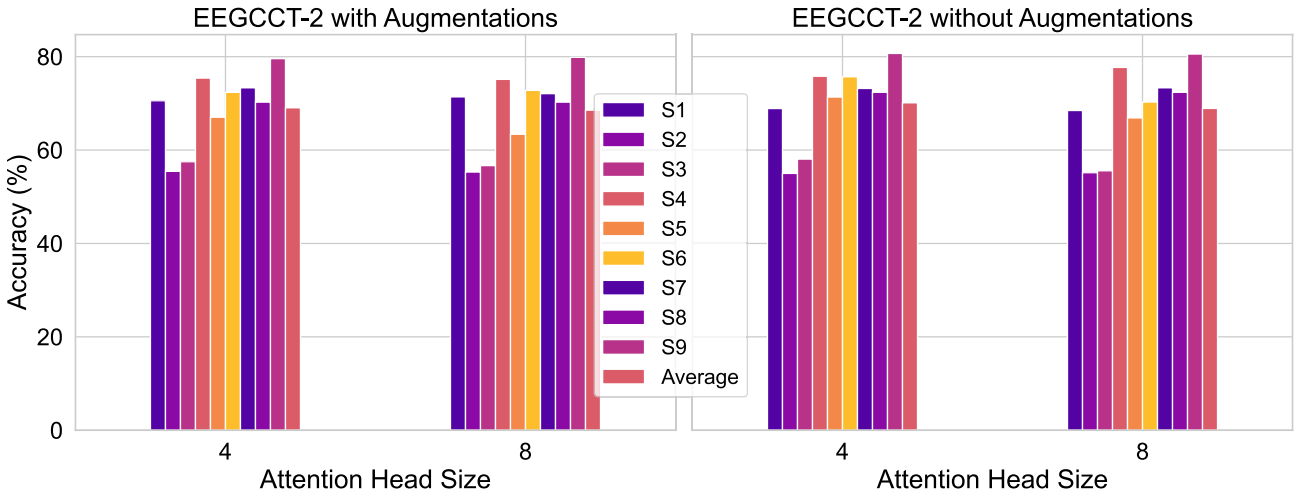

Fig. 4. Impact of transformer head size variation on EEGCCT-2 performance in BCI2b dataset.

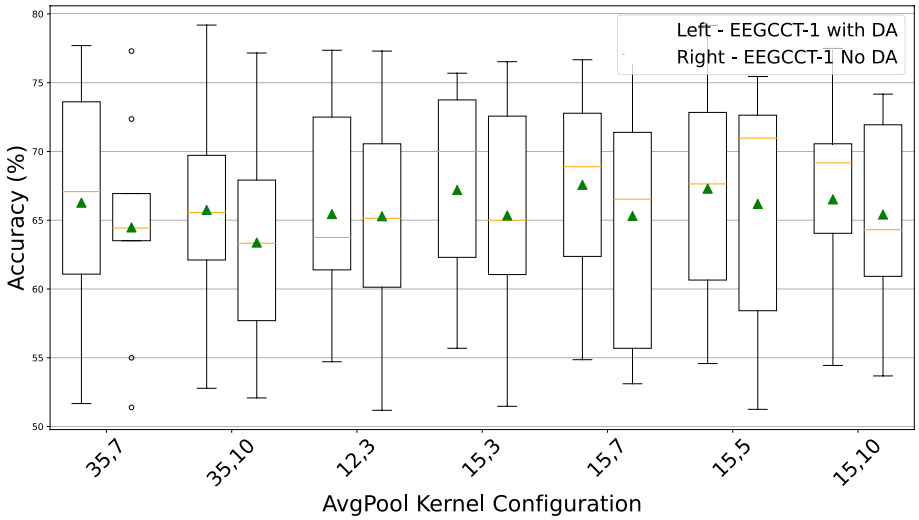

Fig. 5. Influence of average pooling (AvgPool) kernel size alteration on EEGCCT-1 performance in BCI2b dataset.

Model	AvgPool (k, s)	Encoder layer	Head size	Emb. dim.	MLP size	Dropout	Pos. Emb.
EEGCCT-1	(12, 3)	6	8	64	128	50%	Learnable
EEGCCT-2	–	3	4	32	64	50%	Sine

Table 4. Standard configuration for ablation study, providing a basis for understanding the architectural variations in EEGCCT-1 and EEGCCT-2.

performance with a 30% dropout rate under similar conditions. These findings highlight the need for model-specific regularization strategies to prevent overfitting while preserving model efficacy.

8. **Specific Results for EEGCCT-2:** EEGCCT-2 exhibited optimal performance on dataset 2a with augmentation when configured with a single encoder layer, a head size of 4, an embedding dimension of 32, an MLP size of 64, and a dropout rate of 30%. In contrast, for dataset 2b, the model achieved its best performance with three encoder layers and a 50% dropout rate. These findings highlight the significant differences in the model's architectural sensitivity across datasets (refer to Table 2).

This comprehensive ablation study demonstrates how important it is to consider various architectural and hyperparameter options when evaluating the efficacy of EEGCCT models. The findings highlight the need for customized model configurations in order to fully utilize these systems' potential across a variety of datasets.

Discussion

The EEGCCT models' effectiveness and robustness were assessed using the BCI Competition IV datasets 2a and 2b. Across a wide range of participants, EEGCCT-1 and EEGCCT-2 both continuously showed great performance, frequently outperforming both traditional and contemporary models. In particular, these models maintained high average accuracies while achieving remarkable subject-specific performance metrics, an important characteristic for reliable subject-independent motor imagery EEG analysis.

In applications where accurate individual subject analysis is crucial, EEGCCT-1 proved particularly successful in achieving high accuracies for certain participants. On the other hand, Table 3 shows how EEGCCT-2's minimal parameter demands along with its remarkable average accuracy highlight its applicability for more general BCI uses.

Importantly, the study highlighted the role of data augmentation in enhancing model performance. While effective without augmentation, both EEGCCT models showed significant improvements when trained with augmented data, suggesting an increased capability for generalization and robustness.

In comparison with EEGNet, a benchmark model in BCI research, the EEGCCT models often demonstrated enhanced performance. For example, EEGCCT-2 consistently outperformed EEGNet in terms of average accuracy on both datasets, augmented and non-augmented. Specifically, in the BCI2b dataset with augmentation, EEGCCT-2 achieved an average accuracy of 70.28%, slightly higher than EEGNet's 70.12%.

The EEGCCT models also compared favorably against other transformer-based models such as the Conformer, Hybrid s-CViT, and Hybrid t-CViT. Both EEGCCT-1 and EEGCCT-2 maintained competitive average accuracies with substantially fewer parameters, highlighting their efficient and lightweight architecture. This efficiency is particularly beneficial in BCI settings where computational resources and rapid response times are crucial.

The lightweight architecture of EEGCCT models, demonstrated by EEGCCT-2's use of only 65.8*k* parameters, supports not only resource efficiency but also the potential for real-time BCI applications. Collectively, these models with their compact structure and exceptional performance represent a significant advancement in EEG analysis for BCIs, providing a robust solution that generalizes well across subjects and adapts effectively under various conditions.

Conclusion

This study introduced EEGCCT models-Compact Convolutional Transformers optimized for motor imagery-based Brain-Computer Interface applications. Our comprehensive evaluations on the BCI Competition IV datasets 2a and 2b have underscored the efficacy of these models in delivering high-performance subject-independent EEG analysis with an efficient architecture. However, the performance of EEGCCT models might vary across different datasets and motor imagery tasks, suggesting the need for further validation in diverse settings. Additionally, the potential benefits of more sophisticated data augmentation techniques and the deeper integration of domain-specific insights into the model architecture remain areas promising for exploration.

Future research should consider adapting EEGCCT models for real-time BCI applications, employing transfer learning to further improve their subject-independent performances, and extending their application to other types of neural signal processing tasks. Moreover, integrating EEGCCT with other modalities in a multimodal BCI system could offer a path toward more resilient and versatile BCI solutions.

In summary, EEGCCT models represent a significant improvement in developing efficient and effective tools for EEG analysis in BCIs. Their combination of high accuracy, lightweight structure, and adaptability makes them well-suited for accurate BCIs.

Data availability

The datasets used during the current study are available in the BCI Competition IV repository. The specific datasets used are 2a[64]and 2b[65], which can be accessed at https://www.bbci.de/competition/iv/.

Code availability

The code has been released in https://github.com/aikesha/EEGCCT.

Received: 5 July 2024; Accepted: 20 September 2024
Published online: 28 October 2024

References

1. An, J. et al. The beginning of neurohaptics: Controlling cognitive interaction via brain haptic interface. In *2013 International Winter Workshop on Brain-Computer Interface (BCI)*, 103–106. https://doi.org/10.1109/IWW-BCI.2013.6506646 (2013).
2. Borgheai, S. B. *et al.* Enhancing communication for people in late-stage ALS using an fNIRS-based BCI system. *IEEE Trans. Neural Syst. Rehabil. Eng.* **28**, 1198–1207. https://doi.org/10.1109/TNSRE.2020.2980772 (2020).
3. Lesenfants, D. *et al.* An independent ssvep-based brain-computer interface in locked-in syndrome. *J. Neural Eng.* **11**, 035002. https://doi.org/10.1088/1741-2560/11/3/035002 (2014).
4. Muller-Putz, G. R. & Pfurtscheller, G. Control of an Electrical Prosthesis With an SSVEP-Based BCI. *IEEE Trans. Biomed. Eng.* **55**, 361–364. https://doi.org/10.1109/TBME.2007.897815 (2008).
5. van de Laar, B., Gürkök, H., Plass-Oude Bos, D., Poel, M. & Nijholt, A. Experiencing bci control in a popular computer game. *IEEE Trans. Comput. Intell. AI Games* **5**, 176–184. https://doi.org/10.1109/TCIAIG.2013.2253778 (2013).
6. Lier, E., Oosterman, J., Assmann, R., Vries, M. & Goor, H. The effect of virtual reality on evoked potentials following painful electrical stimuli and subjective pain. *Sci. Rep.* **10**, 1. https://doi.org/10.1038/s41598-020-66035-4 (2020).
7. Cui, G., Li, X. & Touyama, H. Emotion recognition based on group phase locking value using convolutional neural network. *Sci. Rep.* **13**, 3769 (2023).

8. Taheri Gorji, H. *et al.* Using machine learning methods and EEG to discriminate aircraft pilot cognitive workload during flight. *Sci. Rep.* **13**, 2507 (2023).
9. Akbari, H. et al. Recognizing seizure using poincaré plot of EEG signals and graphical features in DWT domain. *Brat. Med. J.* **124**, 12–24 (2023). https://doi.org/10.4149/BLL_2023_002
10. Sadiq, M. T., Akbari, H., Siuly, S., Li, Y. & Wen, P. Alcoholic EEG signals recognition based on phase space dynamic and geometrical features. *Chaos Solitons Fract.* **158**, 112036. https://doi.org/10.1016/j.chaos.2022.112036 (2022).
11. Akbari, H. et al. Depression detection based on geometrical features extracted from SODP shape of EEG signals and binary PSO. *Traitement du Signal* **38**, 43–46. https://doi.org/10.18280/ts.380102 (2021).
12. He, B., Baxter, B., Edelman, B. J., Cline, C. C. & Ye, W. W. Noninvasive Brain-Computer Interfaces Based on Sensorimotor Rhythms. In *Proceedings of the IEEE* **103**, 907–925. https://doi.org/10.1109/JPROC.2015.2407272 (2015). Conference Name: Proceedings of the IEEE.
13. Abibullaev, B., Keutayeva, A. & Zollanvari, A. Deep learning in EEG-based BCIs: A comprehensive review of transformer models, advantages, challenges, and applications. *IEEE Access* **11**, 127271–127301. https://doi.org/10.1109/ACCESS.2023.3329678 (2023).
14. Nicolas-Alonso, L. F. & Gomez-Gil, J. Brain computer interfaces, a review. *Sensors* **12**, 1211–1279 (2012).
15. Wan, Z., Yang, R., Huang, M., Zeng, N. & Liu, X. A review on transfer learning in EEG signal analysis. *Neurocomputing* **421**, 1–14. https://doi.org/10.1016/j.neucom.2020.09.017 (2021).
16. Vallabhaneni, R. B. *et al.* Deep learning algorithms in EEG signal decoding application: A review. *IEEE Access* **9**, 125778–125786. https://doi.org/10.1109/ACCESS.2021.3105917 (2021).
17. Saha, S., Ahmed, K. I. U., Mostafa, R., Hadjileontiadis, L. & Khandoker, A. Evidence of variabilities in EEG dynamics during motor imagery-based multiclass brain-computer interface. *IEEE Trans. Neural Syst. Rehabil. Eng.* **26**, 371–382. https://doi.org/10.1109/TNSRE.2017.2778178 (2018).
18. Singh, A., Lal, S. & Guesgen, H. Reduce calibration time in motor imagery using spatially regularized symmetric positives-definite matrices based classification. *Sensors* **19**, 379. https://doi.org/10.3390/s19020379 (2019).
19. Huang, X. *et al.* A review on signal processing approaches to reduce calibration time in EEG-based brain-computer interface. *Front. Neurosci.* **15**, 1. https://doi.org/10.3389/fnins.2021.733546 (2021).
20. Keutayeva, A. & Abibullaev, B. Data Constraints and Performance Optimization for Transformer-Based Models in EEG-Based Brain-Computer Interfaces: A Survey. *IEEE Access* **12**, 62628–62647. https://doi.org/10.1109/ACCESS.2024.3394696 (2024).
21. Mulder, T. Motor imagery and action observation: cognitive tools for rehabilitation. *J. Neural Transm.* **114**, 1265–1278 (2007).
22. Al-Saegh, A., Dawwd, S. A. & Abdul-Jabbar, J. M. Deep learning for motor imagery EEG-based classification: A review. *Biomed. Signal Process. Control* **63**, 102172. https://doi.org/10.1016/j.bspc.2020.102172 (2021).
23. Sadiq, M. T. *et al.* Toward the development of versatile brain-computer interfaces. *IEEE Trans. Artif. Intell.* **2**, 314–328. https://doi.org/10.1109/TAI.2021.3097307 (2021).
24. Yu, X., Aziz, M. Z., Sadiq, M. T., Fan, Z. & Xiao, G. A new framework for automatic detection of motor and mental imagery EEG signals for robust BCI systems. *IEEE Trans. Instrum. Meas.* **70**, 1–12. https://doi.org/10.1109/TIM.2021.3069026 (2021).
25. Dolzhikova, I., Abibullaev, B., Sameni, R. & Zollanvari, A. Subject-Independent Classification of Motor Imagery Tasks in EEG Using Multisubject Ensemble CNN. *IEEE Access* **10**, 81355–81363. https://doi.org/10.1109/ACCESS.2022.3195513 (2022).
26. Dong, Y. *et al.* Subject-independent EEG classification of motor imagery based on dual-branch feature fusion. *Brain Sci.* **13**, 1109. https://doi.org/10.3390/brainsci13071109 (2023).
27. Autthasan, P. *et al.* MIN2Net: End-to-end multi-task learning for subject-independent motor imagery EEG classification. *IEEE Trans. Biomed. Eng.* **69**, 2105–2118. https://doi.org/10.1109/TBME.2021.3137184 (2022).
28. Kwon, O.-Y., Lee, M.-H., Guan, C. & Lee, S.-W. Subject-independent brain-computer interfaces based on deep convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **31**, 3839–3852. https://doi.org/10.1109/TNNLS.2019.2946869 (2020).
29. Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M. & Muller, K.-R. Optimizing Spatial filters for Robust EEG Single-Trial Analysis. *IEEE Signal Process. Mag.* **25**, 41–56. https://doi.org/10.1109/MSP.2008.4408441 (2008).
30. Akhtar, M. T., Mitsuhashi, W. & James, C. J. Employing spatially constrained ICA and wavelet denoising, for automatic removal of artifacts from multichannel EEG data. *Signal Process.* **92**, 401–416. https://doi.org/10.1016/j.sigpro.2011.08.005 (2012).
31. Adeli, H., Zhou, Z. & Dadmehr, N. Analysis of EEG records in an epileptic patient using wavelet transform. *J. Neurosci. Methods* **123**, 69–87. https://doi.org/10.1016/S0165-0270(02)00340-0 (2003).
32. Amin, H. U. *et al.* Feature extraction and classification for EEG signals using wavelet transform and machine learning techniques. *Aust. Phys. Eng. Sci. Med.* **38**, 139–149 (2015).
33. Sadiq, M. T. *et al.* Motor imagery EEG signals decoding by multivariate empirical wavelet transform-based framework for robust brain-computer interfaces. *IEEE Access* **7**, 171431–171451. https://doi.org/10.1109/ACCESS.2019.2956018 (2019).
34. Sadiq, M. T., Yu, X., Yuan, Z. & Aziz, M. Z. Motor imagery BCI classification based on novel two-dimensional modelling in empirical wavelet transform. *Electron. Lett.* **56**, 1367–1369. https://doi.org/10.1049/el.2020.2509 (2020).
35. Janssen, N., Meij, M. V. D., López-Pérez, P. J. & Barber, H. A. Exploring the temporal dynamics of speech production with EEG and group ICA. *Sci. Rep.* **10**, 3667 (2020).
36. Sadiq, M. T. *et al.* Motor imagery BCI classification based on multivariate variational mode decomposition. *IEEE Trans. Emerg. Top. Comput. Intell.* **6**, 1177–1189. https://doi.org/10.1109/TETCI.2022.3147030 (2022).
37. Abiri, R., Borhani, S., Sellers, E., Jiang, Y. & Zhao, X. A comprehensive review of EEG-based brain-computer interface paradigms. *J. Neural Eng.* **16**, 1. https://doi.org/10.1088/1741-2552/aaf12e (2018).
38. Lotte, F. *et al.* A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update. *J. Neural Eng.* **15**, 031005 (2018).
39. Bang, J.-S., Lee, M.-H., Fazli, S., Guan, C. & Lee, S.-W. Spatio-spectral feature representation for motor imagery classification using convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **33**, 3038–3049. https://doi.org/10.1109/TNNLS.2020.3048385 (2022).
40. Xu, J., Zheng, H.-Z., Wang, J., Li, D. & Fang, X. Recognition of EEG signal motor imagery intention based on deep multi-view feature learning. *Sensors* **20**, 1. https://doi.org/10.3390/s20123496 (2020).
41. Ma, X., Wang, D., Liu, D. & Yang, J. DWT and CNN based multi-class motor imagery electroencephalographic signal recognition. *J. Neural Eng.* **17**, 016073. https://doi.org/10.1088/1741-2552/ab6f15 (2020).
42. Schirrmeister, R. T. *et al.* Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* **38**, 5391–5420 (2017).
43. Lawhern, V. J. *et al.* EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *J. Neural Eng.* **15**, 1 (2018).
44. Zhao, X. *et al.* A multi-branch 3D convolutional neural network for EEG-based motor imagery classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **27**, 2164–2177. https://doi.org/10.1109/TNSRE.2019.2938295 (2019).
45. Liu, T. & Yang, D. A densely connected multi-branch 3D convolutional neural network for motor imagery EEG decoding. *Brain Sci.* **11**, 1. https://doi.org/10.3390/brainsci11020197 (2021).
46. Khan, A. *et al.* A survey of the vision transformers and their cnn-transformer based variants. *Artif. Intell. Rev.* **56**, 1–54. https://doi.org/10.1007/s10462-023-10595-0 (2023).
47. Alwasiti, H., Yusoff, M. Z. & Raza, K. Motor imagery classification for brain computer interface using deep metric learning. *IEEE Access* **8**, 109949–109963. https://doi.org/10.1109/ACCESS.2020.3002459 (2020).

48. Xue, J. *et al.* A multifrequency brain network-based deep learning framework for motor imagery decoding. *Neural Plast.* **1–11**, 2020. https://doi.org/10.1155/2020/8863223 (2020).
49. Sadiq, M. T. *et al.* Exploiting pretrained CNN models for the development of an EEG-based robust BCI framework. *Comput. Biol. Med.* **143**, 105242. https://doi.org/10.1016/j.compbiomed.2022.105242 (2022).
50. Shi, X. et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Proc. 28th Int. Conf. Neural. Inf. Process. Sys.* - Volume 1, NIPS'15, 802–810 (MIT Press, Cambridge, MA, USA, 2015).
51. Tayeb, Z. *et al.* Validating deep neural networks for online decoding of motor imagery movements from EEG signals. *Sensors* **19**, 1. https://doi.org/10.3390/s19010210 (2019).
52. Wang, P., Jiang, A., Liu, X., Shang, J. & Zhang, L. LSTM-based EEG classification in motor imagery tasks. *IEEE Trans. Neural Syst. Rehabil. Eng.* **26**, 2086–2095. https://doi.org/10.1109/TNSRE.2018.2876129 (2018).
53. Kumar, S., Sharma, A. & Tsunoda, T. Brain wave classification using long short-term memory network based OPTICAL predictor. *Sci. Rep.* **9**, 1. https://doi.org/10.1038/s41598-019-45605-1 (2019).
54. Hassani, A. et al. Escaping the big data paradigm with compact transformers. arXiv preprint arXiv:2104.05704 (2021).
55. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 1 (2017).
56. Orken, M., Oralbekova, D., Alimhan, K., Tolganay, T. & Othman, M. A study of transformer-based end-to-end speech recognition system for Kazakh language. *Sci. Rep.* **12**, 1. https://doi.org/10.1038/s41598-022-12260-y (2022).
57. So, D. et al. Searching for efficient transformers for language modeling. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W. (eds.) Advances in neural information processing systems, vol. 34, 6010–6022 (Curran Associates, Inc., 2021).
58. Lanchantin, J., Wang, T., Ordonez, V. & Qi, Y. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16478–16488 (2021).
59. He, K. *et al.* Transformers in medical image analysis. *Intell. Med.* **3**, 59–78 (2023).
60. Tao, Y. et al. Gated transformer for decoding human brain EEG signals. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 125–130, https://doi.org/10.1109/EMBC46164.2021.9630210 (2021).
61. Song, Y., Zheng, Q., Liu, B. & Gao, X. EEG Conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Trans. Neural Syst. Rehabil. Eng.* **31**, 710–719. https://doi.org/10.1109/TNSRE.2022.3230250 (2023).
62. Keutayeva, A. & Abibullaev, B. Exploring the potential of attention mechanism-based deep learning for robust subject-independent motor-imagery based bcis. *IEEE Access* **11**, 107562–107580. https://doi.org/10.1109/ACCESS.2023.3320561 (2023).
63. Abibullaev, B., Fedorova, I. & Zollanvari, A. A brute-force CNN model selection for accurate classification of sensorimotor rhythms in BCIs. *IEEE Access* 1–1. https://doi.org/10.1109/ACCESS.2020.2997681 (2020).
64. Tangermann, M. et al. Review of the BCI competition IV. *Front. Neurosci.* **6**, https://doi.org/10.3389/fnins.2012.00055 (2012).
65. Leeb, R. et al. Brain-computer communication: Motivation, aim, and impact of exploring a virtual apartment. *IEEE Trans. Neural Syst. Rehab. Eng.* **15**, 473–482. https://doi.org/10.1109/TNSRE.2007.906956 (2007).
66. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
67. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. CoRR arXiv:abs/1412.6980 (2014).

Acknowledgements

This work was supported by the Institute of Smart Systems and Artificial Intelligence (ISSAI), OPISSAI2020001, Nazarbayev University, Astana 010000, Kazakhstan.

Author contributions

A.K. designed EEGCCT-1 model, performed the experiments, analyzed the data, and took the lead in writing the manuscript. N.F. designed EEGCCT-2 model, conducted the experiments, and created Figs. 1 and 2. B.A. supervised the project, and were in charge of overall direction and planning. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.