Impact of Trial-wise and Test Data Leakage on EEG-Based Emotion Classification

Peihong Lei[1], Mengyao Wu[1], Wenjun Yi[1] and Hanlin Mo[1,*]

[1]School of Artificial Intelligence, Xidian University, 266 Xinglong Section of Xifeng Road, Xi'an, Shaanxi 710126, China

Abstract

Deep learning-based approaches have significantly advanced emotion recognition technology using electroencephalography (EEG) data. However, data leakage poses a major threat to model generalizability. This paper focuses on analyzing two common leakage patterns, test data leakage (test-set-driven hyperparameter tuning) and trial-wise data leakage (where the same trial segment is split between training and test sets), lead to overestimation of deep learning model performance. We systematically quantify the impact of these two leakage types, applying four data processing approaches to the DEAP dataset: normal setting, test data leakage, trial-wise data leakage, and combined test data and trial-wise leakage. Six representative deep learning models were trained and tested under each data processing condition, maintaining identical other factors across all models to control variables. Experimental results demonstrate that under the three leakage conditions, all six models significantly outperform the normal setting: the minimum improvement in valence classification accuracy reached 35.71%, while the minimum improvement in arousal classification accuracy reached 25.00%. Architectures based on convolutional neural networks (CNN) were most affected, while transformer-based models showed smaller but still significant impacts. Further, visualization of the average intermediate features across all EEG data belonging to each class for these models reveals that data leakage induces significant alterations in brain topography patterns. The severity of performance inflation followed the order: combined leakage > trial-wise data leakage > test data leakage. In summary, our findings underscore the critical importance of implementing rigorous data partitioning protocols and leakage-aware experimental designs in both affective computing and neuroscience research. Only in this way can we ensure that artificial intelligence assists researchers in uncovering genuine scientific laws rather than leading them astray.

Keywords

Electroencephalography(EEG), Emotion Classification, Trial-wise Leakage, Test Data Leakage, Inflated Performance.

1. Introduction

Emotional computing has become an increasingly important area in human-computer interaction, cognitive science, and affective neuroscience. Accurate recognition of human emotions enables intelligent systems to better understand and respond to user states, thereby improving user experience and interaction efficiency. Common modalities for emotion recognition include facial expressions, voice, and physiological signals such as electroencephalography (EEG). Among these, EEG-based emotion recognition has unique advantages. It directly measures brain activity, is less susceptible to deliberate masking by the user, and provides fine-grained temporal information. Consequently, EEG-based emotion recognition has broad potential applications in mental health monitoring, adaptive learning, affective gaming, and neurofeedback systems.

Recent advances in EEG-based emotion recognition have been largely driven by deep learning approaches. Methods can be broadly categorized based on network architectures. Convolutional neural networks have been applied to capture spatial-temporal patterns in EEG signals [1, 2]; graph neural networks model the relationships between electrode channels [3, 4]; generative adversarial networks (GANs) have been used to augment EEG data [5]; and Vision Transformers (ViTs) have recently been

4DMR@IJCAI25. International IJCAI Workshop on 1st Challenge and Workshop for 4D Micro-Expression Recognition for Mind Reading, August 29, 2025, Guangzhou, China

*Corresponding author.

23009200885@stu.xidian.edu.cn (P. Lei); 24171214037@stu.xidian.edu.cn (M. Wu); ywj@stu.xidian.edu.cn (W. Yi); mohanlin@xidian.edu.cn (H. Mo)

0000-0002-5155-8157 (H. Mo)

explored for their ability to model long-range dependencies [6, 7]. These models accept various forms of input, including raw EEG time series, frequency-domain features, or two-dimensional image-like transformations of EEG data.

The evaluation of these models typically follows either subject-dependent or subject-independent protocols. In subject-dependent settings, both the training and testing data are derived from the same individuals, whereas in subject-independent settings, the testing set contains only data from individuals unseen during training. Early studies primarily focused on subject-dependent evaluation, achieving increasingly high performance. This success, however, has been partially misleading: recent analysis suggests that many of the apparent improvements were inflated due to hidden data leakage issues, even in supposedly standard evaluation protocols [8, 9].

One common source of leakage in subject-dependent experiments arises from splitting individual trials into multiple segments. While segment-level classification followed by trial-level aggregation (e.g., via voting) is a common practice, segments from the same trial can inadvertently appear in both training and testing sets. A second, often overlooked source of data leakage stems from hyperparameter selection using the test set. Specifically, selecting the number of training epochs or other hyperparameters based on test set performance introduces information from the test set into the training process. Theoretically, such selection should be performed on a separate validation set; however, EEG emotion datasets are often small, prompting researchers to maximize the training data by using the test set for hyperparameter tuning. Tuomas et al. first highlighted this issue in the field of micro-expression recognition, demonstrating that commonly reported model performances were overestimated when proper experimental protocols were enforced [10]. Despite its prevalence, this type of leakage has not been systematically analyzed in EEG-based emotion recognition studies.

In this work, we make the following contributions:

- We systematically analyze two major sources of data leakage in EEG-based emotion recognition—trial-wise leakage and test data leakage—highlighting how these issues can inflate reported performance and potentially mislead neuroscience interpretations.
- Based on the DEAP dataset, we evaluate six widely used EEG emotion classification models under four experimental conditions: normally trained, trial-wise data leakage training, test data leakage training, and combined leakage training. We quantitatively demonstrate the extent to which performance is overestimated.
- Through visualization of intermediate model features, we show that data leakage alters EEG topographic patterns, which may lead to erroneous conclusions in neuroscience and affective computing research.

2. Related Work

This section primarily reviews data leakage in the field of EEG-based emotion classification, with a focus on two types of leakage: trial-wise leakage and test data leakage. Notably, while the former has received attention recently, discussion and analysis of the latter remain scarce.

2.1. Trial-Wise EEG Data Leakage

In EEG analysis tasks (not limited to affective computing), particularly under subject-dependent conditions, it is common to segment continuous EEG signals into shorter epochs for model training and evaluation. Subject-dependent means that the model uses data from the same subjects during both training and testing, i.e., each subject's data is utilized for model learning as well as performance validation. Segmenting long EEG trials into multiple short segments (segmentation or windowing) can effectively increase the number of training samples, enabling deep learning models to learn more robust representations of brain activity patterns and facilitating adaptation to commonly used network architectures [11, 12, 13, 14].

However, special care must be taken during data partitioning: all segments from the same trial must be assigned to the same data split (training, testing, or validation). If segments from the same trial or subject are erroneously distributed across multiple splits, data leakage occurs. In fact, segments from the same subject are far more similar to each other than to segments from different subjects [15]. Such leakage can cause the model to learn subject-specific brain activity patterns rather than abstract representations that generalize to new subjects. As a result, classification accuracy on the test set may appear significantly high, while the model's generalization to new subjects is severely compromised, leading to an overestimation of performance [16, 17]. Recently, some studies have started to address this issue and systematically analyzed the impact of data leakage on model performance and generalizability [8, 9].

2.2. Test EEG Data Leakage

In addition to trial-wise data leakage, another type of data leakage that has long been overlooked in the field of affective computing is test data leakage. A typical scenario occurs when test data are used to select hyperparameters during model training, such as determining the number of training epochs or choosing the optimal model. This effectively allows information from the test set to influence training, leading to overly optimistic performance estimates.

In the micro-expression recognition field, recent studies have systematically analyzed this issue. Kapoor and Narayanan conducted a meta-review of data leakage and reproducibility in machine learning-based sciences, finding that more than 17 research fields and 329 articles were affected by data leakage or similar problems[18]. Specifically, for micro-expression recognition, several articles from 2019–2022 were potentially impacted by test data leakage [10]. The most common case was using test data to determine the number of training epochs—that is, selecting the optimal model during training using the test set. Analyses revealed that some methods reported F1-Scores close to 80 in the original publications, but after correcting for data leakage, the actual performance dropped to around 50 F1-Score. Other issues included feature extraction or preprocessing using test data, which similarly constitute data leakage, causing substantial positive bias. The seriousness of such leakage lies in its potential to mislead researchers regarding the true capabilities of the models.

A similar problem exists in EEG-based affective computing, but it has not yet been systematically identified or analyzed. For example, in 2023, Ding et al. proposed the TSception model [2], a deep learning model for EEG-based emotion classification, and provided detailed experimental protocols and source code on GitHub (https://github.com/yi-ding-cs/TSception). This work has attracted widespread attention in the field, with nearly 300 citations. However, we found that during model training, the authors used the test set, rather than a validation set, to select the optimal training model—i.e., to determine the stopping epoch. This constitutes a typical case of test data leakage, which may lead to overestimation of model performance.

3. Methodology

In this study, we consider an EEG-based emotion classification dataset consisting of N subjects, denoted as:

$$\mathcal{D} = \{S_1, S_2, \ldots, S_N\}$$

where S_i represents all EEG recordings from subject i. Each subject performs M trials, and each trial corresponds to a specific emotional state:

$$S_i = \{T_{i1}, T_{i2}, \ldots, T_{iM}\}$$

Each trial T_{ij} is associated with an emotion label y_{ij}, which can be represented either as one of C discrete emotion classes (e.g., $C = 6$ basic emotions) or as a pair (v_{ij}, a_{ij}) in the valence–arousal (V–A) space.

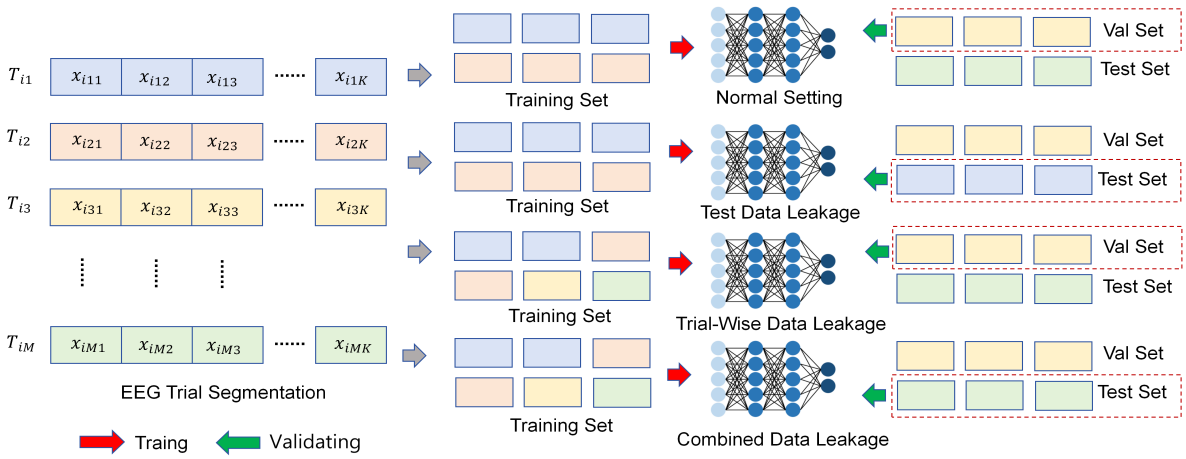

Figure 1: Four experimental configurations based on different data partitioning and training strategies are designed, including normal setting, test data leakage, trial-wise leakage, and combined data leakage.

Under subject-dependent conditions, the model is both trained and tested on data from the same subject. That is, for each subject S_i, the data are divided into disjoint subsets:

$$S_i = S_i^{\text{train}} \cup S_i^{\text{val}} \cup S_i^{\text{test}}, \quad S_i^{\text{train}} \cap S_i^{\text{val}} = \emptyset, \quad S_i^{\text{train}} \cap S_i^{\text{test}} = \emptyset$$

This setting evaluates the model's ability to learn individual-specific patterns of emotional responses, which is a standard approach in EEG-based affective computing. Compared with the subject-independent setting, it is relatively easier to achieve high performance, since the training and testing data come from the same individual and thus share similar physiological and neural characteristics.

Each trial T_{ij} typically contains a long continuous EEG time series. To increase the number of available samples and adapt to deep learning architectures, the trial is divided into K non-overlapping segments:

$$T_{ij} = \{x_{ij1}, x_{ij2}, \ldots, x_{ijK}\}$$

where $x_{ijk} \in \mathbb{R}^{C \times L}$ represents a segment with C EEG channels and L time points per segment. This segmentation process effectively increases the sample size:

$$|\mathcal{D}_{\text{seg}}| = N \times M \times K$$

and enables models to learn more localized and stable representations of emotional brain activity.

However, improper handling of segmented data during dataset partitioning can lead to data leakage, as discussed in Section 2. To systematically investigate the impact of data leakage, we design four experimental configurations based on different data partitioning and training strategies (as shown in Figure 1). Let $\mathcal{D}^{\text{train}}$, $\mathcal{D}^{\text{val}}$, and $\mathcal{D}^{\text{test}}$ denote the training, validation, and testing sets, respectively.

(1) Normal Setting

In the normal (non-leaking) condition, the segmented data are divided in a ratio of $6:2:2$ for training, validation, and testing:

$$|\mathcal{D}^{\text{train}}| : |\mathcal{D}^{\text{val}}| : |\mathcal{D}^{\text{test}}| = 6:2:2$$

All segments originating from the same trial are assigned to the same subset:

$$x_{ijk} \in \mathcal{D}^a \Rightarrow x_{ij(k')} \notin \mathcal{D}^b, \quad a, b \in \{\text{train}, \text{val}, \text{test}\},\ a \neq b$$

During training, the model f_θ is optimized by minimizing the loss $\mathcal{L}$ on $\mathcal{D}^{\text{train}}$, and the best-performing model is selected based on validation accuracy:

$$\theta^* = \arg\min_\theta \mathcal{L}(f_\theta, \mathcal{D}^{\text{val}})$$

The final performance is then evaluated on $\mathcal{D}^{\text{test}}$.

(2) Test Data Leakage In this setup, the data are correctly partitioned as in the normal condition. However, during training, the test set is mistakenly used for model validation, meaning that the best model is selected based on test performance:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(f_\theta, \mathcal{D}^{\text{test}})$$

This causes test data leakage, since the test set influences the training process and model selection, resulting in overly optimistic evaluation results.

(3) Trial-wise Data Leakage This configuration follows the same data split ratio ($6:2:2$) but violates the constraint that segments from the same trial remain in a single subset. In other words, for some trials:

$$\exists\, T_{ij},\ x_{ijk} \in \mathcal{D}^{\text{train}},\ x_{ij(k')} \in \mathcal{D}^{\text{test}}$$

This introduces trial-wise data leakage, allowing information from a trial to appear in both training and testing phases, leading to inflated test accuracy.

(4) Combined Trial-wise and Test Data Leakage This represents the most severe case where both leakage types occur simultaneously. Segments from the same trial are distributed across different subsets, and the test set is also used to determine the optimal model:

$$\begin{cases} \exists\, T_{ij},\ x_{ijk} \in \mathcal{D}^{\text{train}},\ x_{ij(k')} \in \mathcal{D}^{\text{test}} \\ \theta^* = \arg\min_{\theta} \mathcal{L}(f_\theta, \mathcal{D}^{\text{test}}) \end{cases}$$

This configuration leads to severe overestimation of model performance and minimal generalizability to unseen subjects.

The four configurations defined above provide a controlled experimental framework for evaluating how different types of data leakage—trial-wise and test data leakage—affect EEG-based emotion classification. Subsequent experiments quantify the impact of each leakage scenario on model performance and demonstrate how improper dataset handling can mislead conclusions about model effectiveness.

4. Experiments

4.1. Experiment Setting

We conduct experiments based on the widely used EEG emotion recognition benchmark dataset—the DEAP dataset [19]. The dataset contains EEG recordings from 32 participants, with each participant watching 40 one-minute music videos during the experiment to elicit different emotional states. The EEG signals were collected using 32 electrode channels at a sampling rate of 512 Hz. Following the standard preprocessing pipeline provided by the dataset, we downsampled the signals to 128 Hz and applied a 4–45 Hz band-pass filter to remove low-frequency drifts and high-frequency noise. Additionally, the first 3 seconds of each trial were removed to avoid unstable responses from participants at the beginning of the video. Each trial has continuous emotion labels, including the dimensions of Valence and Arousal, with values ranging from 1 to 9. To facilitate classification tasks, these continuous labels were discretized into three levels: low (1–3), medium (4–6), and high (7–9).

We segmented each trial into non-overlapping 1-second segments. Based on the segmented data, we designed four data partitioning and model training approaches as shown in Section 3, including normal setting, test data leakage, trial-wise data leakage and combined trial-wise and test data leakage. These are used to analyze the differences in model performance under various data leakage conditions. We selected six models for our experiments, including EEGNet[1], FCBNet[20], TSception[2], ATCNet[7], VanillaTransformer, and ArjunViT[6]. The first three models (EEGNet, FCBNet, TSception) are based on convolutional neural network (CNN) architectures, while the latter three are built on transformer architectures. All models were implemented using PyTorch and torcheeg frameworks.To isolate the impact of data leakage on model performance while controlling for other variables, we maintained

identical training, validation, and testing procedures across all models. Specifically, we employed the Adam optimizer with a learning rate of 0.001 and a batch size of 64, the epoch number of 100 throughout the training process. All experiments were conducted on an NVIDIA 4090 GPU.

Table 1
Performance comparison of valence classification for 6 models under 4 data leakage conditions (ACC and F1-Score, format: mean ± standard deviation)

Model	Normal Setting		Test Data Leakage		Trial-wise Data Leakage		Combined Leakage	
	ACC	F1-Score	ACC	F1-Score	ACC	F1-Score	ACC	F1-Score
EEGNet	0.45 ± 0.23	0.28 ± 0.17	0.63 ± 0.20	0.43 ± 0.20	0.89 ± 0.11	0.87 ± 0.15	0.90 ± 0.11	0.88 ± 0.13
FCBNet	0.42 ± 0.24	0.25 ± 0.13	0.55 ± 0.20	0.33 ± 0.13	0.57 ± 0.09	0.43 ± 0.10	0.60 ± 0.08	0.46 ± 0.10
TSception	0.41 ± 0.23	0.28 ± 0.20	0.58 ± 0.20	0.38 ± 0.20	0.68 ± 0.13	0.60 ± 0.17	0.70 ± 0.13	0.62 ± 0.17
ATCNet	0.42 ± 0.22	0.27 ± 0.15	0.59 ± 0.19	0.39 ± 0.20	0.87 ± 0.09	0.85 ± 0.12	0.88 ± 0.09	0.86 ± 0.12
VanillaTransformer	0.42 ± 0.25	0.24 ± 0.12	0.54 ± 0.22	0.33 ± 0.19	0.51 ± 0.08	0.29 ± 0.06	0.53 ± 0.08	0.31 ± 0.07
ArjunViT	0.41 ± 0.22	0.28 ± 0.17	0.52 ± 0.22	0.34 ± 0.19	0.50 ± 0.08	0.31 ± 0.06	0.52 ± 0.08	0.31 ± 0.08

4.2. Quantitative and Visualization Analysis

The experimental results presented in Tables 1∼3 reveal several critical observations regarding the impact of data leakage on emotion classification performance. All six models exhibit substantial performance improvements under various data leakage conditions compared to the normal setting. For valence classification (Table 1), the average ACC increases from 0.42 in normal settings to 0.69 in combined leakage conditions - a 64.3% relative improvement. Similarly, for arousal classification (Table 2), average ACC rises from 0.28 to 0.68, representing a 142.9% enhancement.

Table 2
Performance comparison of arousal classification for 6 models under 4 data leakage conditions (ACC and F1-Score, format: mean ± standard deviation)

Model	Normal Setting		Test Data Leakage		Trial-wise Data Leakage		Combined Leakage	
	ACC	F1-Score	ACC	F1-Score	ACC	F1-Score	ACC	F1-Score
EEGNet	0.30 ± 0.18	0.20 ± 0.12	0.42 ± 0.18	0.28 ± 0.12	0.89 ± 0.10	0.88 ± 0.11	0.90 ± 0.10	0.89 ± 0.11
FCBNet	0.27 ± 0.14	0.18 ± 0.10	0.30 ± 0.14	0.20 ± 0.10	0.56 ± 0.09	0.47 ± 0.10	0.59 ± 0.08	0.48 ± 0.10
TSception	0.30 ± 0.16	0.20 ± 0.11	0.40 ± 0.16	0.26 ± 0.09	0.67 ± 0.10	0.63 ± 0.13	0.69 ± 0.10	0.65 ± 0.12
ATCNet	0.27 ± 0.16	0.19 ± 0.11	0.37 ± 0.14	0.26 ± 0.10	0.86 ± 0.08	0.85 ± 0.10	0.88 ± 0.07	0.87 ± 0.09
VanillaTransformer	0.29 ± 0.12	0.20 ± 0.06	0.32 ± 0.12	0.23 ± 0.06	0.48 ± 0.07	0.34 ± 0.08	0.50 ± 0.06	0.34 ± 0.10
ArjunViT	0.26 ± 0.08	0.19 ± 0.05	0.29 ± 0.08	0.22 ± 0.05	0.46 ± 0.07	0.34 ± 0.07	0.49 ± 0.06	0.35 ± 0.09

The severity of performance inflation follows a clear hierarchy: combined leakage > trial-wise data leakage > test data leakage. This pattern is consistent across both emotion dimensions and all evaluation metrics, suggesting that trial-wise leakage has a more pronounced effect than test data contamination alone. However, the impact of test data leakage alone should not be underestimated. For instance, in valence classification, test data leakage caused an average ACC increase from 0.42 (normal setting) to 0.67, while for arousal classification, ACC rose from 0.28 to 0.65. These substantial increments demonstrate that even without trial-wise leakage, test data contamination alone can significantly inflate performance metrics, potentially leading to overly optimistic evaluations and misleading conclusions about model effectiveness.

Traditional CNN-based architectures (EEGNet, ATCNet) show the most dramatic performance boosts under leakage conditions. For instance, EEGNet's ACC for arousal classification jumps from 0.30 to 0.90 under combined leakage. In contrast, transformer-based models (VanillaTransformer, ArjunViT) demonstrate relatively more robustness, though still exhibiting significant inflation. Arousal classification shows greater susceptibility to data leakage effects compared to valence classification. The

performance gap between normal and leaked conditions is more substantial for arousal, particularly in F1-score metrics where arousal shows a 226% improvement versus 111% for valence.

Table 3
Average performance improvement due to data leakage across all models (format: mean ± standard deviation)

Emotion	Normal Setting		Test Data Leakage		Trial-wise Data Leakage		Combined Leakage	
	ACC	F1-Score	ACC	F1-Score	ACC	F1-Score	ACC	F1-Score
Valence	0.42 ± 0.23	0.27 ± 0.16	0.57 ± 0.21	0.37 ± 0.18	0.67 ± 0.18	0.56 ± 0.24	0.69 ± 0.18	0.57 ± 0.24
Arousal	0.28 ± 0.15	0.19 ± 0.09	0.35 ± 0.14	0.24 ± 0.09	0.65 ± 0.19	0.60 ± 0.22	0.68 ± 0.18	0.62 ± 0.22

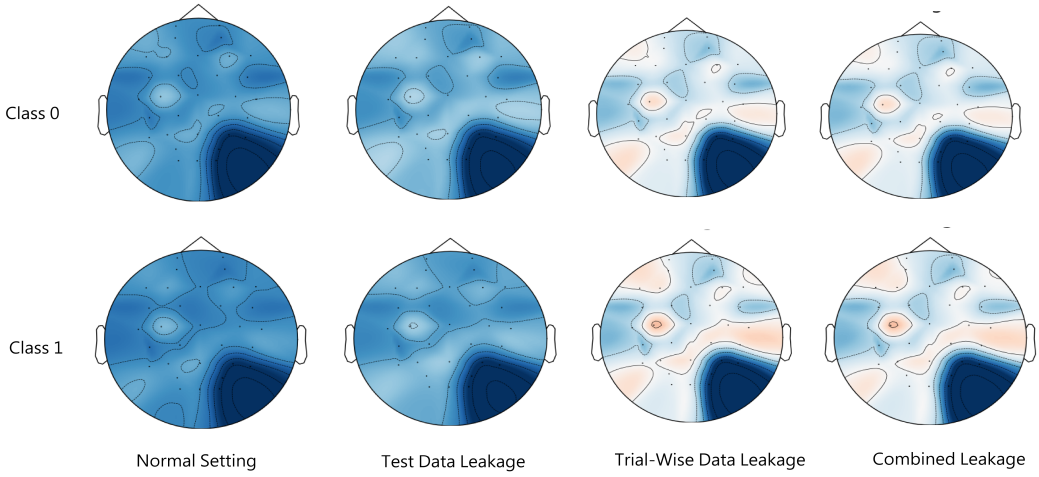

Figure 2: Impact of data leakage on visualization analysis of EEG features from TSception model on valence classification (Class 0: Valence [1,3], Class 1: Valence [4,6]).

The observed performance differences underscore the necessity of rigorous experimental design in EEG-based emotion recognition. The massive performance gaps (e.g., ACC differences up to 0.62 points) highlight how improper data handling can lead to severely overoptimistic results, potentially misleading research conclusions and practical applications.

For the TSception model trained on the valence classification task, we computed the average intermediate features of all EEG data belonging to each class and visualized these averaged features, as shown in Figure 2. The analysis reveals that data leakage induces significant alterations in the brain topography patterns. Notably, trial-wise data leakage produces more pronounced changes compared to test data leakage, which aligns consistently with the quantitative results presented in Table 1∼3.

Previous studies have frequently employed brain topography analysis to investigate the relationships between brain regions and different emotional states, thereby deriving insights that inform cognitive science research. However, our findings demonstrate that when data leakage occurs in experiments, the resulting conclusions regarding brain region correlations may be substantially flawed and unreliable. This underscores the critical importance of rigorous data partitioning protocols in neuroscientific studies involving EEG-based emotion recognition.

5. Conclusion

In this paper, we systematically investigate the critical issue of data leakage in EEG-based emotion recognition research. We identify and analyze two predominant sources of data leakage—test data leakage and trial-wise data leakage—demonstrating how these issues can substantially inflate model performance metrics. Through comprehensive experiments on the DEAP dataset using six established EEG classification models, we quantitatively validate the performance overestimation across four distinct

experimental conditions, revealing the significant impact of different leakage scenarios. Further, our visualization analysis of intermediate model features provides compelling evidence that data leakage fundamentally distorts EEG topographic patterns, thereby challenging the validity of brain region correlations derived from contaminated experimental setups. These findings collectively underscore the critical need for rigorous data partitioning protocols and leakage-aware experimental designs in both affective computing and neuroscience research communities to ensure the reliability and interpretability of future findings.

6. Acknowledgments

This work has partly been funded by the Academy of Finland for Academy Professor project EmotionAI (Grant No.336116), the National Key R&D Program of China (No.2017YFB1002703), and the National Natural Science Foundation of China (Grant No.60873164, 61227802 and 61379082).

7. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and DeepSeek in order to: Grammar and spelling check, Paraphrase and reword. After using these tools and service, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content..

References

[1] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, B. J. Lance, Eegnet: a compact convolutional neural network for eeg-based brain-computer interfaces, Journal of Neural Engineering 15 (2018) 056013.

[2] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, C. Guan, Tsception: capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition, IEEE Transactions on Affective Computing 14 (2023) 2238–2250.

[3] T. Song, W. Zheng, P. Song, Z. Cui, Eeg emotion recognition using dynamical graph convolutional neural networks, IEEE Transactions on Affective Computing 11 (2020) 532–541.

[4] P. Zhong, D. Wang, C. Miao, Eeg-based emotion recognition using regularized graph neural networks, IEEE Transactions on Affective Computing 13 (2022) 1290 – 1301.

[5] Z. Liang, R. Zhou, L. Zhang, L. Li, G. Huang, Z. Zhang, Eegfusenet: hybrid unsupervised deep feature characterization and fusion for high-dimensional eeg with an application to emotion recognition, IEEE Transactions on Neural Systems and Rehabilitation Engineering 29 (2021) 1913 – 1925.

[6] A. Arjun, A. S. Rajpoot, M. R. Panicker, Introducing attention mechanism for eeg signals: emotion recognition with vision transformers, The 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (2021).

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

[8] G. Brookshire1, J. Kasper, N. M. Blauch, Y. C. Wu, R. Glatt, D. A. Merrill, S. Gerrol, K. J. Yoder, C. Quirk, L. Lucero, Data leakage in deep learning studies of translational eeg, Frontiers in Neuroscience 18 (2024).

[9] N. N. Khan, T. Sweet, C. A. Harvey, C. Knapp, D. J. Krusienski, D. E. Thompson, The role of review process failures in affective state estimation: an empirical investigation of deap dataset (2021). https://www.arxiv.org/abs/2508.02417.

[10] T. Varanka, Y. Li, W. Peng, G. Zhao, Data leakage and evaluation issues in micro-expression analysis, IEEE Transactions on Affective Computing 15 (2024) 186–197.

[11] J. Riascos, M. Molinas, F. Lotte, A comparative study on the impacts of data leakage during feature selection using the cic-iot 2023 intrusion detection dataset, The Proceedings of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (2024).
[12] S. Lemm, B. Blankertz, T. Dickhaus, K.-R. Müller, Introduction to machine learning for brain imaging, NeuroImage 56 (2011) 387–399.
[13] H. Chen, J. Li, H. He, J. Zhu, S. Sun, X. Li, B. Hu, Toward the construction of affective brain-computer interface: A systematic review, ACM Computing Surveys 57 (2025) 1–56.
[14] M. de Bardeci, C. T. Ip, S. Olbrich, Deep learning applied to electroencephalogram data in mental disorders: A systematic review, Biological Psychology 162 (2021) 108117.
[15] M. Demuru, M. Fraschini, Eeg fingerprinting: Subject-specific signature based on the aperiodic component of power spectrum, Computers in Biology and Medicine 120 (2020) 103748.
[16] Z. Zhang, J. M. Fort, G. Mateu, Mini review: Challenges in eeg emotion recognition, Frontiers in Psychology 14 (2024) 1289816.
[17] G. Ivucic, S. Pahuja, F. Putze, S. Cai, H. Li, T. Schultz, The impact of cross-validation schemes for eegbased auditory attention detection with deep neural networks, The 46th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (2024).
[18] S. Kapoor, A. Narayanan, Leakage and the reproducibility crisis in machine-learning-based science, Patterns 4 (2023) 100804.
[19] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, Deap: A database for emotion analysis ;using physiological signals, IEEE Transactions on Affective Computing 3 (2012) 18–31.
[20] R. Mane, E. Chew, K. Chua, K. K. Ang, N. Robinson, A. Vinod, S.-W. Lee, C. Guan, Fbcnet: An efficient multi-view convolutional neural network for brain-computer interface (2021). https://arxiv.org/abs/2104.01233.