Highlights

EEG Data Augmentation Method Based on the Gaussian Mixture Model

Chuncheng Liao, Shiyu Zhao, Xiangcun Wang, Jiacai Zhang, Yongzhong Liao, Xia Wu

- Proposing an EEG data augmentation method based on GMM microstate feature reconstruction. Firstly, GMM clustering is performed on same-type data samples to obtain microstate features of each sample type, using Gaussian submodel probabilities. Secondly, random selection of two same-type samples analyzes the similarity of principal components and exchanges principal components with similar features to form new submodel probabilities. Finally, new data is generated based on submodel probabilities, weights, means, and variances.

- By analyzing the newly added EEG samples and original EEG samples from traditional EEG data augmentation methods in terms of time and space features, we demonstrate the differences in spatiotemporal features among the methods.

- Classification tasks are performed on EEG generated by traditional methods and the method proposed in this paper to compare improvements in classification performance.

EEG Data Augmentation Method Based on the Gaussian Mixture Model★

Chuncheng Liao[a,1], Shiyu Zhao[b,1], Xiangcun Wang[a], Jiacai Zhang[a,*], Yongzhong Liao[c] and Xia Wu[d]

[a] *School of Artificial Intelligence, Beijing Normal University, Xinjiekouwai Street 19, Haidian District, 100875, Beijing, China*

[b] *Tianyi Security Technology Co., Ltd., No.88 Yunlongshan Road,Shazhou Street, 210000 Jianye District, Nanjing, China*

[c] *School of Mechanical and Electrical Engineering,Changsha Institute of Technology, NO.366 Wangwang West Road, Gaotangling Street, Wangcheng District, 410200, Changsha, China*

[d] *School of Computer Science And Technology,Beijing Institute of Technology, Zhongguancun South Street 5, Haidian District, 100081, Beijing, China*

ARTICLE INFO

Keywords:
EEG
Gaussian Mixture Model
Microstate
Feature
Data Augmentation

ABSTRACT

Traditional methods of electroencephalograms(EEG) data augmentation, such as segmentation-reassembly and noise mixing, suffer from data distortion that can alter the original temporal and spatial feature distributions of the brain signals. Deep learning-based methods for generating augmentation EEG data, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have shown promising performance but require a large number of comparative learning samples for model training. To address these issues, this paper introduces an EEG data augmentation method based on Gaussian Mixture Model microstates, which retains the spatiotemporal dynamic features of the EEG signals in the generated data. The method first performs Gaussian mixture clustering on data samples of the same class, using the product of the probability coefficients and weight matrices of each Gaussian model as corresponding microstate features. Next, it randomly selects two EEG data samples of the same type, analyzes the similarity of the main components of the microstate features, and swaps the similar main components to form new Gaussian mixture model features. Finally, new data is generated according to the Gaussian mixture model using the respective model probabilities, weights, means, and variances. Experimental results on publicly available datasets demonstrate that the proposed method effectively characterizes the original data's spatiotemporal and microstate features, improving the accuracy of subject task classification.

1. Introduction

Electroencephalograms (EEG) are biological signals generated by brain neural activity, reflecting the physiological state of the brain. EEGs record the spontaneous and rhythmic electrical activity of groups of brain cells through electrodes. Due to their safety, portability, ease of use, high temporal resolution, and low cost, EEGs are widely used in medical and cognitive neuroscience research fields [1].

However, commonly used scalp EEG signals exhibit characteristics such as nonlinearity, non-stationarity, broad frequency bands, and a low signal-to-noise ratio (SNR)[35]. These traits pose significant challenges for research and application. Firstly, the potential signals from within the skull are conducted through layers of tissues, fluids, bones, etc., to reach the scalp, where they are easily interfered with by physiological and non-physiological signals, reducing SNR [5]. Secondly, due to the high level of attention required during data collection, it is difficult to obtain large amounts of brain data, leading to small sample characteristics[16]. Thirdly, the signal data changes over time and is non-stationary and nonlinear [11, 20], resulting in poor generalization performance of computational models for EEG processing. Fourthly, there are significant individual differences in EEG signals, further limiting the generalization performance of EEG processing models.EEG models trained on specific individuals may perform poorly on new individuals [12, 30]. Additionally, a large amount of redundant information or non-specific information unrelated to the target interferes with the key physiological information in the EEG signals. Therefore, it is necessary to increase samples, denoise, and transform EEG data to generate new datasets with better diversity and robustness, thereby improving model generalization, reducing overfitting, and enhancing model analysis accuracy.

Traditional EEG data augmentation methods mainly include the following:

Firstly, methods based on data morphology changes[37, 19]. These methods simulate the effects of factors such as head movement and muscle tension on EEG data, generating data with better representation of temporal and frequency variations based on the original EEG data . However, these geometrically transformed data may destroy the time domain and frequency features[21, 41].

Secondly, methods based on signal segmentation and recombination[26, 32, 18, 44]. This method segments specific time window EEG data according to the temporal characteristics of the EEG signal and reconstructs new data by randomly selecting fragments. Assuming $D = x_i$is an EEG signal set with a specific category, $x_i \in \mathcal{N}$(i=1âĂęN, $|x_i| = N$ indicates the total number of samples), where N represents the number of sampling points in each sample.

[★] This work is funded by the Special Fund for Research on National Major Research Instruments (62227801) of the Nature Science Foundation of China.

[*] Corresponding author

✉ jiacai.zhang@bnu.edu.cn (J. Zhang)

ORCID(s): 0009-0003-1456-6603 (C. Liao)

[1] Authors contribute equally to this work.

Each EEG waveform is divided into k non-overlapping consecutive segments, forming a dataset D∨`& containing $N \times K$ data fragments. New EEG samples $x$∨`& are then generated by randomly concatenating K segments from D∨`&, repeating the operation until the desired number of signals is obtained.These methods are intuitive and simple but may exacerbate model overfitting due to the similarity after augmentation[39, 34, 8, 36]. Another common method is adding noise. This method adds random matrices from Gaussian distributions (Gaussian or salt-and-pepper noise) to the original EEG data to simulate real-world noise interference for data augmentation. This method can effectively increase dataset diversity and improve model robustness [7, 42, 29]. However, introducing artifacts in the original signal makes it difficult to verify the true psychological state response of the new EEG signal, complicating model accuracy validation.

Thirdly, methods based on data transfer. This method transfers existing same-type EEG data to new environments to improve model performance. It usually uses auxiliary data from the source domain to support training in similar domains. This method may lead to data distortion and affect model accuracy [31].

Fourthly, methods based on data generation. There are two main types here: Variational Autoencoders (VAE), consisting of an encoder and a decoder, where the encoder converts original data into latent data, and the decoder maps latent data back to real data. To generate new data, VAE randomly samples points from the learned latent space and passes these samples through the decoder network to reconstruct them as new samples. The drawback of this method is the need for a large number of data samples [23, 10]. Generative Adversarial Networks (GAN) and their variants can generate synthetic data through training generative and discriminative networks[3, 13, 28]. The generative network accepts random noise from a specific distribution (e.g., Gaussian) and attempts to generate realistic-like synthetic data, while the discriminative network is trained to classify between real and synthetic data. These two networks are adversarial; after sufficient training, the generative network produces similar signals. [14, 27, 9] The common drawback of these methods is that they require a certain amount of data to support the adversarial training of generators and discriminators, which conflicts with the goal of data augmentation on small training sets, and also consumes substantial computational resources and presents replication difficulties[33, 17, 4].

Existing methods for increasing EEG data samples either require a large number of comparison learning samples or generate new data that alters the original EEG's spatiotemporal features, leading to data distortion. Researching a new data augmentation method that does not require extensive original data samples while preserving the spatiotemporal dynamic features of similar EEG signals is significant for improving EEG processing algorithms.

Microstate analysis is one of the EEG signal analysis methods, capable of characterizing quasi-steady-state scalp potential fields on a sub-second scale and retaining the temporal dynamics and spatial information of scalp potential distributions, representing a novel form of EEG signal quantification with potential neurophysiological relevance[40, 25]. Our paper reconstructs new EEG data based on Gaussian microstate features. Firstly, Gaussian Mixture Models (GMM) decompose same-type sample EEG signals to obtain microstate feature parameters for each sample, using the probability of Gaussian submodels. Secondly, random selection of two same-type samples analyzes the similarity of principal components, and exchanges and reassembles principal components with similar features to form new submodel probabilities. Finally, new data is generated based on submodel probabilities, weight values, means, and variances.

(1) Proposing an EEG data augmentation method based on GMM microstate feature reconstruction. Firstly, GMM clustering is performed on same-type data samples to obtain microstate features of each sample type, using Gaussian submodel probabilities. Secondly, random selection of two same-type samples analyzes the similarity of principal components and exchanges principal components with similar features to form new submodel probabilities. Finally, new data is generated based on submodel probabilities, weights, means, and variances.

(2) By analyzing the newly added EEG samples and original EEG samples from traditional EEG data augmentation methods in terms of time and space features, we demonstrate the differences in spatiotemporal features among the methods.

(3) Classification tasks are performed on EEG generated by traditional methods and the method proposed in this paper to compare improvements in classification performance.

The rest of this paper is organized as follows. Section 2 mainly introduces our proposed method and experimental data, including the implementation of Gaussian microstate feature reconstruction. Section 3 primarily discusses experiments and results. Section 4 discusses and analyzes experimental results. Finally, Section 5 summarizes this work.

2. Materials And Methods

2.1. The BCI Competition IV Dataset 2a.

This dataset was collected from 9 subjects[6]. Each subject completed four different motor imagery asks: left hand movement (category 1), right hand movement (category 2), feet movement (category 3) and tongue movement (category 4). Each subject underwent a total of 288 trials, divided into 6 sessions. There was a short break midway through each session. Each session comprised 48 trials, with 12 trials per task category.

Data preprocessing: The EEG dataset is pre-processed using Python 3.8.8 and the MNE 0.23.4 toolkit.

Baseline Correction: A common approach was used to correct the baseline. For all datasets, including the training and testing sets, EEG trials with baseline noise were rejected by visual inspection. Then, a high-pass filter was applied to

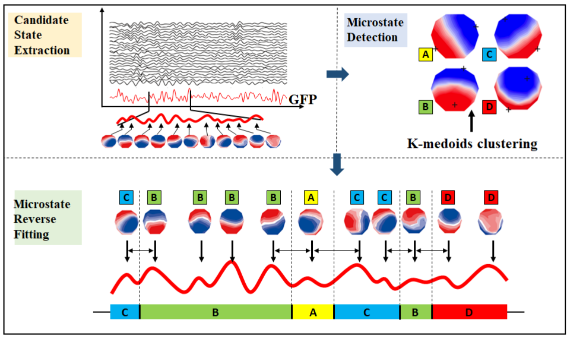


Figure 1: Principle diagram of EEG microstate analysis.

both the training and testing sets; that is, the filters for the training and testing datasets shared the same parameters.

Principal Component Analysis (PCA): To reduce the dimensionality of our data and extract uncorrelated features, the projection space of PCA was learned from the training set. We then projected the testing samples into the same space learned from the training set.

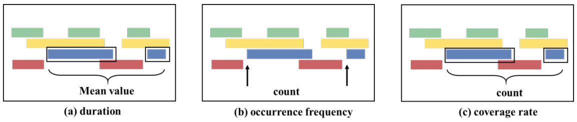
Artifact Rejection: Firstly, multi-channel EEG signals were decomposed into independent components using Independent Component Analysis (ICA). Secondly, correlation coefficients between the time courses of each ICA component and three electrooculogram channels were calculated. Then, threshold processing based on an adaptive Z-score was performed. Components above the threshold (threshold = 3.0) were masked, and the Z-score was recalculated until no components exceeded the threshold. Lastly, the scalp EEGs were reconstructed from the remaining independent components, thereby automatically excluding the effect of the electrooculogram signal. The algorithm was first trained on the training set using a set of known artifacts and then applied to the testing set.

2.2. EEG Microstates and Gaussian Microstate

2.2.1. EEG Microstates

Originating from Lehmann et al. in 1987, multi-channel EEG recordings can be composed of a series of quasi-stable microstates, each characterized by the topological structure of the entire channel terrain map[40, 25]. This representation is an effective method that helps us understand the dynamic processes of the brain during information processing, as well as functional changes under different tasks and states. It holds significant importance for fields such as cognitive neuroscience, clinical diagnosis, and rehabilitation therapy. As shown in Figure 1, EEG microstate analysis includes the following steps:

Microstate Feature Extraction: Key features such as duration, frequency of occurrence, and spatial distribution are extracted from each microstate. These features can reflect functional changes in the brain under different tasks and states.



(a) duration (b) occurrence frequency (c) coverage rate

Figure 2: Feature extraction of microstate hybrid model based on GMM.

Microstate Detection: EEG signals are divided into different microstates through specific algorithms (such as clustering analysis). Each microstate represents a transient functional state of the brain, usually lasting from tens to hundreds of milliseconds.

Statistical Analysis: Extracted microstate features are statistically analyzed to reveal patterns and differences in brain activity under different conditions. For example, comparing microstate features under different cognitive tasks or pathological states can help study abnormalities and plasticity of brain function.

Result Interpretation: Based on the results of statistical analysis, combined with knowledge from neuroscience and psychology, changes in brain activity are interpreted and inferred. This aids in our better understanding of the brain's functions and structure, as well as its role in cognitive and behavioral processes.

2.2.2. Gaussian Microstate

Gaussian Mixture Model-Based EEG Microstates, a distinction from Traditional EEG Microstates Unlike traditional EEG microstates, Gaussian Mixture Model (GMM)-based EEG microstates are distinguished by their use of GMM to decompose EEG microstates into a mixture representation rather than a unique one-hot representation. It has been proven that the classification capability of the GMM-based microstate model under MI tasks is Augmentation [24].

Taking a ten-component Gaussian model(GMM) as an example, the preprocessed EEG data $X \in R^{N\times T}$ serves as the input to the model, where N represents the number of sensors on the EEG device and T is the number of sampling points per sample. $Z \in R^{N\times T}$ represents the output of the model, with K indicating the number of submodels in the GMM, consistent with the number of microstates in the k-medoids set(K=10). The model algorithm is shown in Table 1.

Ultimately, the multi-channel EEG data is decomposed into a linear combination of multiple Gaussian submodels. Similar to the one-hot representation model for MI EEG, the GMM mixture representation model extracts dynamic statistical features based on the probability of each submodel. These features include three characteristics as shown in Figure 2: duration, frequency of occurrence, and coverage.

2.3. Gaussian Mixture Model-Based EEG Data Augmentation Method

This paper decomposes EEG microstates into a mixture representation, meaning that each sampling point of the multi-channel raw EEG data sample can be decomposed into

Table 1
Algorithm Process for EEG Data Augmentation Method Based on Gaussian Mixture Model

```
Algorithm:EEG Data Augmentation Method Based on Gaussian Mixture Model
Input:original_data = (trial,channel,n_samples), original_labels =(trial,)
Output:gmm_data
Step1:Set model parameters.n_components = 10; random_stata = 42;$V_T$ = 0.8
Step2:Cluster the data with the same label and calculate the microstate features of each sample belonging to each cluster.
probs = gmm.predict_proba(original_data)
means = gmm.means_
covariances = gmm. covariances_
weights =gmm.weights_robs
Step3:Sampling.
for{$i \leq trial$
    for{$j \leq channel$
        for{$k \leq n\_components$
            $fitted_value = np.random.(mean = means[i],$
            $cov = np.diag(covariances[i]))\}\}\}$
Step4:Calculate the product matrix.
weighted_probs_values = gmm.weights_ * probs
Step5:Swap similar points.
weighted_probs_values
= swap_columns(weighted_probs_values, $V_T$)
Step6:Fit the data.
data_generate_sampel =
np.matmul(weighted_probs_values ,fitted_values)
Step7:Swap channels and reconstruct the data.
gmm_data =
GMM_FEATURE(probability=probability,random_state=42)
```

a linear combination of multiple Gaussian distributions. The duration and frequency or coverage of each Gaussian submodel at each sampling point form the microstate features, encompassing spatiotemporal dynamic information. Then, points with the same feature across different samples are randomly exchanged while maintaining the overall characteristics of the samples. This is followed by inverse reconstruction to synthesize new EEG data. The overall principle is shown in Figure 3,and the algorithm is presented in Table I. This method increases data diversity and complexity by combining the original EEG signals with noise or distortion signals generated by GMM (Gaussian Mixture Model). As illustrated by the aug_gmm waveform curve in Figure 6 and the brain topographic maps in Figure 10, the signal exhibits more uncertain variations. This approach helps the model learn more robust feature representations, enhancing its ability to recognize different physiological states or cognitive processes.

2.3.1. Gaussian Microstate Decomposition

As shown in Figure 3(a), during the decomposition process, we first preprocess the original EEG data, then apply Gaussian Mixture Model clustering to calculate the product matrix of weights and probabilities for adjacent data points with the same label. Points with a similarity coefficient greater than a threshold are exchanged to obtain a new microstate feature matrix. The steps are as follows:

Step 1: Preprocess the original data with filtering, denoising, and other operations.

Step 2: Conduct Gaussian clustering on the data based on labels to obtain the probability, weight, mean, and variance for each category.

Step 3: Multiply the weight of each sample by the probability matrix to obtain the microstate feature matrix.

Step 4: Calculate the similarity coefficients of the microstate feature matrix points for adjacent data with the same label. For points with a similarity coefficient greater than 0.8, exchange their positions to generate a new microstate feature matrix. The purpose is to reduce overfitting in the generated data.

2.3.2. Gaussian Microstate Reconstruction

As shown in Figure 3(b), during the reconstruction process, we set a random seed to ensure the reproducibility of the results. For each sample, a channel is randomly selected, and the original data and fitted data are exchanged for that channel, ultimately generating new EEG data. The steps are as follows:

Step 1: Combine the microstate feature matrix with the mean and variance to refit the EEG data.

Step 2: Set a random seed to ensure the reproducibility of the results.

Step 3: Randomly select a channel for each sample according to a certain probability.

Step 4: Exchange the original data and fitted data for the corresponding channel.

Step 5: Reconstruct to obtain new EEG data.

2.4. Other Augmentation Methods

To evaluate the effectiveness of the EEG data augmentation method proposed in this paper, we employed nine different data augmentation techniques. These techniques cover time domain, frequency domain, spatial domain, or other comprehensive transformations.

2.4.1. Time-Domain Augmentation Methods

Noise Addition: By superimposing noise signals onto the original signal, we simulate interference and distortion in real-world environments[42, 29]. This accounts for variations in different environments, device differences, or signal transmission disturbances. As shown by the aug_noise curve in Figure 4, the waveform changes after adding noise. The brain topographic maps in Figure 7 reflect how the intensity and distribution of signals in different regions are affected by noise to varying degrees, thereby increasing data complexity and diversity. This helps the model learn more robust feature representations, enhancing its ability to recognize different physiological states or cognitive processes.

Assuming the original signal is $x(t)$, where t represents time, and ϵ represents noise, which could be Gaussian white noise, uniform white noise, salt-and-pepper noise, etc. The

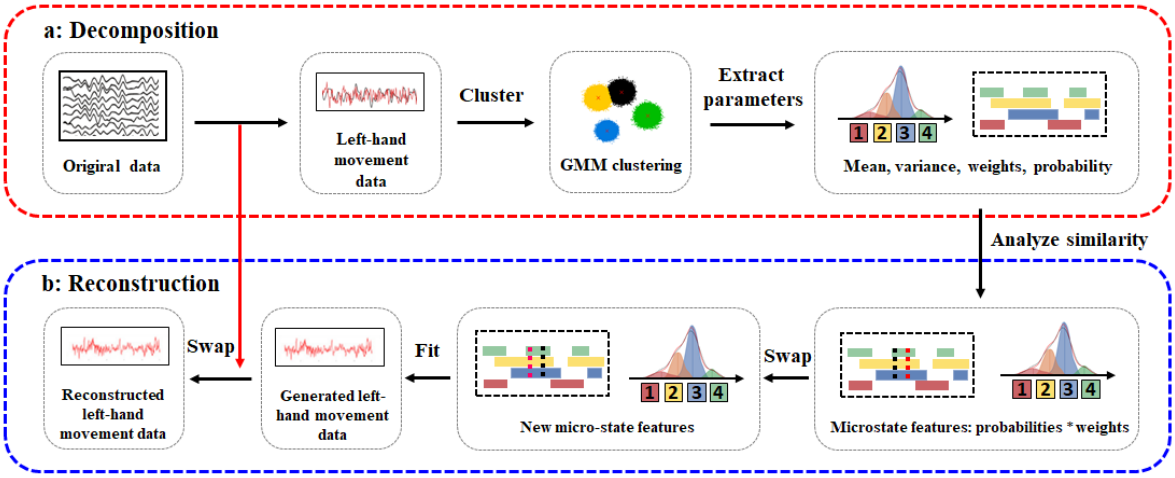


Figure 3: Principle diagram of the EEG data augmentation method based on Gaussian mixture models.Data preprocessing mainly includes filtering(0-38HZ), segmentation and baseline correction, removing artifacts, and referencing. (a) Decomposition.First, preprocess the original EEG data, then use the Gaussian Mixture Model (GMM) for clustering. Calculate the product matrix of weights and probabilities for adjacent data points with the same label. Swap points with similarity coefficients greater than a threshold to obtain a new microstate feature matrix. (b) Reconstruction.First, set a random seed to ensure the reproducibility of results. For each sample, randomly select a channel, swap channels between the original data and the fitted data, and finally generate new EEG data.

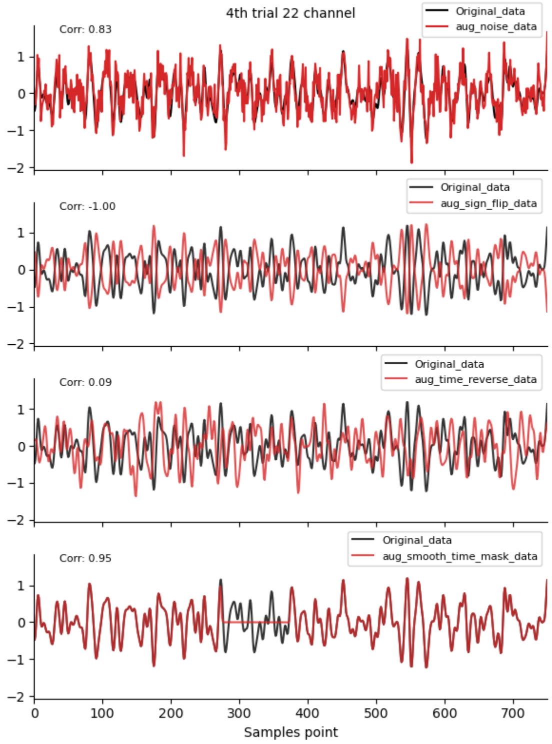


Figure 4: Time-domain augmentation method waveform diagram.Data source: Subject 1, Trial 4, Left-hand Motor Imagery, Channel 22.

augmented signal can be expressed as:

$$y(t) = x(t) + \epsilon \tag{1}$$

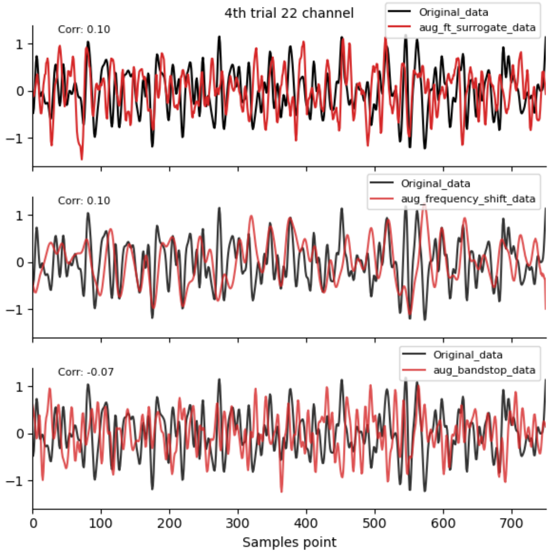


Figure 5: Frequency-Domain augmentation method waveform diagram.Data source: Subject 1, Trial 4, Left-hand Motor Imagery, Channel 22.

Time Reverse: By reversing the signal in time, we increase data diversity[29]. This is demonstrated by the aug_time _reverse curve in Figure 4. The mathematical representation is relatively straightforward, primarily involving reversing the time sequence of the signal.

Assuming the original signal is $x(t)$, where t represents time, the time-reversed signal can be expressed as:

$$y(t) = x(T - t) \tag{2}$$

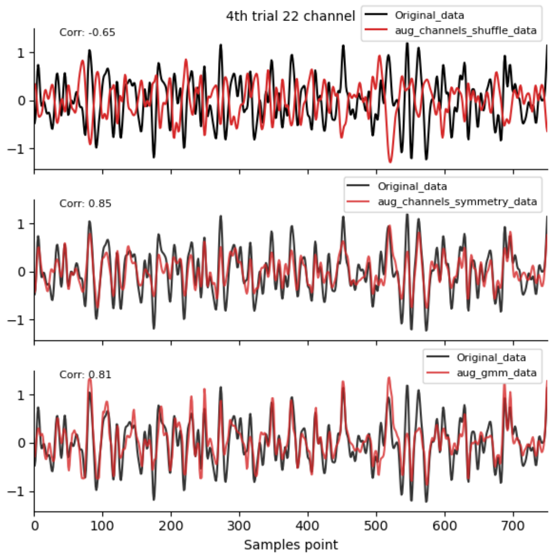


Figure 6: Spatial-Domain and GMM augmentation methods waveform diagram.Data source: Subject 1, Trial 4, Left-hand Motor Imagery, Channel 22.

Time Masking: By randomly selecting a continuous segment on the time axis and setting its values to zero, we simulate noise or distortion[29].As shown by the aug_smooth _time_mask curve in Figure 4, the signal is completely set to zero within a certain time interval. This treatment may appear as a sudden drop or disappearance of signal intensity in specific regions during the corresponding time intervals in the brain topographic maps in Figure 5, thereby simulating signal interference or loss scenarios.

Assuming the original signal is $x(t)$, where t represents time, the mathematical representation of the symmetric transformation is:

$$y(t) = x(t) \times (1 - M(t) \tag{3}$$

where $M(t)$ is the masking function, taking values of 0 or 1.

Sign Flip: By flipping the sign of the signal, we convert positive signals to negative and vice versa[29]. As shown by the aug_sign_flip curve in Figure 4, the waveform undergoes a reversal between positive and negative values. In the brain topographic maps in Figure 7, this change manifests as an opposite distribution of positive and negative signals in certain regions compared to the original signal, increasing data diversity.

Assuming the original signal is $x(t)$, where t represents time, the mathematical representation of the sign flip transformation is:

$$y(t) = -x(t) \tag{4}$$

2.4.2. *Frequency-Domain Augmentation Methods*

Frequency Shift: By shifting the frequency of the signal, we alter its frequency components[29]. As shown by the aug_frequency_shift curve in Figure 5, the frequency components of the waveform undergo a shift. In the brain topographic maps in Figure 8, this may reflect changes in the spatial distribution of different frequency components, indicating that the dominant frequencies in certain regions differ from the original signal, thereby simulating different physiological states or cognitive processes.

Assuming the original signal is $x(t)$ and Δ is the amount of frequency shift, the mathematical representation of the frequency shift transformation is:

$$y(t) = x(t) \cdot e^{j2\pi\Delta ft} \tag{5}$$

Fourier Transform Surrogate: By computing the Fourier transform of the EEG signal and generating a surrogate signal, we alter the frequency-domain characteristics of the signal[39]. As shown by the aug_ft_surrogate curve in Figure 5, the waveform undergoes changes in its frequency components. In the brain topographic maps in Figure 8, this may reflect variations in the spatial distribution of different frequency components, aiding the model in learning richer frequency-domain information.

Assuming $X(t)$ is the original signal and $Y(\omega)$ is the frequency-domain signal after the Fourier transform, the mathematical representation of the Fourier transform is:

$$Y(\omega) = F[X(t)] \tag{6}$$

Bandstop Filter: A signal processing technique used to block signals within a specific frequency range while allowing other frequencies to pass through[29]. It is useful in many applications, such as removing power line noise from electrocardiogram (ECG) signals or eliminating interference in specific frequency bands in electroencephalogram (EEG) signals. This is demonstrated by the aug_bandstop_filter curve in Figure 5. The transfer function of a bandstop filter typically consists of two low-pass filters and one high-pass filter.

Assuming we have a bandstop filter with a center frequency of fc and a bandwidth of B, its transfer function can be expressed as:

$$H(f) = HLP(f) - HHP(f) \tag{7}$$

where: $HLP(f)$ is the transfer function of the low-pass filter with a cutoff frequency of $2f_c - 2B$. $HHP(f)$ is the transfer function of the high-pass filter with a cutoff frequency of $2f_c + 2B$.

2.4.3. *Spatial-Domain Augmentation Methods*

Channel Symmetry: This method involves applying a form of symmetric transformation to signals recorded from different electrodes[8]. As shown by the aug_channels_symmetry curve in Figure 6, this may result in waveforms exhibiting some degree of symmetry between different electrodes. In the brain topographic maps in Figure 9, this change may cause the signal distribution between different regions to exhibit symmetric patterns, aiding the model in learning the interrelationships and independence among various channels.

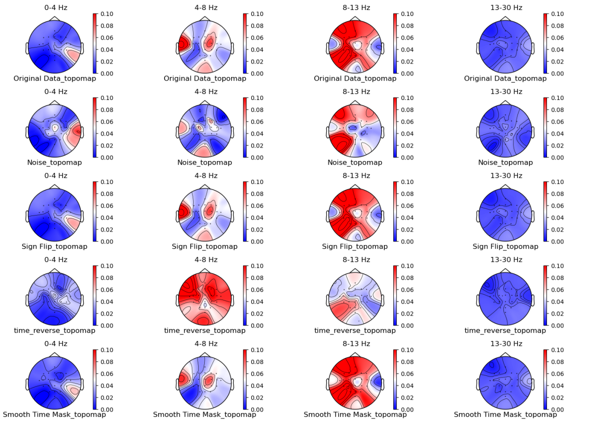


Figure 7: Time-domain augmentation method for brain topographic maps using left-hand motor imagery.Data source: Subject 1, Trial 4.

Assuming the original signal is $x(t)_i$, where t represents time and i indicates the electrode number, the mathematical representation of the symmetric transformation is:

$$y(t)_i = x(t)_{n+1-i} \tag{8}$$

where n is the total number of electrodes.

Channel Shuffle: By randomly shuffling the order of signals recorded from different electrodes, we alter the correspondence between the electrodes[36]. As shown by the aug_channels_shuffle curve in Figure 6, this results in changes to the correspondence between waveforms across different electrodes. In the brain topographic maps in Figure 9, this change may cause the signal distribution in different regions to become chaotic or disordered, increasing the complexity and diversity of the data.

Assuming the original signal is $x(t)_i$, where t represents time and i indicates the electrode number, the mathematical representation of the shuffling transformation is:

$$y(t)\delta(i) = x(t)_i \tag{9}$$

where δ is a random permutation function.

3. EXPERIMENT AND RESULTS

3.1. Comparison of Data Characteristics Generated by Different Augmentation Methods

Figure 4 shows the time-domain waveforms of the first participant, the first trial, imagined left-hand movement, and channel 22 for data generated by ten EEG data augmentation methods. Figure 10 displays the brain topographic maps of the first subject, the fourth trial, imagined left-hand movement, and frequency bands 0-4 Hz (Delta), 4-8 Hz (Theta), 8-13 Hz (Alpha), and 13-30 Hz (Beta) for data generated by the same ten EEG data augmentation methods.

From the original data brain topographic maps, it is evident that different frequency bands exhibit significant variations in activity across various brain regions. Specifically, the α wave band (8-13 Hz) shows the most pronounced activity in the frontal and occipital lobes. The activation of these areas is closely related to the planning, simulation, and execution of imagined left-hand movements. In contrast, other frequency bands such as δ waves, θ waves, and β waves show relatively weaker activity during imagined left-hand movement tasks and have less distinct associations with specific brain regions compared to the α wave band.

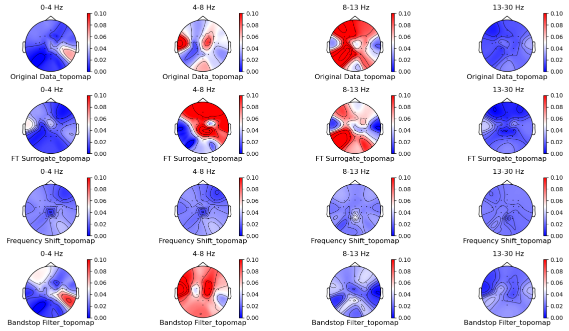


Figure 8: Frequency-Domain augmentation method for brain topographic maps using left-hand motor imagery.Data source: Subject 1, Trial 4.

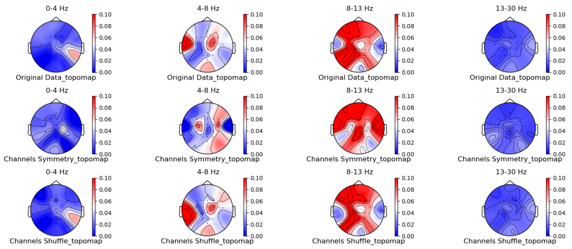


Figure 9: Spatial-Domain augmentation method for brain topographic maps using left-hand motor imagery.Data source: Subject 1, Trial 4.

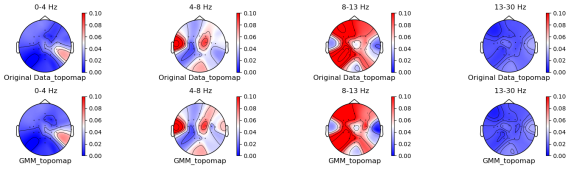


Figure 10: GMM augmentation method for brain topographic maps using left-hand motor imagery.Data source: Subject 1, Trial 4, Left-hand Motor Imagery.

Looking at the brain topographic maps from each augmentation method, methods such as signal flipping, time masking, channel shuffling, and GMM generally maintain characteristics similar to those of the original data brain topographic maps. Other augmentation methods exhibit larger deviations.

Of course, the above description is based on a general understanding. Due to significant individual differences, actual research results often vary greatly depending on factors such

Table 2
Comparison of average classification accuracy in K-fold cross-validation between data generated by different augmentation methods and the original data accuracy mean on dataset $2a$.

Method	FBCSP[2]	LSTM[15]	EEGNet[22]	ShallowNet[43]	Deep4Net[38]	Avg	SD
Original data	67.75	48.17	46.07	48.91	52.89	52.76	8.74
Noise Addition	73.22	80.72	75.29	80.32	83.08	78.53	4.10
Sign Flip	74.63	80.72	74.50	81.84	82.75	78.89	4.01
Time reverse	78.27	79.32	76.39	79.73	79.90	78.72	1.45
Time Masking	75.46	79.88	75.52	80.17	80.92	78.39	2.67
FT Surrogate	81.26	77.60	73.34	80.13	82.19	78.90	3.55
Frequency Shift	81.18	74.40	68.47	72.79	76.83	74.73	4.72
Bandstop fliter	76.32	78.39	76.98	78.16	81.13	78.20	1.85
Channel Symmetry	77.87	75.86	73.93	79.47	81.61	77.75	3.00
Channel Shuffle	78.59	76.51	68.29	75.06	79.86	75.66	4.52
GMM Augmentation	79.67	80.53	77.70	82.61	82.73	80.64	2.11

as specific task requirements, personal physiological characteristics, psychological states, and the analytical methods used for data analysis.

3.2. Comparison of the Effectiveness of Data Generated by Different Augmentation Methods on Classification Models

To accurately assess the performance of each data augmentation method on various models, we will employ a 5-fold cross-validation approach. Specifically, for each participant, the dataset is divided into five subsets. Each time, one subset is used as the test set and the remaining subsets are used as the training set. This process is repeated k times, and the average classification accuracy for each participant under the model is calculated. Finally, the average classification accuracy across nine participants is computed as the performance metric.

For each model (FBCSP, LSTM, EEGNet, ShallowNet, Deep4Net), we record the average classification accuracy of each data augmentation method across nine participants. As shown in Table 2.

Through the analysis of the average classification accuracies of different augmentation methods across five models, we obtained the following results:

GMM Augmentation: It achieved the highest average accuracy of 80.64% with a standard deviation of 2.11, indicating that among all augmentation methods, GMM Augmentation provided the best performance improvement. It delivered the highest average classification accuracy on all models, particularly excelling on Deep4Net.

FT Surrogate: The average accuracy was 78.90% with a standard deviation of 3.55, making it the second-best enhancement method, closely approaching the performance of GMM Augmentation. This method significantly improved the accuracy on FBCSP and Deep4Net but had minimal impact on EEGNet.

Time Reverse: The average accuracy was 78.72% with a standard deviation of 1.45. This method performed well across all models and also showed relatively good stability in terms of data consistency, although it had a narrow range and limited variation.

Sign Flip: The average accuracy was 78.89% with a standard deviation of 4.01, similar to Channel Symmetry, and demonstrated good consistency. While it performed well on ShallowNet and Deep4Net, its improvements on FBCSP and EEGNet were minimal.

Noise Addition: The average accuracy was 78.53% with a standard deviation of 4.10, performing well overall and significantly enhancing the accuracy on FBCSP and EEGNet, but having almost no effect on ShallowNet while showing significant improvement on Deep4Net.

Time Masking: The average accuracy was 78.39% with a standard deviation of 2.67. Although ranked middling overall, it improved performance on all models, especially on ShallowNet and Deep4Net.

Bandstop Filter: The average accuracy was 78.20% with a standard deviation of 1.85. This method significantly Augmentation the accuracy on Deep4Net and maintained relatively stable performance.

Channel Symmetry: The average accuracy was 77.75% with a standard deviation of 3.00, indicating consistent and stable performance across all models.

Channel Shuffle: The average accuracy was 75.66% with a standard deviation of 4.52, showing the most variability among all methods. Its performance varied across models, providing some improvement on FBCSP and Deep4Net.

Frequency Shift: The average accuracy was 74.73% with a standard deviation of 4.72, similar to Noise Addition, performing well but with larger fluctuations. It performed best on FBCSP but offered limited improvements on other models.

Original: As a baseline, the average accuracy was 52.76% with a standard deviation of 8.74, the lowest among all methods.

From these analyses, it is evident that GMM Augmentation and FT Surrogate are the two most effective data augmentation methods, significantly improving the model's classification accuracy. Time Reverse and Bandstop Filter also provided relatively stable performance improvements.

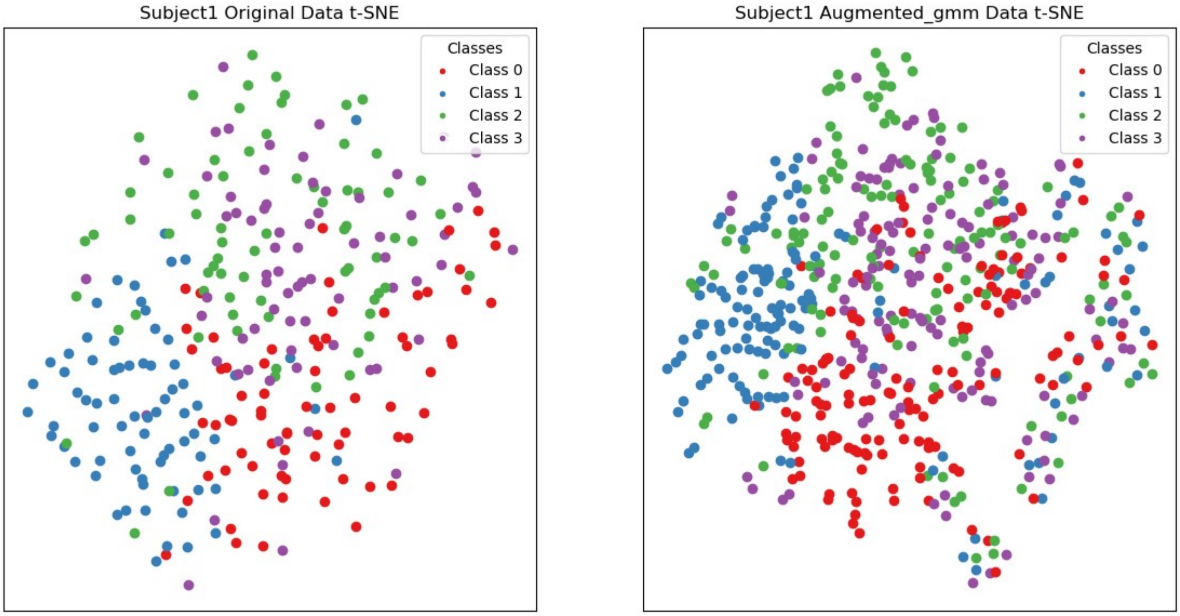


Figure 11: t-SNE visualization of Subject 1's original data and the data augmented using the Gaussian Mixture Model (GMM).Red represents the data for left-hand motor imagery, blue represents the data for right-hand motor imagery, green represents the data for both feet motor imagery, and purple represents the data for tongue motor imagery.

In contrast, the original data (without augmentation) performed the worst, indicating that employing appropriate data augmentation techniques can significantly enhance the classification performance of models.

3.3. Comparison of Visualization Results between Original Data and the data augmented using the Gaussian Mixture Model (GMM)

As shown in Figure 11, by comparing the t-SNE visualization of the original data and the data generated by augmentation methods for the first participant, we found that:

3.3.1. *Clarity of Clusters*

Original Data: In the t-SNE visualization of the original data, points from different classes do not have clear boundaries, leading to insufficient distinction between classes.

Augmentation Data: The data processed by the augmentation method described in this paper shows clearer cluster structures in t-SNE visualization. Points from each category are relatively clustered together, forming distinct clusters with certain boundaries between them. This indicates that the augmentation method effectively improves the distinguishability between different categories.

3.3.2. *Local Structure Compactness*

Original Data: In the t-SNE visualization of the original data, points from the same class are more dispersed and do not form tight clusters.

Augmentation Data: The data processed by the augmentation method described in this paper exhibits higher local structural compactness in t-SNE visualization. Points from the same category are closely clustered together, forming more compact clusters. This indicates that the augmentation method effectively improves the intra-class consistency.

In summary, the augmentation methods in this paper improve the visualization of the data, making points from different classes more clearly clustered together, and increasing the separation between different classes. This improvement helps enhance the performance of subsequent machine learning models, as clearer class boundaries and higher local structure compactness aid the model in learning and classifying data more accurately.

4. Discussion

The purpose of data augmentation is to improve the generalization ability of models by increasing data diversity, reduce overfitting, and ultimately enhance model performance on unseen data. Different augmentation methods can have varying impacts on model performance. In practical applications, the choice of which augmentation method to use depends on the specific classification task and objectives. Ideally, the specific impact of different augmentation methods on model performance should be evaluated through techniques such as cross-validation.

Based on the comparison of different augmentation methods and their effects on classification models discussed above, we can conclude:

(1)The method proposed in this paper demonstrates significant improvements across all models, particularly on FBCSP and Deep4Net, where the enhancements are most substantial, with increases of 11.92% and 29.84%, respectively. This indicates that the proposed method is an effective data augmentation technique that can significantly improve the generalization ability of models.

(2)Noise injection also shows considerable improvement on most models, especially on LSTM and ShallowNet, where the enhancement exceeds 30%, demonstrating the potential of noise addition in improving model robustness. However, the level of noise needs careful tuning to avoid obscuring useful information within the signal.

(3)Data expansion is beneficial for all types of models, especially for deep learning models that require large amounts of data to optimize parameters, as it can significantly improve the statistical performance of models and reduce overfitting.

(4)Combining multiple augmentation methods generally provides the best performance enhancement, but care must be taken to avoid excessive augmentation that may distort the signal, thereby reducing model performance. Therefore, the selection of augmentation methods and models needs to be carefully adjusted based on the characteristics of the data and the requirements of the task.

5. Conclusion

In this paper, we introduce an augmentation scheme based on Gaussian microstate feature reconstruction for EEG data to address issues related to the disruption of spatiotemporal dynamic features or reliance on a large number of real samples. The scheme involves Gaussian clustering of similar samples to extract mixed representations of microstates, followed by exchanging and reconstructing new Gaussian mixture model probability features based on the similarity of probability features between two samples of the same type. New sample data is generated through sampling based on probability, mean, and variance. Its performance was tested and evaluated on datasets. Experimental results demonstrate that our proposed scheme is an effective data augmentation technique in terms of motor imagery recognition accuracy, significantly enhancing the generalization ability of models.

Acknowledgment

This work is funded by the Special Fund for Research on National Major Research Instruments (62227801) of the Nature Science Foundation of China.

References

[1] Abdul Hussain, A., Singh, A., Guesgen, H., Lal, S., 2021. A comprehensive review on critical issues and possible solutions of motor imagery based electroencephalography brain-computer interface. Sensors 21. doi:10.3390/s21062173.

[2] Ang, K.K., Chin, Z.Y., Wang, C., Guan, C., Zhang, H., 2012. Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. Frontiers in Neuroscience 6. URL: https://www.frontiersin.org/articles/10.3389/fnins.2012.00039, doi:10.3389/fnins.2012.00039.

[3] Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, JMLR.org. p. 214âĂŞ223. doi:https://doi.10.5555/3305381.3305404.

[4] Bao, G., Yan, B., Tong, L., Shu, J., Wang, L., Yang, K., Zeng, Y., 2021. Data augmentation for eeg-based emotion recognition using generative adversarial networks. Frontiers in Computational Neuroscience 15. URL: https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2021.723843, doi:10.3389/fncom.2021.723843.

[5] Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.M., Robbins, K.A., 2015. The prep pipeline: standardized preprocessing for large-scale eeg analysis. Frontiers in Neuroinformatics 9. URL: https://doi.org/10.3389/fninf.2015.00016, doi:10.3389/fninf.2015.00016.

[6] Blanchard, G., Blankertz, B., 2004. Bci competition 2003–data set iia: spatial patterns of self-controlled brain rhythm modulations. IEEE Trans Biomed Eng 51, 1062–1066. doi:https://lampz.tugraz.at/~bci/database/001-2014/description.pdf.

[7] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. J. Artif. Int. Res. 16, 321âĂŞ357.

[8] Deiss, O., Biswal, S., Jing, J., Sun, H., Westover, M.B., Sun, J., 2018. Hamlet: Interpretable human and machine co-learning technique. ArXiv doi:10.48550/arXiv.1803.09702.

[9] Fahimi, F., Dosen, S., Ang, K.K., Mrachacz-Kersting, N., Guan, C., 2021. Generative adversarial networks-based data augmentation for brainâĂŞcomputer interface. IEEE Transactions on Neural Networks and Learning Systems 32, 4039–4051. doi:10.1109/TNNLS.2020.3016666.

[10] Fu, R., Wang, Y., Jia, C., 2022. A new data augmentation method for eeg features based on the hybrid model of broad-deep networks. Expert Systems with Applications 202, 117386. URL: https://www.sciencedirect.com/science/article/pii/S0957417422007321, doi:https://doi.org/10.1016/j.eswa.2022.117386.

[11] Gramfort, A., e, D.S., Haueisen, J., HÃďmÃďlÃďinen, M., Kowalskig, M., 2013. Time-frequency mixed-norm estimates: Sparse m/eeg imaging with non-stationary source activations. NeuroImage 70, 410–422. URL: https://www.sciencedirect.com/science/article/pii/S1053811912012372, doi:https://doi.org/10.1016/j.neuroimage.2012.12.051.

[12] Gramfort, A., Strohmeier, D., J. Haueisen e, f., d, M.H., g, M.K., 2019. BrainâĂŞmachine interfaces from motor to mood. Nature Neuroscience 22, 1554–1564. URL: https://doi.org/10.1038/s41593-019-0488-y, doi:https://doi.org/10.1038/s41593-019-0488-y.

[13] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A., 2017. Improved training of wasserstein gans, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. p. 5769âĂŞ5779. doi:10.5555/3295222.3295327.

[14] Hartmann, K.G., Schirrmeister, R.T., Ball, T., 2018. Eeg-gan: Generative adversarial networks for electroencephalograhic (eeg) brain signals. ArXiv doi:https://doi.10.48550/arXiv.1806.01875.

[15] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[16] Jas, M., Engemann, D.A., Bekhti, Y., Raimondo, F., Gramfort, A., 2017. Autoreject: Automated artifact rejection for meg and eeg data. NeuroImage 159, 417–429. URL: https://www.sciencedirect.com/science/article/pii/S1053811917305013, doi:https://doi.org/10.1016/j.neuroimage.2017.06.030.

[17] JOUR, Fei, Z., Fei, Y., Yuchen, F., Quan, L., Bairong, S., 2019. Electrocardiogram generation with a bidirectional lstm-cnn generative adversarial network. Scientific Reports 9. URL: https://doi.org/10.1038/s41598-019-42516-z, doi:10.1038/s41598-019-42516-z.

[18] Kim, S.J., Lee, D.H., Choi, Y.W., 2023. Cropcat: Data augmentation for smoothing the feature distribution of eeg signals, in: 2023 11th International Winter Conference on Brain-Computer Interface (BCI), pp. 1–4. doi:10.1109/BCI57258.2023.10078539.

[19] Krell, M.M., Kim, S.K., 2017. Rotational data augmentation for electroencephalographic data, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 471–474. doi:10.1109/EMBC.2017.8036864.

[20] Kuanar, S., Athitsos, V., Pradhan, N., Mishra, A., Rao, K., 2018. Cognitive analysis of working memory load from eeg, by a deep recurrent neural network, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2576–2580. doi:10.1109/ICASSP.2018.8462243.

[21] Lashgari, E., Liang, D., Maoz, U., 2020. Data augmentation for deep-learning-based electroencephalography. Journal of Neuroscience Methods 346, 108885. URL: https://www.sciencedirect.com/science/article/pii/S0165027020303083, doi:https://doi.org/10.1016/j.jneumeth.2020.108885.

[22] Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J., 2018. Eegnet: a compact convolutional neural network for eeg-based brainâĂŞcomputer interfaces. Journal of Neural Engineering 15, 056013. URL: https://dx.doi.org/10.1088/1741-2552/aace8c, doi:10.1088/1741-2552/aace8c.

[23] Li, Y., Huang, J., Zhou, H., Zhong, N., 2017. Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks. Applied Sciences 7. URL: https://www.mdpi.com/2076-3417/7/10/1060, doi:10.3390/app7101060.

[24] Liao, C., Zhao, S., Zhang, J., 2024. Motor imagery recognition based on gmm-jcsfe model. IEEE Transactions on Neural Systems and Rehabilitation Engineering 32, 3348–3357. doi:10.1109/TNSRE.2024.3451716.

[25] Liu, W., Liu, X., Dai, R., Tang, X., 2017. Exploring differences between left and right hand motor imagery via spatio-temporal eeg microstate. Computer Assisted Surgery 22, 258–266. URL: https://doi.org/10.1080/24699322.2017.1389404, doi:10.1080/24699322.2017.1389404, arXiv:https://doi.org/10.1080/24699322.2017.1389404. pMID: 29096552.

[26] Lotte, Fabien, 2015. Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brainâĂŞcomputer interfaces. Proceedings of the IEEE 103, 871–890. doi:10.1109/JPROC.2015.2404941.

[27] Luo, Y., Lu, B.L., 2018. Eeg data augmentation for emotion recognition using a conditional wasserstein gan, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2535–2538. doi:10.1109/EMBC.2018.8512865.

[28] Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P., 2017. Least squares generative adversarial networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2813–2821. doi:10.1109/ICCV.2017.304.

[29] Mohsenvand, M.N., Izadi, M.R., Maes, P., 2020. Contrastive representation learning for electroencephalogram classification, in: Alsentzer, E., McDermott, M.B.A., Falck, F., Sarkar, S.K., Roy, S., Hyland, S.L. (Eds.), Proceedings of the Machine Learning for Health NeurIPS Workshop, PMLR. pp. 238–253. URL: https://proceedings.mlr.press/v136/mohsenvand20a.html.

[30] more, O.S.â.Y.Y.â.M.L.â.â.H.D.â.E.C.â.M.S.S., 2020. Neural decoding and control of mood to treat neuropsychiatric disorders. Biological Psychiatry 87, s95–s96. URL: https://doi.org/10.1016/j.biopsych.2020.02.265, doi:https://doi.org/10.1016/j.biopsych.2020.02.265.

[31] Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22, 1345–1359. doi:10.1109/TKDE.2009.191.

[32] Pei, Y., Luo, Z., Yan, Y., Yan, H., Jiang, J., Li, W., Xie, L., Yin, E., 2021. Data augmentation: Using channel-level recombination to improve classification performance for motor imagery eeg. Proceedings of the IEEE 15, 645–952. doi:10.3389/fnhum.2021.645952.

[33] Ramponi, G., Protopapas, P., Brambilla, M., Janssen, R., 2018. T-CGAN: conditional generative adversarial network for data augmentationin noisy time series with irregular sampling. CoRR abs/1811.08295. URL: http://arxiv.org/abs/1811.08295, arXiv:1811.08295.

[34] Rommel, C., Moreau, T., Gramfort, A., 2021. Cadda: Class-wise automatic differentiable data augmentation for eeg signals. ArXiv abs/2106.13695. URL: https://api.semanticscholar.org/CorpusID:235652429.

[35] Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T.H., Faubert, J., 2019. Deep learning-based electroencephalography analysis: a systematic review. Journal of Neural Engineering 16, 051001. URL: https://dx.doi.org/10.1088/1741-2552/ab260c, doi:10.1088/1741-2552/ab260c.

[36] Saeed, A., Grangier, D., Pietquin, O., Zeghidour, N., 2021. Learning from heterogeneous eeg signals with differentiable channel reordering, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1255–1259. doi:10.1109/ICASSP39728.2021.9413712.

[37] Sakai, A., Minoda, Y., Morikawa, K., 2017. Data augmentation methods for machine-learning-based classification of bio-signals, in: 2017 10th Biomedical Engineering International Conference (BMEiCON), pp. 1–4. doi:10.1109/BMEiCON.2017.8229109.

[38] Schirrmeister, R., Gemein, L., Eggensperger, K., Hutter, F., Ball, T., 2017. Deep learning with convolutional neural networks for decoding and visualization of eeg pathology, in: 2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp. 1–7. doi:10.1109/SPMB.2017.8257015.

[39] Schwabedal, J.T.C., Snyder, J.C., Cakmak, A., Nemati, S., Clifford, G.D., 2018. Addressing class imbalance in classification problems of noisy signals by using fourier transform surrogates. ArXiv , 1–8doi:10.48550/arXiv.1806.08675.

[40] Tait, L., Zhang, J., 2022. Meg cortical microstates: Spatiotemporal characteristics, dynamic functional connectivity and stimulus-evoked responses. NeuroImage 251, 119006. URL: https://www.sciencedirect.com/science/article/pii/S1053811922001355, doi:https://doi.org/10.1016/j.neuroimage.2022.119006.

[41] Um, T.T., Pfister, F.M.J., Pichler, D., Endo, S., Lang, M., Hirche, S., Fietzek, U., Kulić, D., 2017. Data augmentation of wearable sensor data for parkinsonâĂŹs disease monitoring using convolutional neural networks, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, Association for Computing Machinery, New York, NY, USA. p. 216âĂŞ220. URL: https://doi.org/10.1145/3136755.3136817, doi:10.1145/3136755.3136817.

[42] Wang, F., Zhong, S.h., Peng, J., Jiang, J., Liu, Y., 2018. Data augmentation for eeg-based emotion recognition with deep convolutional neural networks, in: MultiMedia Modeling, Springer International Publishing, Cham. pp. 82–93. doi:https://doi.org/10.1007/978-3-319-73600-6_8.

[43] Xingfei, H., Deyang, W., Haiyan, L., Fei, J., Hongtao, L., 2021. Shallownet: An efficient lightweight text detection network based on instance count-aware supervision information, in: Neural Information Processing, Springer International Publishing, Cham. pp. 633–644.

[44] Y, L., LZ, Z., ZY, W., BL, L., 2020. Data augmentation for enhancing eeg-based emotion recognition with deep generative models. Journal of neural engineering 17. doi:10.1088/1741-2552/abb580.