Development of a CNN-Based Motor Imagery Classification System from Electroencephalography (EEG) Signals Using Multi-Source Datasets

Tobias Natalio Sianipar*, Nur Ahmadi*†, Dessi Puji Lestari*†
*School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung, Indonesia
†Center for Artificial Intelligence, Bandung Institute of Technology, Bandung, Indonesia
Email: 13521090@std.stei.itb.ac.id, nahmadi@itb.ac.id, dessipuji@itb.ac.id

***Abstract*—A Convolutional Neural Network (CNN) is one of the methods that can be used to build a motor imagery classification system for electroencephalography (EEG) signals. Unfortunately, CNN, as a deep learning method, often shows poor classification performance on datasets with a limited number of trials. Although several techniques for increasing the number of trials have been proposed in previous studies, this research explores a new approach by combining trials from multiple source datasets (multi-source dataset) to improve the classification performance of models trained on small datasets. Since trials from different datasets have different EEG channel configurations, a channel harmonization method is applied in the proposed model. Meanwhile, the increased subject variation in the multi-source dataset is addressed through the implementation of the domain generalization method in the proposed model. The experiments begin by building a baseline model and evaluating the impact of various trial-augmentation methods on the dataset used to train the baseline model in terms of classification performance. The proposed model is then trained using the multi-source dataset. The results show that the baseline model achieves an average test accuracy of 0.6702, which is lower than the proposed model's average test accuracy of 0.7113 on the same test data. The use of multi-source datasets is also shown to be effective as a method for increasing the number of trials in the baseline model, although it does not outperform other trial-augmentation methods. Nevertheless, the proposed model is required to process multi-source datasets more effectively, thereby achieving a higher average accuracy compared to other trial-augmentation methods.**

***Index Terms*—CNN, motor imagery, multi-source dataset, EEG**

I. Introduction

Electroencephalography (EEG) are one of the methods used to record brain activity [1]. Motor imagery refers to the activity that requires the activation of brain signals responsible for motor movements, accompanied by the awareness to refrain from moving the corresponding body part [2]. Consciously performed motor imagery has a function equivalent to the unconscious preparation of body movements [3]. Therefore, motor imagery classification is a perfect BCI paradigm to restore body functionality through prosthetic limbs that can be controlled using brain signals.

The use of EEG signals for motor imagery classification has become an active research field in recent years [4]–[8]. The wide availability of open-source EEG datasets is one of the advantages of using EEG signals in research compared to other types of brain signal recordings. A trial in an EEG dataset is represented as a time series from multiple electrodes (channels) placed in different areas of the scalp. Each trial in the dataset records brain signal activity corresponding to a specific task. EEG recording is usually conducted with the help of visual cues that indicate the task the subject must perform [9].

The use of deep learning in classification models has been proven effective for EEG signal classification [10], [11]. A study benchmarking different deep learning architectures for motor imagery classification demonstrated that CNN architectures achieve the highest accuracy among other deep learning models [11]. To date, most CNN architectures for EEG signal classification published in the literature perform optimally when trained with datasets from a single source [11].

One study that investigated the combination of different classification models and multiple datasets hypothesized that deep learning models can achieve good accuracy if the number of trials per class per subject in the dataset exceeds 150 [10]. This is a significant limitation, as recording 150 trials per class per subject requires a considerable amount of time. Another drawback of CNN models trained on a single-source dataset is the possibility that the model may classify features in EEG signals based on BCI paradigms other than motor imagery [9].

A proposed solution to these issues is to combine datasets from different sources (multi-source dataset) with the same classification classes to increase the number of trials per class and enhance the variation of visual cues in the training data. However, differences in the number of channels, sampling rates, sessions, visual cues used, and subjects across multi-source datasets may lead to feature mismatches and reduced CNN accuracy. The problem caused by varying channel numbers is not trivial, a method must be implemented to enable CNN models to accept inputs with different channel configurations without losing spatial information.

The challenge posed by increased variation in sessions and subjects across multi-source datasets is also not straightforward; a method must be developed to train models that can capture information independent of subject and session. Developing CNN models for motor imagery classification on EEG signals with multi-source datasets may lead to findings on the significance of universal EEG features relevant to

motor imagery tasks. This underlines the importance of further research in this field to develop effective solutions for utilizing EEG datasets in motor imagery classification.

II. Related Works

Although no studies have yet developed an EEG signal classification model using multi-source datasets, several studies have explored related approaches. One approach to building an EEG signal classification model for motor imagery is to use a CNN. EEGNet is a CNN based architecture designed for EEG signals classification, consisting of three main convolutional layers: Conv2D, DepthwiseConv2D, and SeparableConv2D. In the Conv2D layer, the entire time series from each EEG channel is convolved temporally with N convolution kernels. The outputs from the Conv2D layer are then convolved spatially across EEG channels in the DepthwiseConv2D layer, producing a one-dimensional output. In the SeparableConv2D layer, the convolution results from the previous layer are combined in parallel and then convolved with M convolution kernels to generate features used for classification [12]. Experiments with the EEGNet model have shown that it can perform well without data augmentation. EEGNet is also effective at learning feature variations across different types of classification tasks in EEG datasets [12]. This method has been proven to work well for motor imagery classification.

Research on methods for handling EEG data with different channel numbers as well as methods for improving classification performance across different data domains has been published. Spatial attention is one of the methods developed to represent EEG signal data in such a way that the location and number of channels used to record the EEG signals can be generalized. A recent study on the effect of using spatial attention methods for EEG signal classification demonstrated improved performance on EEG datasets with varying numbers of channels [13]. However, the classification model in that study was built using spectral images, which discard part of the temporal information from the EEG signals being classified. Furthermore, the EEG datasets with different channel numbers used in that study were still derived from a single source, created by removing certain channels from some trials in the dataset.

Another study proposed a framework that can reduce the impact of domain differences on the extracted features of the dataset [4]. However, that research only focused on domain differences within a single-source dataset. In addition, the study only performed classification on the same subjects across different recording sessions. The domain generalization method can minimize the marginal distribution $P(X)$ of extracted features across different domains. However, this method is still less suitable for EEG signal classification models, which need to preserve the separability of the conditional distribution $P(Y \mid X)$ in the features extracted from different classification classes within a given domain. The use of this method has been proven to improve classification accuracy on EEG signals recorded in different sessions.

III. Datasets

The development of motor imagery classification models using EEG signals requires sufficiently large and well-known datasets in order to enable comparison between the accuracy of existing models and the newly developed model. Difficulties in accessing, as well as the limited knowledge of how to use EEG recording devices, are also reasons for relying on existing datasets. To simplify access to the existing dataset, the MOABB framework is utilized for efficient retrieval and standardized handling of the data [10]. The datasets used in this research are as follows:

A. BNCI2014_001

This dataset contains 4 classification classes: imagining the movement of the right hand, left hand, both feet, and tongue. Recordings were conducted using visual cues in the form of arrows pointing up, down, left, or right, which respectively represent tongue, both feet, left hand, and right hand. In addition, an acoustic tone was used as a marker to indicate the start of each recording. Each trial was recorded using 22 Ag/AgCl electrodes with an inter-electrode distance of 3.5 cm, along with 3 monopolar EOG electrodes for denoising eye movement artifacts during preprocessing. Data were collected from 9 subjects. Each subject participated in 12 trials for each classification class, repeated 6 times per session. Two recording sessions were conducted on different days. The EEG signal time series in this dataset was sampled at 250 Hz for 4 seconds. The data were processed with a bandpass filter between 0.5 Hz and 100 Hz, as well as a notch filter at 50 Hz [14].

The EEG data in this dataset are divided into two parts: data from the first recording session and data from the second session. In each session, 6 long EEG signals were recorded for each subject. Each long EEG signal consists of 48 motor imagery trials of a given class, each lasting 4 seconds and separated by 2 seconds of preparation and 1.5 seconds of rest. Timestamps were provided to indicate trial segments and non-trial segments, facilitating the extraction of trials from each long signal to be used as rows in the motor imagery dataset. Each trial in this dataset contains 22 rows of time series, each storing 1000 floating-point values of the electrical potential captured by the EEG channels.

B. Zhou2016

This dataset contains 3 classification classes: imagining the movement of the right hand, left hand, and feet. Each trial was recorded using 14 Ag/AgCl electrodes. Data were collected from 4 subjects using visual cues in the form of arrows pointing right, left, or down, which respectively represent right-hand, left-hand, and foot movements. Each subject participated in 25 trials for each classification class, with 2 repetitions per session, across 3 sessions. The intervals between sessions varied from day to month. The EEG time-series signals in this dataset were recorded at a frequency of 250 Hz for a duration of 5 seconds [15].

The EEG data in this dataset is divided into three sections. In each recording session, 2 long EEG signals were collected from each subject. Each long EEG signal consists of 75 motor imagery trials of a given class, each lasting 5 seconds and separated by 1 second of preparation and 4 seconds of rest. Timestamps were provided to indicate trial segments and non-trial segments, making it easier to extract trials from each long signal to be used as rows in the motor imagery dataset. Each trial in this dataset contains 14 rows of time series, each storing 1250 floating-point values of the electrical potential captured by the EEG channels.

C. Weibo2014

This dataset contains 6 classification classes: imagining the movement of the right hand, left hand, right foot, both hands, a combination of right hand and right foot, and a combination of left hand and right foot. Recordings were conducted using a red circle as a visual cue for preparation, followed by special characters to indicate the task to be performed. Data were collected from 10 subjects. Each trial was recorded using 64 electrodes. Each subject participated in 80 trials per classification class. The EEG time-series signals in this dataset were recorded at a frequency of 200 Hz for 4 seconds [16].

The EEG data in this dataset consists of only one session. Each subject has one long EEG recording. Each long EEG signal consists of 80 motor imagery trials of a given class, each lasting 4 seconds and separated by 3 seconds of preparation and 1 second of rest. Timestamps were provided to mark trial segments and other parts, making it easier to extract trials from each long signal to be used as rows in the motor imagery dataset. Each trial in this dataset contains 60 rows of time series, each storing 800 floating-point values of the electrical potential captured by the EEG channels.

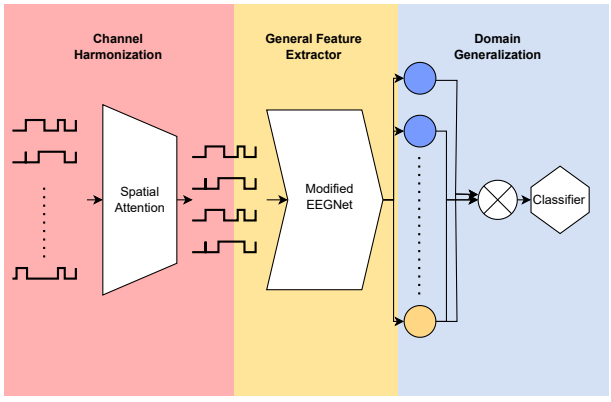

Fig. 1. Proposed Model Pipeline.

IV. Methodology

To develop the proposed model capable of efficiently processing multi-source datasets, a pipeline must be designed that can handle input trials with varying numbers of channels while remaining robust to signal variations across the multi-source dataset. The proposed model pipelines are shown in Figure 1. The proposed model consists of three main components as follows:

A. Channel Harmonization

This component serves as a bottleneck that connects various channel configurations from different datasets into a single configuration that can be processed by the subsequent component. It utilizes the spatial attention method described in [17]. The spatial attention method is more suitable because it can handle different numbers of channels without expert intervention by learning the spatial relationships of EEG signal channels during the training process. Spatial attention takes channel positions as input information, making differences in electrode placement standards across datasets less problematic. The use of spatial attention also enables an end-to-end model implementation, allowing for more flexible dataset combinations.

$$a(x,y) = \sum_{k=1}^{K}\sum_{l=1}^{K} \Re\Big(Z_j^{(k,l)}\Big)\cos\big(2\pi(kx+ly)\big) + \Im\Big(Z_j^{(k,l)}\Big)\sin\big(2\pi(kx+ly)\big) \quad (1)$$

Spatial attention takes trial with any number of channels and output virtual trial with predetermined number of channels. The time series for each channel are weighted according to the location of the channel using Equation (1).

$$\forall j \in \{1,\ldots,D_1\},\ SA(X)^{(j)} = \frac{\frac{1}{C}\sum_{i=1}^{C} e^{\,a_j(x_i,y_i)} X^{(i)}}{\sum_i e^{\,a(x_i,y_i)}} \quad (2)$$

Then a virtual time series is generated by all time series in a trial with the corresponding complex matrix Z_j. Each virtual time series is generated using Equation (2). There are two hyperparameters tuned in this component: the size of the complex matrix Z_j and the number of output channels produced by this component.

B. General Feature Extractor

This component functions to extract features from the output of the previous component. EEGNet has been proven to be the most robust CNN method [12]. Based on this knowledge, the general feature extractor component adopts the EEGNet architecture, modified by adding kernels of different sizes in the initial Conv2D layer, following the approach proposed in [4].

C. Domain Generalization

This component utilizes the EEG-DG method described [4]. To address the differences in EEG signal characteristics caused by variations in subjects and recording sessions, several methods that can be considered as solutions include domain adaptation and domain generalization. One drawback of domain adaptation is the potential loss of the conditional distribution in the features extracted by deep learning models. This issue is one of the reasons for the emergence of the domain generalization method, which is designed to preserve

the conditional distribution. Based on this understanding, domain generalization is a more suitable choice for addressing the problem of differences in EEG signal characteristics. This component takes the features extracted by the previous component and produces the final features, along with several domain specific features used to calculate the values of various loss functions required for model training. We use each unique subjects to describe domain in this study. The final features extracted in this component are then used as input to the softmax function, which performs classification. There are three hyperparameters tuned in this component: the weights for the margin-invariant loss, condition-invariant loss, and domain loss functions.

V. Experiments Details and Analysis

In this study, we conduct experiments to evaluate the effectiveness of using multi-source datasets as a trial augmentation method, with EEGNet serving as the baseline model. Furthermore, we experiment with the proposed model to assess its improvements over the baseline in processing multi-source datasets.

A. Dataset Preparation

The EEG data underwent several preprocessing steps: bandpass filtering between 8–30 Hz, resampling to 200 Hz, and the addition of domain labels by embedding subject identifiers into trial labels. The frequency band 8-30 Hz was chosen as it contain most information related to human motor system [18]. Only the overlapping classes—feet, left hand, and right hand—were retained, with the first 4 seconds of each trial preserved. Baseline correction was applied using 0.2 s of pre-trial activity, followed by Z-score normalization to reduce noise while maintaining relevant temporal patterns and ensuring robustness across subjects and datasets. For dataset splitting, Zhou2016 used the third session of each subject as test data, BNCI2014_001 used the second session, while Weibo2014, which only has one session, was randomly divided into 50% test and 50% training/validation (seed=42). Training and validation data were further split into 70% and 30% respectively (seed=42). Across all datasets, the number of trials and class distributions were kept balanced among training, validation, and test sets.

B. Multi-source Dataset Effectiveness Comparisons

To demonstrate the effectiveness of the proposed technique, several baseline models were constructed. Rebuilding the baseline models allowed tuning and experimentation to be carried out more flexibly. The use of multi-source datasets to train the baseline models was also intended to compare the effectiveness of the proposed method against the baseline in processing multi-source data. Therefore, to evaluate the effectiveness of using multi-source datasets, the baseline experiments were divided into the following scenarios:

- The baseline model was trained using the common method. In this experiment, one model was built and trained per subject in each dataset.
- The baseline training process followed the previous experimental scenario, but with the addition of data augmentation techniques applied to the training data. The augmentation methods used were recombination in time [19].
- The baseline model was trained using all subjects within each dataset. In this experiment, one model was built per dataset.
- The baseline model was trained using multi-source datasets. Differences in input sizes caused by using multi-source datasets were resolved through padding.

Each model is trained for 500 epochs, while retaining only the version that achieves the best validation accuracy. A mini-batch of size 30 is used for each epoch. The dropout rate for each dropout layer in the EEGNet architecture is set to the recommended value of 0.25.

The application of a multi-source dataset as training data for the model has also proven to be quite effective compared to training without trial augmentation as shown in Table I. However, based on the experiments conducted on the baseline model, it can be seen that the accuracy improvement obtained is still insufficient to match other trial augmentation methods when classifying test data from all dataset sources. This is assumed to be due to the suboptimal performance of the baseline model in addressing the challenges that arise in processing multi-source datasets.

The performance gap of the baseline model when trained on multi-source datasets can be attributed to two major factors. First, the heterogeneity in EEG channel configurations across datasets. The three datasets employed in this study differ in both electrode placement and channel count, resulting in trials with inconsistent input dimensions. In addition to dimensional discrepancies, each channel conveys essential spatial information reflecting scalp electrode positions, which is critical for classification. Addressing this issue requires a method capable of generalizing EEG signals with varying channel counts while preserving both spatial and temporal information.

Second, the intrinsic variability in EEG signal characteristics across datasets. Such variability arises from biological differences among subjects, session-specific recording conditions, and the nature of the visual cues presented during acquisition. Even EEG signals recorded in different sessions from the same subject can exhibit distinct characteristics due to factors such as changes in electrode placement, mental state, or environmental conditions [20]. These differences complicate the task of discovering stable discriminative patterns across datasets. Consequently, more advanced methods are necessary to enable classifiers to generalize effectively despite multi-source dataset variability.

C. Proposed Model on Multi-source Dataset

The proposed model pipeline is specifically designed to address the challenges that limit the effectiveness of the baseline model in processing multi-source datasets. Unlike the baseline, which struggles with heterogeneous channel configurations and signal variability, the proposed pipeline

TABLE I
MODELS CLASSIFICATION ACCURACY FOR VARIOUS TEST DATA SOURCE

No.	Test Data	EEGNet				Proposed Model
		Training Method				
		Per Subject	Per Subject + Recombination in Time	Per Dataset	Multi-source dataset	Multi-source dataset
		Accuracy				
1	Weibo2014	0.5202	0.5792	0.6077	0.5945	**0.6645**
2	BCI2014_001	0.6471	**0.6692**	0.6492	0.6019	0.6512
3	Zhou2016	0.8433	0.8550	**0.8767**	0.8450	0.8183
4	Rata-rata	0.6702	0.7011	0.7112	0.6805	**0.7113**

uses spatial attention to normalize input dimensions across datasets while preserving the spatial and temporal information embedded in EEG signals. Furthermore, it incorporates EEG-DG to enhance robustness against inter-dataset variability, enabling the model to identify consistent discriminative patterns despite differences in biological factors, recording sessions, and experimental conditions.

The proposed model uses a combination of weights (0.1, 0.1, 0.15) for the margin-invariant loss, condition-invariant loss, and domain loss, respectively. The number of channel for each virtual trial generated by Channel Harmonization component is set to 46 and the size of each complex matrix Z_j is set to 48×48. The model is trained for 500 epochs, while retaining only the version that achieves the best validation accuracy. A mini-batch of size 30 is used for each epoch. The training process utilizes Adam Optimizer with a learning rate of 0.0001.

TABLE II
STATISTICS COMPARISON OF THE ZHOU2016 DATASET FEATURE EXTRACTED BY THE PROPOSED MODEL AND THE BASELINE MODEL

Statistic	Proposed Model Features	Baseline Model Features
F-statistic	58.4206	218.1528
p-value	6.7338×10^{-24}	7.6041×10^{-72}

Based on Table I, the training of the proposed model using a multi-source dataset successfully improved classification accuracy on test data from the dataset with the fewest trials (Weibo2014) much more effectively compared to other trial augmentation methods. However, the classification accuracy of the proposed model trained on the multi-source dataset experienced a significant decrease on the Zhou2016 dataset. This is assumed to occur because the model ignores dataset-specific artifacts that are quite useful for classification on Zhou2016, which already has a large number of trials per class. It can also be seen in Table II that the features from the Zhou2016 dataset extracted by the proposed model have smaller F-statistics and larger p-values compared to the features extracted by the baseline model. This indicates that the class features of Zhou2016 extracted by the baseline model are easier to classify than those extracted by the proposed model.

D. Proposed Model on Single Dataset

In this section, experiments are conducted to determine the influence of multi-source datasets on the proposed model. These experiments aim to demonstrate that the improvement in classification performance of the proposed model is not merely due to increased model complexity, but because the model can process multi-source datasets more efficiently. The evaluation is performed by training the proposed model using subsets of multi-source datasets that were used in the previous scenarios.

In line with the assumption being tested—regarding the improvement in classification performance for subjects with a low number of trials, the data subset used to train the model in this scenario comes from the dataset with the fewest trials per class (Weibo2014). The hyperparameters of the proposed model are set according to the configuration that achieved the best performance in previous evaluation scenarios. The resulting classification accuracy on the same test data from Weibo2014 is **0.5344**, which is significantly lower than when the model is trained using a multi-source dataset.

Based on that result it can be concluded that the proposed model cannot match the baseline model performance if it is trained without using multi-source datasets. it can be observed that the model trained using multi-source datasets achieves significantly better classification performance than the model trained on a single dataset for each subject in the dataset with a low number of trials per class. This demonstrates that the improvement in classification performance on that dataset is not solely due to increased model complexity. This evaluation scenario shows that multi-source datasets have a substantial impact on improving accuracy for datasets with a low number of trials per class.

E. Proposed Model Extracted Feature

By using a multi-source dataset, the proposed model successfully identified common characteristics of EEG signals across datasets. However, the dominant common characteristics are assumed to originate from the EEG signals of subjects who already achieved relatively high classification accuracy before being trained with the multi-source dataset. This assumption is supported by the classification class regions in the t-SNE extracted feature projection of the subjects with the highest accuracy in each dataset shown in Figure 2, which are not only well-clustered but also exhibit

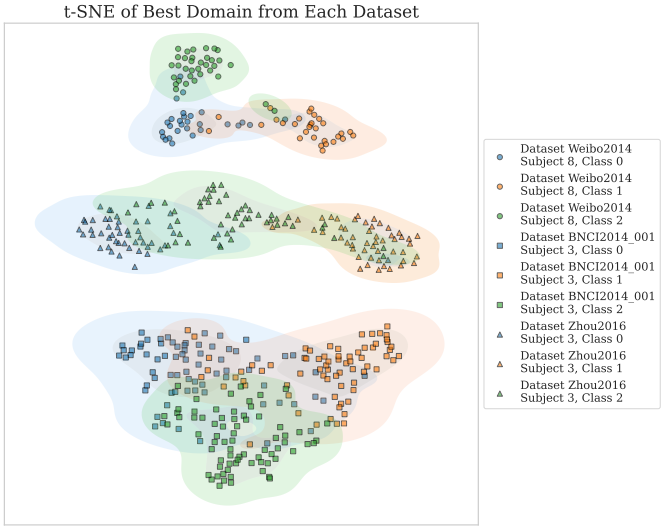

Fig. 2. Proposed Model Extracted Feature t-SNE Map.

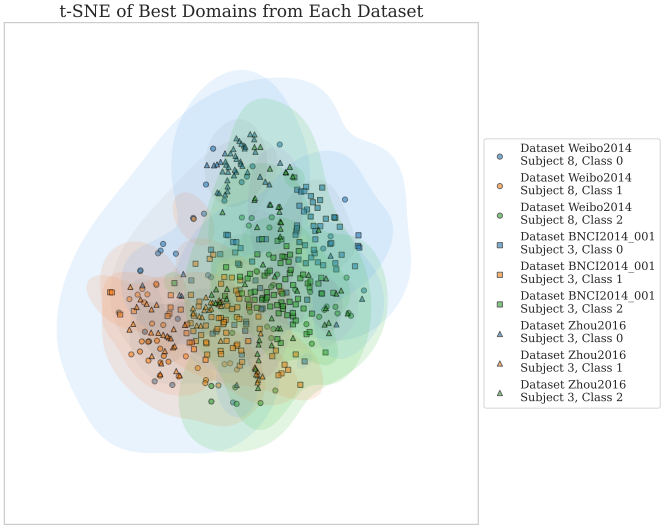

Fig. 3. Proposed Model Extracted Feature t-SNE Map.

a consistent ordering pattern of classification classes. The feature used in this analysis is the output from the last layer before the classification layer. To ensure that the clarity of the classification class regions is indeed caused by general characteristics, the distribution of the t-SNE projection of subject features as extracted by the baseline model trained using the entire dataset is compared with the distribution of the extracted feature projections from the proposed model. Based on the t-SNE projection results shown in Figure 2 and Figure 3, the distribution of feature projections produced by the proposed model exhibits clearer class-specific regions compared to the extracted features generated by the baseline model. The cause of bias toward general characteristics in subjects with relatively good classification accuracy is difficult to determine, as all EEG signals produced by each subject have similar variance, and the subject with the lowest variance does not necessarily guarantee the best classification accuracy.

TABLE III
PROPOSED MODEL CLASSIFICATION ACCURACY AND F1-SCORE PER SUBJECT FOR EACH DATASET

Subject	Accuracy	F1-Score		
		feet	left_hand	right_hand
Weibo2014				
1	0.5612	0.7671	0.5385	0.2667
2	0.5882	0.6301	0.5556	0.5581
3	0.4894	0.7273	0.4615	0.1818
4	0.6869	0.8493	0.5902	0.5938
5	0.6957	0.8197	0.6038	0.6571
6	0.6386	0.4898	0.7119	0.6897
7	0.7816	0.7419	0.7857	0.8214
8	0.9239	0.8966	0.9310	0.9412
9	0.8085	0.9474	0.8052	0.6667
10	0.4725	0.5714	0.3256	0.4737
Overall	0.6645	0.7465	0.6353	0.6055
BNCI2014_001				
1	0.7963	0.8696	0.7719	0.7480
2	0.6296	0.7451	0.5696	0.5620
3	0.8519	0.8000	0.8497	0.9028
4	0.5972	0.6621	0.5963	0.5238
5	0.4259	0.1758	0.4565	0.5350
6	0.5370	0.5714	0.6108	0.3784
7	0.6574	0.8402	0.5634	0.5124
8	0.6574	0.5570	0.7361	0.6923
9	0.7083	0.7093	0.8750	0.5000
Overall	0.6512	0.6798	0.6643	0.6023
Zhou2016				
1	0.8467	0.8367	0.8800	0.8235
2	0.7533	0.9263	0.7244	0.5897
3	0.8133	0.7556	0.9143	0.7619
4	0.8600	0.8909	0.8952	0.7765
Overall	0.8183	0.8550	0.8467	0.7459

F. Subjects Centroid Influence on Classification

The accuracy of each subject, even within the same dataset, varies considerably. To analyze the differences in accuracy for each subject, the centroid visualization of the subject's trial features for each classification class was carried out using t-SNE. The feature used in this analysis is the trial feature directly from the EEG signal.

As shown in Figure 4, there are several subjects from each dataset whose feature centroids are relatively far from the global centroid of the 'feet' classification class. In the Weibo2014 dataset, subjects 2, 3, and 6 have centroids that are positioned relatively far from the global centroid of the 'feet' classification class. In the BNCI2014_001 dataset, subjects 2, 5, 7, and 9 also have centroids located relatively far from the global centroid of the 'feet' classification class. Nevertheless, being far from the global centroid does not necessarily mean that these subjects are difficult to classify for the 'feet' class. This can be confirmed by the relatively good F1-scores for the 'feet' class obtained by most of these subjects, as shown in Table III. Based on this analysis, it can be concluded that

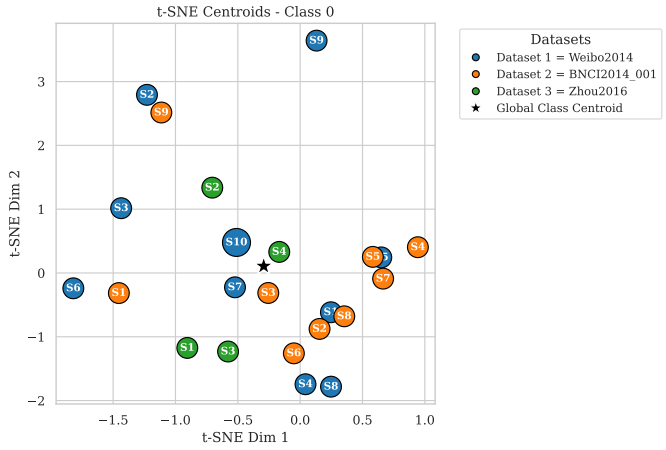

Fig. 4. Feet Class Centroid for Each Subject.

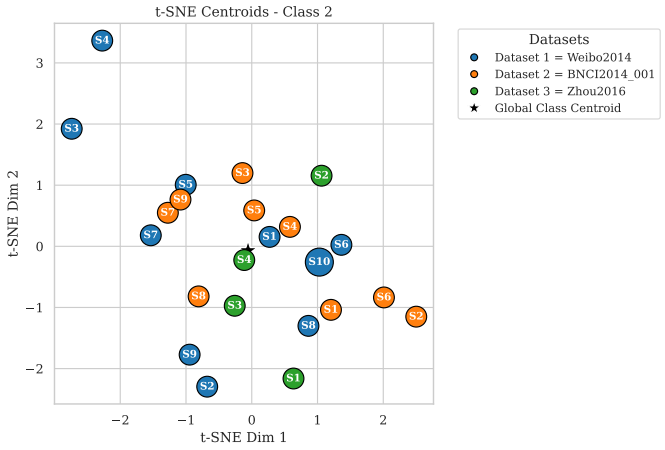

Fig. 6. Right Hand Class Centroid for Each Subject.

the classification region of the 'feet' class is not significantly influenced by the global centroid of the class features.

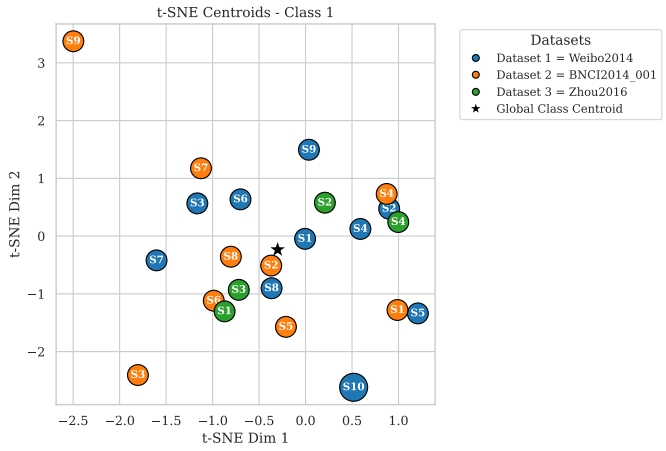

Fig. 5. Left Hand Class Centroid for Each Subject.

Similar to Figure 5, there are several subjects from each dataset whose feature centroids are quite far from the global centroid of the left_hand classification class. In the Weibo2014 dataset, subjects 3, 4, 5, 6, and 9 have centroids located considerably far from the global centroid of the left_hand classification class. In the BNCI2014_001 dataset, subjects 1, 4, 7, and 9 have centroids positioned relatively far from the global centroid of the left_hand classification class. In the Zhou2016 dataset, subject 4 has a centroid that is relatively far from the global centroid of the left_hand classification class. However, being far from the global centroid does not necessarily mean that these subjects are difficult to classify for the left_hand class. This can be confirmed by the F1-scores of the left_hand classification class for most of these subjects, which are relatively good as shown in Table III. Based on this analysis, it can be concluded that the classification region of the left_hand class is not significantly influenced by the global centroid of the classification features.

Likewise, in Figure 6, there are several subjects from each dataset whose feature centroids are relatively far from the global centroid of the right_hand classification class. In the Weibo2014 dataset, subjects 3, 4, 5, 6, and 9 have centroids located considerably far from the global centroid of the right_hand classification class. In the BNCI2014_001 dataset, subjects 2, 3, 7, and 9 also have centroids positioned relatively far from the global centroid of the right_hand classification class. In the Zhou2016 dataset, subject 1 and 2 has a centroid that is relatively far from the global centroid of the right_hand classification class. However, being far from the global centroid does not necessarily imply that these subjects are difficult to classify for the right_hand class. This can be verified by the relatively good F1-scores of the right_hand classification class for most of these subjects, as presented in Table III. Based on this analysis, it can be concluded that the classification region of the right_hand class is not significantly affected by the global centroid of the classification features.

G. Cross-Subject Classification

The training data used to train the proposed model contains trial subsets from all subjects included in the validation and test data. This causes the classification model built to be subject-dependent. Therefore, an evaluation in a cross-subject scenario is also conducted to test the model's performance in classifying unseen subjects.

In this evaluation scenario, the training data uses all trials from subjects 1 and 2 of the Zhou2016 dataset, subjects 1 to 4 of the BNCI2014_001 dataset, and subjects 1 to 5 of the Weibo2014 dataset. Meanwhile, the validation data uses subject 3 of Zhou2016, subjects 5 and 6 of BNCI2014_001, and subjects 6 and 7 of Weibo2014. Trials from the remaining subjects are used as test data. This configuration ensures that the evaluation scenario is fully cross-subject.

As shown in Table IV, the classification accuracy on each dataset drops significantly compared to the subject-dependent scenario. This indicates that the generalization techniques used are still insufficient to capture universal motor imagery char-

acteristics. One reason for the accuracy drop in this scenario is assumed to be closely related to the low signal-to-noise ratio (SNR) in EEG signals for motor imagery tasks. Additionally, it is assumed that the size of the training data used is still not large enough to capture more general signal characteristics.

TABLE IV
CROSS-SUBJECT ACCURACY

Method	Weibo2014	BNCI2014_001	Zhou2016
Cross-Subject	0.4972	0.5417	0.7747
Subject-Dependent	**0.6645**	**0.6512**	**0.8183**

VI. CONCLUSION

Based on the conducted research, the following conclusions can be drawn:

1) A CNN model architecture for a motor imagery classification system capable of handling multi-source datasets was successfully built. The sampling rates of the multi-source datasets used were adjusted to match the dataset with the lowest sampling rate to standardize input dimensions. Differences in the number of channels were efficiently handled using a spatial attention method that incorporates channel position information. Increased signal variability in multi-source datasets was managed using the EEG-DG method, which employs three new loss functions to help the model extract more general features.
2) The evaluation comparing the proposed model with the baseline model shows that the proposed model outperforms the baseline trained according to the literature as well as using multi-source datasets for several datasets. The comparison results are as follows).
   - An increase in classification accuracy of 0.1443 for the proposed model compared to the baseline trained without trial augmentation for the Weibo2014 dataset.
   - An increase in classification accuracy of 0.0041 for the proposed model compared to the baseline without trial augmentation for the BNCI2014_001 dataset.
   - A decrease in classification accuracy of 0.0250 for the proposed model compared to the baseline trained without trial augmentation for the Zhou2016 dataset.
   - An increase in classification accuracy of 0.0700 for the proposed model compared to the baseline trained using multi-source datasets for the Weibo2014 dataset.
   - An increase in classification accuracy of 0.0493 for the proposed model compared to the baseline trained using multi-source datasets for the BNCI2014_001 dataset.
   - A decrease in classification accuracy of 0.0267 for the proposed model compared to the baseline trained using multi-source datasets for the Zhou2016 dataset.

   Nevertheless, the proposed model yields significantly lower accuracy than the baseline when trained using a single dataset.
3) The evaluation results testing the effectiveness of using multi-source datasets show that this method is reasonably effective in improving the classification accuracy of the trained model. However, other trial-increasing methods proved more effective on the baseline model. A specialized model capable of efficiently processing multi-source datasets is required to further improve the effectiveness of this method. The proposed model developed in this study has proven capable of enhancing the effectiveness of this approach.

VII. FUTURE WORKS

Based on the discussion results, the following suggestions are proposed for future research:

- The trial-increasing method has significant potential to improve the classification accuracy of trained EEG signal models. Future research could investigate the performance of classification models when various data augmentation methods are applied to multi-source datasets used for training.
- The domain generalization component used in the proposed model exhibits linear complexity growth with the number of domains in the dataset. Therefore, further studies on redefining domains or exploring alternative generalization methods are recommended.
- GPU optimization for the training process was not applied to the spatial attention method to keep its implementation as close as possible to the literature used. Future research could explore optimizations for the spatial attention method.
- Research on the effect of multi-source datasets for EEG signal classification in other BCI paradigms can also be conducted.

REFERENCES

[1] J. W. Britton, L. C. Frey, J. L. Hopp, P. Korb, M. Z. Koubeissi, W. E. Lievens, E. M. Pestana-Knight, and E. K. St Louis, "Electroencephalography (eeg): An introductory text and atlas of normal and abnormal findings in adults, children, and infants," 2016.

[2] M. Lotze and U. Halsband, "Motor imagery," *Journal of Physiology-paris*, vol. 99, no. 4-6, pp. 386–395, 2006.

[3] M. Jeannerod, "Mental imagery in the motor context," *Neuropsychologia*, vol. 33, no. 11, pp. 1419–1432, 1995.

[4] X.-C. Zhong, Q. Wang, D. Liu, Z. Chen, J.-X. Liao, J. Sun, Y. Zhang, and F.-L. Fan, "Eeg-dg: A multi-source domain generalization framework for motor imagery eeg classification," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[5] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, and M. S. Hossain, "Deep learning for eeg motor imagery classification based on multi-layer cnns feature fusion," *Future Generation computer systems*, vol. 101, pp. 542–554, 2019.

[6] T.-j. Luo, C.-l. Zhou, and F. Chao, "Exploring spatial-frequency-sequential relationships for motor imagery classification with recurrent neural network," *BMC bioinformatics*, vol. 19, no. 1, p. 344, 2018.

[7] S. Omari, A. Omari, F. Abu-Dakka, and M. Abderrahim, "Eeg motor imagery classification: tangent space with gate-generated weight classifier," *Biomimetics*, vol. 9, no. 8, p. 459, 2024.

[8] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[9] P. Suwandjieff and G. R. Müller-Putz, "Eeg analyses of visual cue effects on executed movements," *Journal of Neuroscience Methods*, vol. 410, p. 110241, 2024.

[10] S. Chevallier, I. Carrara, B. Aristimunha, P. Guetschel, S. Sedlar, B. Lopes, S. Velut, S. Khazem, and T. Moreau, "The largest eeg-based bci reproducibility study for open science: the moabb benchmark," *arXiv preprint arXiv:2404.15319*, 2024.

[11] H. Altaheri, G. Muhammad, M. Alsulaiman, S. U. Amin, G. A. Altuwaijri, W. Abdul, M. A. Bencherif, and M. Faisal, "Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review," *Neural Computing and Applications*, vol. 35, no. 20, pp. 14 681–14 722, 2023.

[12] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.

[13] D. Truong, M. A. Khalid, and A. Delorme, "Deep learning applied to eeg data with different montages using spatial attention," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2023, pp. 2587–2593.

[14] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz *et al.*, "Review of the bci competition iv," *Frontiers in neuroscience*, vol. 6, p. 55, 2012.

[15] B. Zhou, X. Wu, Z. Lv, L. Zhang, and X. Guo, "A fully automated trial selection method for optimization of motor imagery based brain-computer interface," *PloS one*, vol. 11, no. 9, p. e0162657, 2016.

[16] W. Yi, S. Qiu, K. Wang, H. Qi, L. Zhang, P. Zhou, F. He, and D. Ming, "Evaluation of eeg oscillatory patterns and cognitive process during simple and compound limb motor imagery," *PloS one*, vol. 9, no. 12, p. e114853, 2014.

[17] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J.-R. King, "Decoding speech perception from non-invasive brain recordings," *Nature Machine Intelligence*, vol. 5, no. 10, pp. 1097–1107, 2023.

[18] V. S. Kardam, S. Taran, and A. Pandey, "Motor imagery tasks based electroencephalogram signals classification using data-driven features," *Neuroscience Informatics*, vol. 3, no. 2, p. 100128, 2023.

[19] O. George, R. Smith, P. Madiraju, N. Yahyasoltani, and S. I. Ahamed, "Data augmentation strategies for eeg-based motor imagery decoding," *Heliyon*, vol. 8, no. 8, 2022.

[20] A. Melnik, P. Legkov, K. Izdebski, S. M. Kärcher, W. D. Hairston, D. P. Ferris, and P. König, "Systems, subjects, sessions: to what extent do these factors influence eeg data?" *Frontiers in human neuroscience*, vol. 11, p. 150, 2017.