ID-RemovalNet: Identity Removal Network for EEG Privacy Protection with Enhancing Decoding Tasks

Huabin Wang, **Jie Ruan**, **Cunhang Fan** and **Yingfan Cheng** and **Zhao Lv**[*]
Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China
{wanghuabin, cunhang.fan, kjlz}@ahu.edu.cn,{e23201009, e23301217}@stu.ahu.edu.cn

Abstract

Electroencephalogram (EEG) contains not only decoding task information but also personal identity privacy information. If it is stolen or attacked, the user's brain-computer interaction behavior may be maliciously manipulated. Existing EEG identity privacy protection generally adopts generative or adding tiny perturbation methods, which can protect the identity privacy in EEG signals to some extent. However, these methods also damage the performance of decoding task. In order to solve these problems, this paper proposes an identity removal network (ID-RemovalNet) to achieve EEG privacy protection while improving the classification accuracy of decoding task. Firstly, an identity decorrelation separation module is constructed to accurately remove the identity features to achieve privacy protection while reducing the interference with the task decoding features. Secondly, a multi-domain multi-level fusion feature extraction module is designed to extract the high-quality EEG time-frequency features. Finally, the feature enhancement module is used to compensate for the loss of task decoding features and excitation of dominant feature selection during identity feature removal. The experimental results show that ID-RemovalNet removes identity information to 0.43% on four EEG datasets with two different paradigms, and significantly improves the EEG task decoding accuracy by 3.28%, and achieves the state-of-the-art performance in cross-subject EEG experiment.

1 Introduction

Electroencephalography (EEG) as the primary input method for Brain-Computer Interface (BCI), is convenient and cost-effective, but it may leak personal privacy and medical information [Gu *et al.*, 2021]. Studies have shown that BCI has potential in authentication systems [Neupane *et al.*, 2019], but EEG signals collected by consumer-grade BCI devices can expose users' private information, such as credit card details, PIN codes, etc [Martinovic *et al.*, 2012; Choi *et al.*, 2018]. Additionally, EEG signals collected by different BCI devices may be related to each other [Kong *et al.*, 2018]. These findings highlight the urgency of protecting user privacy in BCI applications, especially in preventing EEG data from being stolen or manipulated [Zhang and Wu, 2019]. Therefore, developing techniques that can extract EEG task features without personal identity information is crucial for the further application of BCI.

Initially, methods for the overall protection of EEG signals are proposed; however, traditional anonymization cannot fully prevent the leakage of EEG privacy data. As a result, researchers have explored more complex EEG data protection methods, primarily including data encryption and privacy-preserving machine learning. Data encryption is a relatively traditional protection approach, including homomorphic encryption (HE) [Robinson and Varghese, 2016], secure multiparty computation (SMC) [Agarwal and others, 2018], and secure processors. However, encryption algorithms may disrupt the task-specific features of EEG and still carry the risk of being cracked. Privacy-preserving machine learning methods, including federated learning [Jia *et al.*, 2024] and source-free transfer learning [Xia *et al.*, 2022; Wu *et al.*, 2024; Zhang *et al.*, 2022; Gu *et al.*, 2022], store the original EEG data locally and only transmit model parameters or use API interfaces to execute EEG decoding tasks. However, since the model itself may contain sensitive information, attackers can still reconstruct privacy-related information from the original data through reverse engineering or model theft.

Based on this, there has been a growing emphasis on the safeguarding of identity features within EEG signals. Research on brainwave identity privacy protection is relatively limited, with two main approaches: one is the use of generative networks to synthesize EEG data [Pascual *et al.*, 2021; Singh *et al.*, 2023], where the generated data resembles real data but cannot be used to identify individual identities; the second approach involves adding small perturbations to the brainwave signals, which disrupt the identity-related features in the signals, thus protecting the user's identity information [Meng *et al.*, 2023; Chen *et al.*, 2024]. However, both generative and perturbation methods also impair task classification features, leading to a certain degree of reduction in task classification accuracy. For example, in [Meng *et al.*, 2023], while the identity recognition rate of the EEG signals decreased by 49.51%, the task decoding rate also dropped by

[*]Corresponding author: kjlz@ahu.edu.cn

2.43%.

To address these issues, this paper proposes a brainwave privacy protection method called ID-RemovalNet. The core of this method lies in constructing a brainwave signal decomposition framework. Unlike general brainwave protection methods, the proposed method decomposes the brainwave signal into task-related features, identity privacy features, and noise components, with a focus on protecting identity privacy features. On the one hand, the proposed method uses the multi-Domain multi-Level fusion feature extraction (MDML) module to extract high-quality and comprehensive feature information from brainwave signals. On the other hand, the proposed method designs the identity decorrelation separation (IDS) module. Existing generative and subtle perturbation methods not only protect identity privacy but also interfere with task classification features. Our method directly removes the identity privacy features that are decorrelated from the task, ensuring no impact on subsequent task. Finally, to balance both brainwave privacy protection and decoding enhancement, the proposed method introduces a task feature enhancement module. By using the attention-based adversarial feature selection (AAFS) module to stimulate the selection of dominant features, the proposed method enhances decoding accuracy and cross-domain generalization ability. Additionally, a loss-guided identity-level task feature re-fusion (LITFR) module is designed to further compensate for the loss of task features after the removal of identity information, ensuring the accuracy of task decoding. The major contributions of our paper are outlined as follows:

- The proposed ID-RemovalNet designs the Identity Decorrelation Separation (IDS) module to remove the identity information to protect the EEG identity privacy, and the Feature Enhancement (AAFS&LITFR) module, LITFR effectively compensates for the partial loss of the EEG task features during the removal process, and AAFS stimulates the selection of the dominant features of the EEG task to realize the EEG task decoding enhancement and generalization.
- The proposed ID-RemovalNet designs a multi-domain multilevel fusion feature extraction (MDML) module, which extracts rich EEG features by designing the fusion of global and local features in the time-frequency domain to realize the enhancement of EEG task decoding task.
- ID-RemovalNet reduces identity information to 0.43% across four EEG datasets with two different paradigms, significantly improving brainwave task decoding accuracy by 3.28%, and achieves state-of-the-art recognition performance in cross-subject EEG experiments.

2 Our Proposed ID-RemoveNet Method

As shown in Figure 1, the designed method ID-removalNet works sequentially through three key modules: (1) EEG Data Align effectively reduces subject differences between sessions and improves EEG decoding accuracy; (2) Identity Decorrelation Separation (IDS) Module extracts high-quality task-relevant features and identity features using multi-domain multi-level Fusion Feature Extraction (MDML) Module, and through the decorrelation items (Dec) and feature subtraction (Sub), separates and removes the EEG identity information, and clean task features are obtained; (3) Feature Enhancement Module utilizes the loss-guided identity-level task feature fusion (LITFR) module to compensate for the loss of task features, and the attention-based adversarial feature selection (AAFS) module to capture cross-session stable task features.

2.1 EEG Data Align

Given an EEG dataset $D = \{x_i, y_i, u_i\}_{i=1}^{N}$, where is the sample of the $x_i \in X \subset R^{c\times t}$ is the i-th EEG trial,c denotes the sample channel, t denotes the sample time, $y_i \in Y = \{1, \ldots, K\}$ is the label corresponding to the task, $u_i \in U = \{1, \ldots, U\}$ is the user label of the i-th trial, and N is the number of EEG trials.

Inspired by [He and Wu, 2020], we found that alignment can effectively reduce subject differences across sessions, where EA methods are efficient, completely unsupervised, and show excellent performance in multiple BCI paradigms. As shown in Figure1 1 (A), where for N EEG trials in a given domain, the EA first calculates the Euclidean arithmetic mean $\overline{R}$ of all N spatial covariance matrices:

$$\overline{R} = \frac{1}{N}\sum_{n=1}^{N} X_n (X_n)^T \tag{1}$$

Then, it performs the alignment by:

$$\overline{X}_n = \overline{R}^{-1/2} X_n, \quad n = 1, \ldots, N \tag{2}$$

Thus, EEG data distributions from different domains become more consistent.

2.2 Identity Decorrelation Separation (IDS) Module

EEG signals, as a unique biological feature, contain a wealth of personal information. During cognitive tasks, the extracted task features often contain a large amount of identity information, which not only leads to privacy leakage but also affects the accuracy of task decoding due to the presence of subject variability. As shown in Figure1 1 (B), we propose an identity de-correlation separation module to realize the removal of task-irrelevant identity information, thus obtaining task features containing only a small amount of identity information for EEG decoding and privacy protection. The method focuses on achieving the removal of identity information by decomposing the EEG feature F into three key components, as shown in Equation 3:

$$F = F_{task} + F_{id} + F_{noise} \tag{3}$$

where F_{task} denotes task-related features, F_{id} denotes identity characteristics of the subject, and F_{noise} denotes noise features in EEG. After MDML extraction of features only a small number of noise features are retained and Eq. 3 is further simplified to:

$$F = F_{task} + F_{id} \tag{4}$$

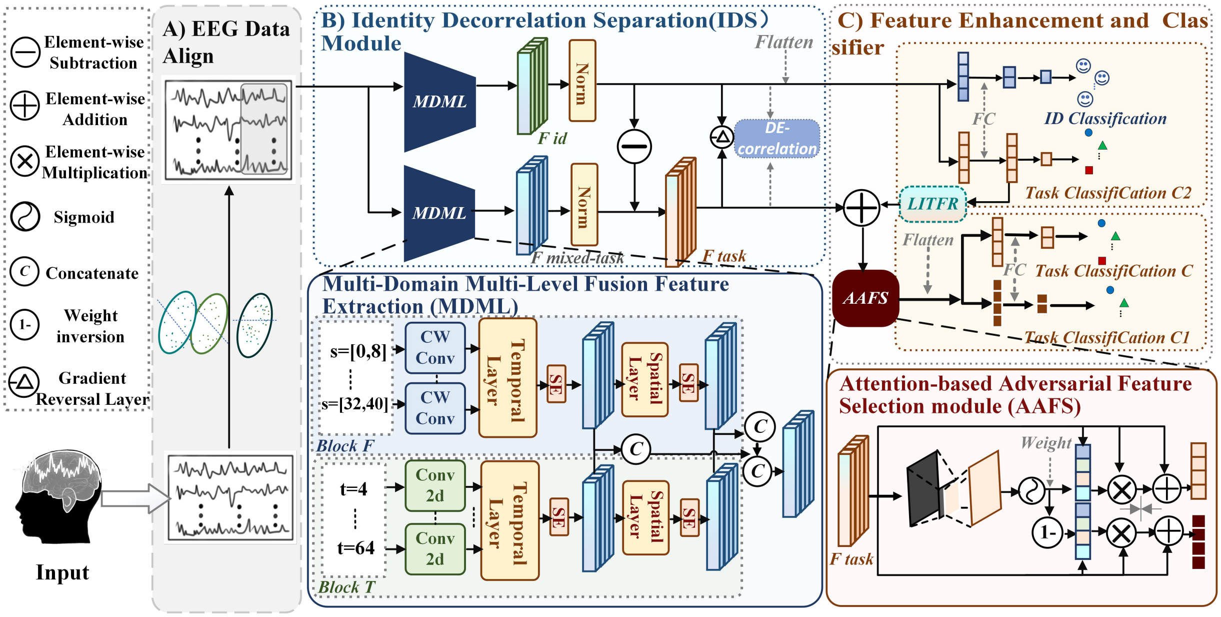

Figure 1: ID-RemovalNet Overview: (A) EEG Data Alignment; (B) Identity Decorrelation Separation (IDS) module, which extracts features through the Multi-Domain Multi-Level Fusion (MDML) module. This module consists of two time-domain and frequency-domain multi-branch parallel networks, focusing on the fusion of global and local features from different domains. It then uses a gradient reversal layer and distance constraints to separate task features and identity features, and identity features are removed by feature subtraction; (C) Feature Enhancement and Classifier, where LITFR supplements task features with identity-level task features guided by loss. AAFS generates attention weights for adversarial training.

Multi-domain Multi-level Fusion Feature Extraction (MDML) To obtain F_{task}, F_{id}, we use a unified backbone network to extract identity features and task features.The designed multi-domain multi-level fusion feature extraction module is shown in Figure1 1: (1) A branch-time scale CNN for converting EEG signals into time-domain discriminative SST representations; (2) A branch-frequency scale learnable continuous wavelet for converting signals into frequency-domain discriminative SST representations; (3) Multi-level fusion of local and global time-frequency features.

Based on [Jia *et al.*, 2021], dynamic adjustment of attention is used to analyze signals at different time scales. ”Block T” designs five parallel temporal convolutional layers to extract time features at different scales. After undergoing Squeeze and Excitation (SE) [Hu *et al.*, 2018] operations, the output local discriminative SST F^1 and global discriminative SST F^2 represent different levels of temporal dependencies. ”Block F”, based on the research from [Liu *et al.*, 2023], adopts five parallel learnable continuous wavelet transformations [Li and al., 2022] to extract features corresponding to specific frequency bands, also outputting local discriminative SST F_f^1 and global discriminative SST F_f^2 through SE.

Finally, unlike fusion methods that only focus on different domain features, the fusion strategy we use combines both temporal and frequency domain features, integrating both local and global features, effectively capturing multi-level and multi-domain feature information. As shown in equation 5:

$$F = \text{concat}((F^1 \oplus F_f^1), (F^2 \oplus F_f^2)) \tag{5}$$

Decorrelation Items (Dec) Due to the complexity of the EEG features, there is a coupled correlation between the task features extracted by MDML and the identity features, so we introduce a decorrelation regularisation term to compute the correlation rho between the two, and by minimising this coefficient. As expressed in Eq. 6, linear dimensionality reduction is performed on task features and identity features to obtain vectors h_{task} and h_{id} , where $[h_{\text{task}}, h_{id}] \subseteq \mathbb{R}^{1\times 1}$

$$rho = \frac{(f(h_{task}) \times f(h_{id}))^2}{g(h_{task}) \times g(h_{id})} \tag{6}$$

where $f(h_{task})$ and $f(h_{id})$ denote doing mean processing on the vectors to make them centred. $g(h_{task})$ and $g(h_{task})$ denote calculating the vector variance.

And the introduction of Gradient Reversal Layer (GRL) [Ganin and Lempitsky, 2015] for domain adversarial training further enhances feature separation. As shown in Equation 7, the GRL reverses the direction of the gradient, prompting the network to optimise the features during the training process so that the task-related features are as distinct as possible from the identity-related features. Where the output of the task classifier is $f(x)$ and the loss of the task classifier is L_d.

$$\text{Gradient Reversal Layer: } \frac{\partial L_d}{\partial f(F_{\text{task}})} = -\frac{\partial L_d}{\partial f(F_{id})} \tag{7}$$

Feature Subtraction (Sub) At this point,the task features obtained are highly decorrelated with the identity features. Finally, by subtracting the features as shown in Eq. 8, the task features containing only minimal identity information can be obtained.

$$F_{\text{task}} = F - F_{id}, \quad [F, F_{\text{task}}, F_{id}] \subset R^{m\times n} \tag{8}$$

2.3 Feature Enhancement and Classifiers

Loss-guided Identity-level Task Feature Re-fusion (LITFR) Since task features and identity features are not perfectly linearly related, simple removal of identity information may result in the loss of task decoding features. As shown in Figure1 1 (C), the loss-guided identity-level task feature re-fusion module calculates the loss of identity features to the task and dynamically adjusts the feature fusion weights to balance the removal of identity information with the integrity of task decoding, as shown in Equation 9:

$$F_{\text{total}} = F_{\text{task}} + \sum_{i=1}^{M} \omega_i f_{\text{id}}(x_i) \tag{9}$$

where $f_{id}(x)$ is the identity-level task feature and F_{task} is the fused task characteristic, ω denotes the feature fusion weight. As shown in Equation 10, τ is the temperature coefficient that controls the weight smoothing degree.

$$\omega = \frac{1}{1 + \exp\left(\frac{\ell_{C_2}}{\tau}\right)} \tag{10}$$

Attention-based Adversarial Feature Selection (AAFS) During the training of a neural network, the model usually activates the main features associated with the labels. However, these features may not work when faced with unseen test data, leading to performance degradation. To address this problem, inspired by the self-challenge mechanism [Huang *et al.*, 2020], the design of an attentional confrontation-based feature selection module reinforces the impact of key features, forcing the model to better mine the features that are most useful for decoding in the task, and at the same time enhances the robustness of the model, thus improving its ability to generalise over unseen data. The attention weights are calculated as shown in Equation 11, where α_i represents the output weight.

$$A_i = \frac{\exp(\log \alpha_i - \log(-\log \varepsilon_i))/\tau)}{\sum_{j=1}^{N} \exp(\log \alpha_j - \log(-\log \varepsilon_j))/\tau)}, \quad \varepsilon \sim U(0,1) \tag{11}$$

where N represents the feature dimension, Gumbel noise is used to introduce randomness, and the temperature of Gumbel-Softmax controls the smoothness of Softmax. We further optimise the model by calculating the dominant feature F_{TD} and the disadvantageous feature F_{TI}.

$$F_{TD} = A \otimes F_{task} \tag{12}$$

$$F_{TI} = (1 - A) \otimes F_{task} \tag{13}$$

where the weights A are learnable parameters that feed F_{TD} and F_{TI} into the the primary classifier C and the secondary classifier C_1 for training, respectively.

As shown in Eq. 14-Eq. 18, under the supervision of cross entropy loss, we train the main feature generator and classifier to predict the correct labels, here ℓ_{CE} cross entropy loss, y is the task label and s is the user label. The loss terms are as follows:

$$\ell_{cls} = \ell_{CE}(H_C(x) \circ A, y) \tag{14}$$

$$\ell_{dcls} = \ell_{CE}(H_D(x), s) \tag{15}$$

$$\ell_{acls} = \ell_{CE}(H_{C_1}(x) \circ (1 - A), y) \tag{16}$$

$$\ell_{tcls} = \ell_{CE}(H_{C_2}(x), y) \tag{17}$$

$$\ell_{GRL} = \ell_{CE}(GRL(H_c(x), y)) \tag{18}$$

Ultimately, we formulate the learning of the model as the following optimisation problem,Here λ is the trade-off coefficient:

$$\begin{aligned} \min_{H_C,H_Q,H_D,C_1,C_2,D} =&\lambda_1\ell_{cls} + \lambda_2\ell_{GRL} + \lambda_3\ell_{dcls} \\ &- \lambda_4 rho - \lambda_5\ell_{acls} + \lambda_6\ell_{tcls} \end{aligned} \tag{19}$$

Datasets	Subjects	Points	Channels	Trails	Sessions
MI4C	9	1000	22	576	2
MI2C	9	1000	3	240	2
P300	8	128	32	3300	4
ERN	16	166	56	300	5

Table 1: Details of the four datasets used in the experiment

3 Experiments

3.1 Datasets

The following four public databases are used in this experiment. See Table 1:

(1)Four-class motor imagery dataset (MI4) [Tangermann *et al.*, 2012]: Derived from BCI Competition IV dataset 2a. data were sampled at 250 Hz for 22 EEG channels. Data were extracted within 0-4 seconds after each imagery cue and bandpass filtered at 8-32 Hz.

(2)P300 evoked potentials (P300) [Hoffmann *et al.*, 2008]: Data were recorded on 32 channels at a frequency of 2048 Hz. subsequently downsampled to 128 Hz. the duration of each EEG signal epoch was 0-1 seconds.

(3)Two-class motor imagery dataset (MI2C) [Abelson *et al.*, 1985]: Derived from the BCI Competition IV dataset 2b.The first two sessions without visual feedback were used in this paper, sampled at 250 Hz and containing 3 EEG channels. Data within 0-4 seconds were extracted after each imagery cue and band-pass filtered at 1-40 Hz.

(4)Feedback Error-Related Negativity (ERN) [Leeb *et al.*, 2008]: Sourced from a competition at the 2015 IEEE Neuroengineering Conference, this paper uses a training set from 16 users. Data was recorded on 56 channels at 200 Hz, post downsampling to 128 Hz. epochs of EEG signals from 0-1.3 seconds were extracted.

3.2 Baseline Methods

We use the following three CNN models as baseline feature extractors and retain the last fully-connected layer of each model as a task classifier, while using two fully-connected layers as identifiers:

(1) EEGNet [Lawhern *et al.*, 2018]: A compact CNN architecture designed for EEG classification tasks with deeply separable convolution.

Datasets	Models	Uncorrelated EEG(%)		Decorrelated EEG(%)		Reducation(%)	
		BCA	UIA	BCA	UIA	BCA	UIA
MI4C	EEGNet	61.90	91.50	68.40	**0.04**	6.50	-91.48
	ShallowCNN	61.95	90.17	69.74	**0.04**	**7.79**	-90.13
	DeepCNN	58.12	81.32	65.12	0.12	7.00	-81.20
	MDML	**69.48**	**92.37**	**72.45**	0.25	2.97	**-92.12**
MI2C	EEGNet	59.00	64.52	61.19	**0.37**	2.19	-64.15
	ShallowCNN	58.69	48.43	59.58	2.08	0.89	-46.35
	DeepCNN	56.90	37.08	59.72	0.73	2.82	-36.35
	MDML	**61.30**	**74.17**	**65.57**	0.55	**4.27**	**-73.62**
ERN	EEGNet	64.29	59.90	67.21	0.50	2.92	-59.40
	ShallowCNN	66.03	77.82	67.57	0.54	1.54	-76.54
	DeepCNN	67.51	65.14	68.64	0.38	1.13	-64.76
	MDML	**67.75**	**80.28**	**71.38**	**0.24**	**3.63**	**-80.04**
P300	EEGNet	67.89	91.70	69.47	0.55	1.58	-91.15
	ShallowCNN	65.14	**97.21**	67.64	0.20	2.50	**-97.01**
	DeepCNN	66.60	96.08	69.42	0.20	**2.82**	-95.88
	MDML	**68.89**	91.09	**70.85**	**0.06**	1.96	-91.06
Average		63.84	77.42	67.12	0.43	3.28	77.05

Table 2: Task Recognition Accuracy and Identity Recognition Accuracy of Raw EEG Signals and Identity-related Features

(2) ShallowCNN [Schirrmeister *et al.*, 2017]:A shallow version of DeepCNN with only one convolutional block and with a larger kernel and different pooling.

(3) DeepCNN [Schirrmeister *et al.*, 2017]: contains four convolutional blocks, the first one is designed for EEG inputs and the remaining three are standard convolutional blocks.

3.3 Hyperparameterization

In our experiments, we used a batch size of 128, an initial learning rate of 0.01 and adjusted the learning rate to 0.001 after 50 rounds, model training was performed a total of 100 times and the best model was selected for testing. We evaluate the performance of the task classifier using Balanced Classification Accuracy (BCA) and the performance of the identity classifier using Identity Accuracy (UIA). For each database, a leave-one-session cross-validation is performed, and the mean value of the three experiments is reported.For each loss function of joint multi-task learning, we test the hyperparameters in the range of 0.01-1, and use a stepwise tuning approach, i.e., adjusting each hyperparameter from smallest to largest while keeping the other hyperparameters unchanged until the optimal value is found. After each hyperparameter is determined to be optimal, the value is fixed to ensure that each hyperparameter is optimized. In addition, based on experience, we designed generic parameter settings for each database: 1.00, 0.05, 1.00, 0.05, 0.01, 1.00.

4 Results

4.1 Identity Protection and Task Decoding

To evaluate the performance of the proposed model in EEG identity privacy protection and task decoding, we conducted experiments using four models on four datasets. The specific data are shown in Table 2. The data display the EEG decoding rate and user recognition rate under the original data, as well as the EEG decoding rate and user recognition rate of task features after identity-related features have been removed.

A higher UIA value in raw EEG indicates that the raw data contains a higher degree of identity information, meaning that it is relatively easy to identify the source of a segment of EEG features

Experiments have proved that ID-RemovalNet performs well in identity protection, and the UIA value of Decorrelated EEG decreases by up to 92.12% for MI4C, 73.62% for MI2C, 80.04% for ERN, and 97.01% for P300.

Experiments have demonstrated that ID-RemovalNet enhances the decoding of EEG tasks, and the BCA value of Decorrelated EEG increased up to 7.79% for MI4C, 4.27% for MI2C, 3.63% for ERN and 2.82% for P300;

4.2 Ablation Experiment

We performed an ablation experiment by adding each component of the method step-by-step, and the results of the ablation experiment are listed in Table 3.

The baseline data are the results without any added manipulation, raw EEG decoded by the task as well as user recognition.

We added the alignment treatment and BCA values were substantially increased and UIA values showed a decrease in the motor imagery MI4C and MI2C databases. This suggests that alignment is more favorable for task decoding of EEG data.

We added the IDS component, where ”Dec” represents the feature decorrelation component and ”Sub” represents the feature subtraction component. The data from the four databases combined shows that after adding ”Dec”, the task-related information and identity-related information in the EEG signals were separated, resulting in a significant decrease in user identification performance, dropping to 5%-20%. But the BCA impact was minimal. After adding ”Sub”,

Models	MI4C(%)								MI2C(%)							
	EEGNet		ShallowCNN		DeepCNN		MDML		EEGNet		ShallowCNN		DeepCNN		MDML	
	BCA	UIA	BCA	UIA	BCA	UIA	BCA	UIA	BCA	UIA	BCA	UIA	BCA	UIA	BCA	UIA
Baseline	53.22	92.67	51.81	91.05	47.42	78.61	54.92	94.37	58.06	76.39	55.69	67.82	55.11	56.02	58.18	81.84
+Align	61.91	91.50	61.95	90.17	58.12	81.32	69.48	92.37	59.00	64.52	58.68	48.43	56.90	37.08	61.30	74.17
+Dec.	62.11	12.87	61.24	14.72	56.78	8.89	68.38	15.14	57.61	9.74	57.62	17.02	56.79	11.45	60.14	12.76
+Sub.	62.11	0.42	61.24	0.14	56.78	0.02	68.38	0.01	57.61	0.14	57.62	0.37	56.79	1.68	60.14	0.52
+AAFS.	67.84	0.05	68.98	0.05	64.64	0.35	71.39	0.16	59.91	0.19	59.44	0.93	58.47	0.14	64.26	0.45
All w/LITFR.	68.40	0.04	69.74	0.04	65.12	0.12	72.45	0.25	61.19	0.37	59.58	2.08	59.72	0.23	65.57	0.55

Models	P300(%)								ERN(%)							
	EEGNet		ShallowCNN		DeepCNN		MDML		EEGNet		ShallowCNN		DeepCNN		MDML	
	BCA	UIA	BCA	UIA	BCA	UIA	BCA	UIA	BCA	UIA	BCA	UIA	BCA	UIA	BCA	UIA
Baseline	50.86	41.72	60.46	84.93	62.37	68.83	65.32	93.34	63.92	66.76	64.07	68.23	63.57	64.09	64.41	59.47
+Align	67.89	91.70	65.14	97.21	66.60	96.08	68.69	91.09	64.29	59.90	66.03	77.82	67.51	65.14	67.75	80.28
+Dec.	65.64	10.79	66.75	14.34	67.31	16.44	68.40	13.49	64.83	6.86	67.11	7.65	67.68	6.70	69.10	6.08
+Sub.	65.64	0.18	66.75	0.13	67.30	0.13	68.40	0.06	64.83	1.05	67.11	0.24	67.69	0.25	68.74	0.57
+AAFS.	68.59	0.03	67.45	0.06	67.86	0.02	69.44	0.03	66.32	0.40	67.51	0.32	68.61	0.37	70.64	0.13
All w/LITFR.	69.47	0.55	67.64	0.20	69.42	0.20	70.85	0.06	67.21	0.50	67.57	0.54	68.64	0.38	71.38	0.24

Table 3: In the ablation study on MI4C, MI2C, ERN, and P300, the table presents task and user recognition accuracies for the three basic models and MDML. The third and fourth rows show results with IDS, while the fifth and sixth rows show results with feature enhancement (AAFS&LITFR).

Models	MI4C(%)		ERN(%)		MI2C(%)		P300(%)		TD	FD	LC	GC
	BCA	UIA	BCA	UIA	BCA	UIA	BCA	UIA				
MDML	72.45	0.25	71.38	0.24	65.57	0.55	70.85	0.06	✓	✓	✓	✓
MDML w/o Time-domain	71.97	0.12	70.74	0.57	64.44	0.23	70.01	0.19		✓		
MDML w/o Frequency-domain	71.01	0.10	68.36	1.56	64.20	0.29	70.26	0.08	✓			
MDML w/o Local concat	67.84	0.95	70.31	0.59	64.81	0.77	70.28	0.18	✓	✓		✓
MDML w/o global concat	70.94	0.35	69.88	0.63	65.03	2.96	68.96	0.20	✓	✓	✓	

Table 4: In the ablation studies on MI4C, MI2C, ERN, and P300, the task and user recognition accuracies achieved by MDMLNet are shown. ”TD” refers to the time-domain branch, ”FD” to the frequency-domain branch, ”GC” to global fusion, and ”LC” to local fusion.

the BCA remained almost unchanged, but the UIA once again dropped significantly to below 2.5%.

We continue to add feature enhancement components, including attention-based adversarial feature selection (AAFS) and Loss-guided Identity-level Task Feature Re-fusion (LITFR). The data from the four databases combined shows that after adding only the AAFS component, the UIA value experiences a slight fluctuation while maintaining a low value, but the BCA value sees a significant increase. Meanwhile, after adding the LITFR component, there is a slight increase across different databases.

4.3 Ablation Experiment of MDML

To verify the effectiveness of our proposed MDML in extracting high-quality EEG task features and identity features, we conducted an ablation experiment by adding each branch (TD&FD) and fusion method (FC&GC) in the method step by step, and the results of the ablation experiment are listed in Table 4.

As shown in Table 4, the multi-scale frequency domain convolution outperforms the multi-scale time domain convolution, except for P300, where globally fused features are better for recognition on motion imagery MI4C, MI2C, but local features perform better on event-related potentials ERN, P300.

The experiments show that MDML with complete structure exhibits optimal results for BCA on all four databases, its optimal BCA is about 0.77% higher on average than the rest of the results, and it exhibits the lowest UIA values on the ERN and P300 databases.

In addition we can get clean features for the task with any network model either in the training phase or in the testing phase. This makes our method highly applicable and can achieve the desired purpose with any database and feature extractor.

5 Discussion

5.1 Cross-subject Comparison Experiments

Cross-subject comparison experiments in EEG are a significant challenge in Brain-Computer Interface (BCI) research. However, due to physiological differences between individuals, there is considerable variability in EEG data across subjects, which introduces many challenges for cross-subject

Approach	Subject(%)									Avg
	1	2	3	4	5	6	7	8	9	
EEGNet [Lawhern *et al.*, 2018]	79.9	57.6	89.9	63.9	58.3	60.1	81.6	72.2	64.6	69.8
ShallowCNN [Schirrmeister *et al.*, 2017]	78.8	48.6	87.5	63.9	64.9	62.5	81.9	79.2	74.7	71.3
DeepCNN [Schirrmeister *et al.*, 2017]	67.0	55.6	76.4	61.8	61.8	61.5	82.6	64.6	59.0	65.6
EEGNet* [Lawhern *et al.*, 2018]	82.3	63.5	93.1	70.5	72.2	67.4	88.5	83.7	72.2	77.0
ShallowCNN* [Schirrmeister *et al.*, 2017]	81.9	62.2	90.3	70.8	72.6	66.3	84.4	82.6	76.7	76.4
DeepCNN* [Schirrmeister *et al.*, 2017]	79.2	56.3	84.7	67.4	66.8	60.8	84.4	75.7	70.5	71.8
DRDA [Zhao *et al.*, 2021]	83.2	55.1	87.4	75.3	62.3	57.2	86.2	83.6	82.0	74.8
TCNet-Fusion [Musallam *et al.*, 2021]	72.1	53.0	84.2	65.1	64.8	57.0	78.0	83.0	72.9	70.0
ATCNet [Altaheri *et al.*, 2023]	74.0	54.7	87.9	62.1	63.4	55.7	77.3	86.9	75.3	71.0
TMSANet [Zhao and Zhu, 2025]	82.3	49.7	74.0	64.2	44.4	63.2	72.2	76.7	72.2	66.5
BDAN-SPD [Wei *et al.*, 2024]	89.0	57.2	92.7	74.6	55.8	58.6	93.1	88.9	87.5	77.5
ID-RemovalNet	**90.3**	**63.9**	**96.2**	**83.7**	**72.6**	**73.6**	87.5	**92.7**	80.2	**82.3**

Table 5: BCI 2a Cross-Subject Experimental Comparison, where the representation of the model with * uses ID-RemovalNet

Approach	Average		
	ERN	P300	BCI2b
EEGNet [Lawhern *et al.*, 2018]	68.0	73.8	60.1
ShallowCNN [Schirrmeister *et al.*, 2017]	71.1	71.4	59.2
DeepCNN [Schirrmeister *et al.*, 2017]	72.0	73.5	62.7
EEGNet* [Lawhern *et al.*, 2018]	70.6	74.3	67.7
ShallowCNN* [Schirrmeister *et al.*, 2017]	72.8	73.1	63.7
DeepCNN* [Schirrmeister *et al.*, 2017]	73.5	76.9	64.6
EEG-TCNet [Ingolfsson *et al.*, 2020]	68.9	76.9	65.8
TCNet-Fusion [Musallam *et al.*, 2021]	68.1	77.1	65.8
ATCNet [Altaheri *et al.*, 2023]	69.0	76.7	66.0
JMNet [Kim *et al.*, 2024]	73.1	78.3	64.1
TMSANet [Zhao and Zhu, 2025]	70.3	76.1	65.8
ID-RemovalNet	**76.5**	**79.9**	**70.8**

Table 6: ERN,P300,BCI 2b Cross-Subject Experimental Comparison, where the representation of the model with * uses ID-RemovalNet

decoding. ID-RemovalNet has been shown to effectively separate identity features, protecting privacy while reducing interference with task-related features. We reproduced the code from existing papers and conducted cross-subject experiments on the BCI 2a, BCI 2b, ERN, and P300 databases. To ensure experimental fairness, the ”cross-subject comparison experiment” was conducted by applying the model trained on data from one session to the remaining sessions, using test data features for cross-subject experiments. DRDA [Zhao *et al.*, 2021] and BDAN-SPD [Wei *et al.*, 2024] used the data from the original papers.

Baseline Model with ID-RemovalNet Table 5 shows that we selected the base models EEGNet, ShallowCNN, and DeepCNN for the experiments, where each subject becomes the target domain once, and the experiments show that the cross-subject performance of EEGNet, ShallowCNN, and DeepCNN with ID-RemovalNet outperforms the original data, specifically, for BCI2a , EEGNet improved from 69.8% to 77.0%, ShallowCNN from 71.3% to 76.4%, and DeepCNN from 65.6% to 71.8%; per-subject improvement was also achieved, and in addition, Table 6 further demonstrates the applicability of ID-RemovalNet on the ERN, P300, and BCI 2b databases, the results of all baseline models using ID-RemovalNet still outperform the original data.

ID-RemovalNet and Others In addition, the proposed ID-RemovalNet achieves a high accuracy of 82.3%, 76.5%, 79.9% and 70.8% on BCI2a, ERN, P300, and BCI2b, respectively, a result that outperforms all the other experimental results we reproduced, which demonstrates that our model extracts the task features that effectively removes the interference of the identity information and the feature enhancement module superiority, and also demonstrates that MDML extracts discriminative features that further enhance the classification performance.

6 Conclusion

In this paper, we propose a new privacy-preserving framework for EEG signals, ID-RemovalNet, which aims to address the challenge of existing privacy-preserving methods to maintain decoding performance while protecting the privacy of EEG data. The framework efficiently extracts high-quality EEG features through a multidomain multilevel fusion feature extraction module and strips identity features using an identity de-correlation separation module, which protects individual privacy and reduces the interference of identity information on task features. In addition, ID-RemovalNet further optimises the task feature selection through feature enhancement to compensate for feature loss and effectively enhance the decoding accuracy. Experiments show that ID-RemovalNet removes identity information to 0.43% on four EEG datasets in two different paradigms, while significantly improving the EEG task decoding accuracy by 3.28% and reaching the optimum in cross-subject experiments. Meanwhile, in future research, we will focus on the remaining privacy information in EEG and use multi-feature removal networks to further protect EEG.

Acknowledgments

This work was supported by the Natural Science Foundation for the Higher Education Institutions of Anhui Province (Grant No. 2022AH050091), the Science and Technology Innovation Strategy and Soft Science Research Project of Anhui Province (Grant NO.202306f01050020)

References

[Abelson *et al.*, 1985] Harold Abelson, Gerald Jay Sussman, and Julie Sussman. *Structure and Interpretation of Computer Programs*. MIT Press, Cambridge, Massachusetts, 1985.

[Agarwal and others, 2018] A. Agarwal et al. Privacy-preserving linear regression for brain-computer interface applications. In *Proc. IEEE Int. Conf. Big Data*, pages 5277–5278, December 2018.

[Altaheri *et al.*, 2023] H. Altaheri, G. Muhammad, and M. Alsulaiman. Physics-informed attention temporal convolutional network for eeg-based motor imagery classification. *IEEE Transactions on Industrial Informatics*, 19(2):2249–2258, February 2023.

[Chen *et al.*, 2024] X. Chen, S. Li, Y. Tu, Z. Wang, and D. Wu. User-wise perturbations for user identity protection in eeg-based bcis. *Journal of Neural Engineering*, October 2024.

[Choi *et al.*, 2018] G.-Y. Choi, S.-I. Choi, and H.-J. Hwang. Individual identification based on resting-state eeg. In *Proc. Int. Conf. Brain-Comput. Interface*, pages 1–4, January 2018.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15)*, pages 1180–1189, 2015.

[Gu *et al.*, 2021] Xiaotong Gu, Zehong Cao, Alireza Jolfaei, Peng Xu, Dongrui Wu, Tzyy-Ping Jung, and Chin-Teng Lin. Eeg-based brain-computer interfaces (bcis): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 18(5):1645–1666, September 2021.

[Gu *et al.*, 2022] T. Gu, Z. Wang, X. Xu, D. Li, H. Yang, and W. Du. Frame-level teacher-student learning with data privacy for eeg emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, April 2022. early access.

[He and Wu, 2020] H. He and D. Wu. Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, February 2020.

[Hoffmann *et al.*, 2008] U. Hoffmann, J.-M. Vesin, T. Ebrahimi, and K. Diserens. An efficient p300-based brain–computer interface for disabled subjects. *J. Neurosci. Methods*, 167(1):115–125, January 2008.

[Hu *et al.*, 2018] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, Salt Lake City, UT, USA, 2018.

[Huang *et al.*, 2020] Z. Huang, H. Wang, E. P. Xing, and D. Huang. Self-challenging improves cross-domain generalization. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 124–140, 2020.

[Ingolfsson *et al.*, 2020] T. M. Ingolfsson, M. Hersche, X. Wang, N. Kobayashi, L. Cavigelli, and L. Benini. Eeg-tcnet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2958–2965, Toronto, ON, Canada, 2020.

[Jia *et al.*, 2021] Ziyu Jia, Youfang Lin, Jing Wang, Xuehui Wang, Peiyi Xie, and Yingbin Zhang. Salientsleepnet: Multimodal salient wave detection network for sleep staging. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 2614–2620, 2021.

[Jia *et al.*, 2024] T. Jia, L. Meng, S. Li, J. Liu, and D. Wu. Federated motor imagery classification for privacy-preserving brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:3442–3451, 2024.

[Kim *et al.*, 2024] Jun-Mo Kim, Keun-Soo Heo, Dong-Hee Shin, Hyeonyeong Nam, Dong-Ok Won, Ji-Hoon Jeong, and Tae-Eui Kam. A learnable continuous wavelet-based multi-branch attentive convolutional neural network for spatio–spectral–temporal eeg signal decoding. *Expert Systems with Applications*, 251:123975, 2024.

[Kong *et al.*, 2018] X. Kong, W. Kong, Q. Fan, Q. Zhao, and A. Cichocki. Task-independent eeg identification via low-rank matrix decomposition. In *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, pages 412–419, December 2018.

[Lawhern *et al.*, 2018] 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*, 15(5):056013, October 2018.

[Leeb *et al.*, 2008] R. Leeb, C. Brunner, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller. Bci competition 2008–graz data set b. *Graz University of Technology, Austria*, 16:1–6, 2008.

[Li and al., 2022] T. Li and al. Waveletkernelnet: An interpretable deep neural network for industrial intelligent diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(4):2302–2312, April 2022.

[Liu *et al.*, 2023] K. Liu, M. Yang, Z. Yu, G. Wang, and W. Wu. Fbmsnet: A filter-bank multi-scale convolutional neural network for eeg-based motor imagery decoding. *IEEE Trans. Biomed. Eng.*, 70(2):436–445, 2023.

[Martinovic *et al.*, 2012] I. Martinovic, D. Davies, M. Frank, D. Perito, T. Ros, and D. Song. On the feasibility of sidechannel attacks with brain-computer interfaces. *Proc. 21st USENIX Secur. Symp.*, pages 143–158, August 2012.

[Meng *et al.*, 2023] L. Meng, X. Jiang, J. Huang, W. Li, H. Luo, and D. Wu. User identity protection in eeg-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3576–3586, 2023.

[Musallam *et al.*, 2021] Y. K. Musallam, N. I. AlFassam, G. Muhammad, S. U. Amin, M. Alsulaiman, W. Abdul, H. Altaheri, M. A. Bencherif, and M. Algabri. Electroencephalography-based motor imagery classification using temporal convolutional network fusion. *Biological Signal Processing and Control*, 69:102826, August 2021.

[Neupane *et al.*, 2019] A. Neupane, K. Satvat, M. Hosseini, and N. Saxena. Brain hemorrhage: When brainwaves leak sensitive medical conditions and personal information. In *Proc. 17th Int. Conf. Privacy Secur. Trust (PST)*, pages 1–10, August 2019.

[Pascual *et al.*, 2021] D. Pascual, A. Amirshahi, A. Aminifar, D. Atienza, P. Ryvlin, and R. Wattenhofer. Epilepsygan: Synthetic epileptic brain activities with privacy preservation. *IEEE Transactions on Biomedical Engineering*, 68(8):2435–2446, August 2021.

[Robinson and Varghese, 2016] V. Robinson and E. B. Varghese. A novel approach for ensuring the privacy of eeg signals using application-specific feature extraction and aes algorithm. In *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, pages 1–6, August 2016.

[Schirrmeister *et al.*, 2017] R. Schirrmeister, L. Gemein, K. Eggensperger, F. Hutter, and T. Ball. Deep learning with convolutional neural networks for decoding and visualization of eeg pathology. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–7, Philadelphia, PA, USA, 2017.

[Singh *et al.*, 2023] G. Singh, P. Patel, M. Asaduzzaman, and G. Bajwa. Selective eeg signal anonymization using multi-objective autoencoders. In *2023 20th Annual International Conference on Privacy, Security and Trust (PST)*, pages 1–7, Copenhagen, Denmark, 2023.

[Tangermann *et al.*, 2012] M. Tangermann, K. R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz. Review of the bci competition iv. *Frontiers in Neuroscience*, 6:55, July 2012.

[Wei *et al.*, 2024] F. Wei, X. Xu, X. Li, and X. Wu. Bdan-spd: A brain decoding adversarial network guided by spatiotemporal pattern differences for cross-subject mi-bci. *IEEE Transactions on Industrial Informatics*, 20(12):14321–14329, 2024.

[Wu *et al.*, 2024] H. Wu, Z. Ma, Z. Guo, Y. Wu, J. Zhang, and G. Zhou. Online privacy-preserving eeg classification by source-free transfer learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:3059–3070, 2024.

[Xia *et al.*, 2022] K. Xia, L. Deng, W. Duch, and D. Wu. Privacy-preserving domain adaptation for motor imagery-based brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*, 69(11):3365–3376, November 2022.

[Zhang and Wu, 2019] X. Zhang and D. Wu. On the vulnerability of cnn classifiers in eeg-based bcis. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 27(5):814–825, May 2019.

[Zhang *et al.*, 2022] W. Zhang, Z. Wang, and D. Wu. Multi-source decentralized transfer for privacy-preserving BCIs. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 30:2710–2720, 2022.

[Zhao and Zhu, 2025] Qian Zhao and Weina Zhu. Tmsa-net: A novel attention mechanism for improved motor imagery eeg signal processing. *Biomedical Signal Processing and Control*, 102:107189, 2025.

[Zhao *et al.*, 2021] H. Zhao, Q. Zheng, K. Ma, H. Li, and Y. Zheng. Deep representation-based domain adaptation for nonstationary eeg classification. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(2):535–545, Feb. 2021.