OPEN ACCESS
International Journal of Neural Systems, Vol. 36, No. 5 (2026) 2650006 (14 pages)
© The Author(s)
DOI: 10.1142/S0129065726500061

World Scientific
www.worldscientific.com

Int. J. Neur. Syst. 2026.36. Downloaded from www.worldscientific.com
by 79.54.49.253 on 03/02/26. Re-use and distribution is strictly not permitted, except for Open Access articles.

Data-related Ablation for Reinforcing Deep Learning in Explaining Complex Phenomena

Romeo Lanzino* and Luigi Cinque†
Department of Computer Science,
Sapienza University of Rome,
Via Salaria 113, Rome 00198, Italy
**lanzino@di.uniroma1.it*
†cinque@di.uniroma1.it

Gian Luca Foresti
Department of Mathematics, Computer Science and Physics
University of Udine, Via delle Scienze,
Udine 33100, Italy
gianluca.foresti@uniud.it

Giuseppe Placidi‡
A²VI-Lab c/o Department of Life,
Health and Environmental Sciences, University of L'Aquila,
Via Vetoio Coppito, L'Aquila 67100, Italy
giuseppe.placidi@univaq.it

Received 3 September 2025
Accepted 2 December 2025
Published Online 30 January 2026

Deep Learning (DL) models excel at automatically learning intricate patterns within complex data, but their black box nature undermines human trust. To address this, current validation strategies typically focus on the model itself, modifying its architecture to assess the role and importance of the components. However, this model-centric view overlooks the critical learning substrate, which is represented by the data, implicitly assuming that it accurately represents the target phenomenon. This implicit trust in data means that evaluation may fail to detect whether high performance stems from exploiting biases or data quirks rather than learning relevant patterns. We present a novel *data-related ablation* as a complement to the traditional architectural ablation. Using this framework for Electroencephalography (EEG) signals of Emotional Recognition (ER) and Motor Execution (ME) as a case study, we show that seemingly high-accuracy models often rely heavily on process-irrelevant features, maintaining performance even when key information is eliminated. This shows that a standard, data-independent evaluation can be misleading about whether a model truly captured the intended process; the proposed approach helps distinguish robust learning from leaning on incidental characteristics. Therefore, incorporating data-related ablation is essential for developing reliable and generalizable DL models in fields that rely on data derived from complex and often not completely known phenomena.

Keywords: Deep learning; bias; artifacts; explainable AI; robust AI; ablation.

‡Corresponding author.

This is an Open Access article published by World Scientific Publishing Company. It is distributed under the terms of the Creative Commons Attribution 4.0 (CC BY) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Int. J. Neur. Syst. 2026.36. Downloaded from www.worldscientific.com
by 79.54.49.253 on 03/02/26. Re-use and distribution is strictly not permitted, except for Open Access articles.

1. Introduction

Unraveling complex phenomena across various domains using Artificial Intelligence (AI) presents a significant scientific challenge,[1] and at its heart lies the need to develop robust and reliable models that can accurately capture underlying patterns and relationships in the data collected on the investigated physical phenomenon. Sophisticated multimodal data collection technologies have led diverse fields into a data-rich era, enabling applications from signal analysis to generative models,[2] from biomedical applications[3–6] to decision-making systems[7] and so on. However, the complexity and variability of the explored phenomenon pose significant interpretative challenges[8] which are not addressed by conventional pipelines.[9,10]

To address these challenges, researchers have increasingly employed Deep Learning (DL) models,[11] which excel at learning hierarchical representations directly from raw data; this makes them a promising tool for various applications,[12] mainly when the phenomenon explored is not fully understood. In fact, when applied to processes with subjective[6,13–16] or ambiguous labels,[6,14–18] or to data that are too contextualized, conventional evaluation pipelines may be fundamentally flawed, as they do not account for complex relationships between labels and data, obscuring their true validity.[8,19] In these cases, a validation strategy can involve splitting the data into several components such that phenomenon-specific, useful information is filtered out to check if the trained DL model continues to yield good performance, as shown in Fig. 1. If this happens, as sketched in Fig. 1(b), it means that the model is based on misleading information rather than on what it should be based on. This is the core of the *data-related ablation* paradigm we define therein.

In fact, the limitations of conventional pipelines underscore the need for a more nuanced approach to *validation*. The traditional paradigm is composed of what we call a *model-related ablation* (Fig. 2(a)), which evaluates the performance of the model by systematically modifying its components or hyperparameters,[20–25] implicitly assuming that the data are free from spurious, subtle correlations[8]; this assumption is too strong when dealing with variable, noisy data coming from partially unknown phenomena. This ablation lacks what we term "*data explainability*", defined as the evaluation of whether model decisions rely on data features that are effective for the target phenomenon.

To account for data variability and uncertainty, the proposed paradigm complements the *model-related ablation* with *data-related ablation* (Fig. 2(b)). The latter systematically perturbs the input data, allowing one to determine whether the predictions are based on meaningful and process-relevant characteristics rather than on the side effects affecting the data.[26] *Data-related ablation* precedes *model-related ablation* to ensure that the model is learning meaningful features before optimizing the model architecture.

To demonstrate the effectiveness of the proposed paradigm, as a case study, we applied it to two Electroencephalography (EEG) classification tasks: one related to the complex task of Emotion Recognition (ER),[13] the other represented by the benchmark Motor Execution task (ME).[27] The findings not only expose the vulnerabilities of current DL models but also pave the way for the development of robust and explainable AI systems that can be applied in virtually every domain, from medicine to computer vision, natural language processing and beyond.

The main contributions of this work are summarized as follows:

- The integration of the traditional *model-related ablation* with the concept of *data-related ablation* to complete the ablation pipeline. This provides an innovative paradigm for evaluating whether DL models rely on meaningful features instead of biases and artifacts. This ensures that the learned representations are based on relevant and process-related characteristics;
- A detailed case study of the pipeline in action on the EEG domain, testing the proposed method on the complex task of ER compared to the benchmark task of ME. The experiments show key weaknesses in ER of current DL methods and highlight the usefulness of the proposed paradigm;
- A discussion of the implications of the *data-related ablation* in building robust, explainable and reliable DL systems;
- The extension of the concept of explainability to go beyond the model architecture for including the data explainability in the loop;

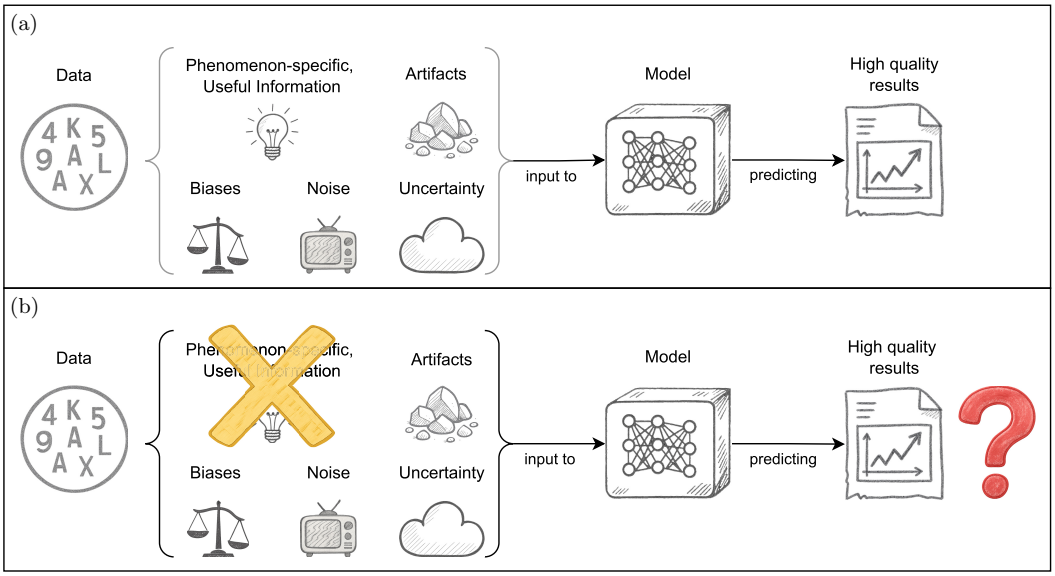

Fig. 1. The prediction pipeline before (a) and after (b) Data-related ablation. If the useful information is removed, as in (b), why is the quality of the nearly unaltered results so high? This comparison shows the main issue discussed in this work. If a model maintains high accuracy even when key information related to specific phenomena is removed (scenario (b)), it might suggest that the model depends on misleading correlations instead of genuine learning. In this context, "*Phenomenon-specific, useful information*" is defined beforehand based on established knowledge in the domain. "*Artifacts*" are spurious features affecting the domain that have the appearance of useful information but are generated outside the domain itself.

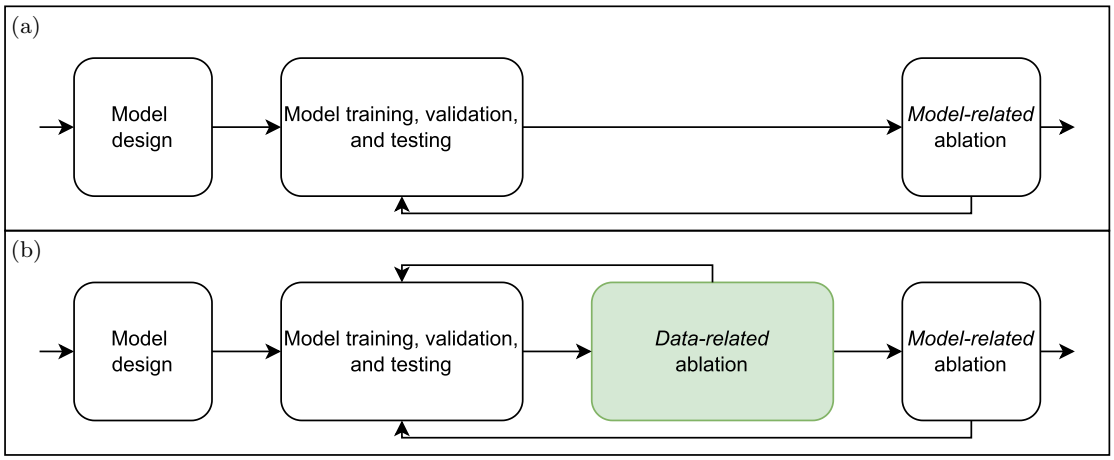

Fig. 2. Comparison between standard, *model-related ablation* pipeline (a) and the proposed integration of *data-related* and *model-related ablations* into a single paradigm (b). This structural change prioritizes checking the validity of the data source, making sure that the model is based on relevant features before significant resources are spent on improving the architecture through *model-related ablation*.

- The contribution of the data-related paradigm in explaining complex phenomena.

The rest of this paper is composed of four main sections, each focusing on a specific part of our investigation. In Sec. 2, we explain the need for *data-related ablation* as a useful tool to examine the role of different components of a signal in shaping the performance of the model.

Section 3 provides a detailed account of the methods used to conduct *data-related ablation* experiments; this section describes how we designed the ablation strategies, why we chose them, and the steps we took to systematically remove or

Int. J. Neur. Syst. 2026.36. Downloaded from www.worldscientific.com
by 79.54.49.253 on 03/02/26. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Neur. Syst. 2026.36. Downloaded from www.worldscientific.com
by 79.54.49.253 on 03/02/26. Re-use and distribution is strictly not permitted, except for Open Access articles.

manipulate specific data components, allowing us to observe the effects on model behavior.

In Sec. 4, we present the experimental testbed used to evaluate the proposed ablation pipeline. This section introduces the datasets and tasks considered within the EEG domain and shares the findings we obtained from our experiments.

Finally, Sec. 5, offers a reflective discussion on the broader implications of our pipeline, providing insights on its strengths and limitations, thereby discussing how the insights gained from data ablation can shape future research.

2. Motivation for Data-related Ablation

Understanding how to explain complex phenomena is crucial to building robust automated decisions[28,29]; however, as the complexity of processable phenomena increases and DL models become more sophisticated,[11] it can be challenging to grasp what drives the behavior of DL. To address this challenge, researchers use a technique called *ablation*, which involves modifying specific components or characteristics of the model to assess their role and importance.[20,30,31] To distinguish this traditional approach from the data-centric paradigm we introduce later, we will refer to it as *model-related ablation*. Thus, in this section, we first revisit the limitations of conventional *model-related ablation*, and then introduce the complementary idea of *data-related ablation*, showing its advantages through examples. Finally, we show how combining these two approaches gives us a more complete paradigm for understanding the behavior of the model and the observed phenomenon.

Model-related ablation can provide valuable information, but its main limitation is that it does not consider the data on which it operates.[32,33] As an example, consider a model trained to recognize chairs from other furniture, using just images of four-legged chairs made of aluminum or wood: the model's choice could be based on the four-legged shape of the chairs, the material, the color, or any other combination of them. With these assumptions, if we test the model with a four-legged wooden table, the classification could result in a chair because the model lacks features capable of distinguishing chairs from tables. Analogously, if we test the model with single-base chairs or with plastic four-legged chairs, the model would inevitably fail.

To refine this process, we propose a *data-related ablation* to complement *model-related* one. This ablation perturbs the data on which the system is trained and evaluated on; by doing so, we can assess whether the model's predictions are driven by meaningful patterns in the data or by superficial characteristics that are irrelevant to the process at hand.[8] If a model performs well with spurious features, then changes to the model architecture may just reinforce a dependence on these features instead of preferring domain-specific features. This is equivalent to testing the above model with different furniture or chairs of different types, colors and materials.

By combining *data-* and *model-related ablation*, as in Fig.2(b), it is possible to gain a more comprehensive understanding of complex phenomena and models,[9] as well as the features on which these models base their decision, improving explainability. Both kinds of ablation are fundamental: the first provides a detailed understanding of the model's internal workings, but may overlook the impact of the data on the model's behavior[33]; the second sheds light on how the model interacts with the data, but may not provide direct insight into the model's internal mechanics.[26] The synergistic combination of the two approaches allows one to develop a more complete picture of how models work in complex processes and ultimately build more robust and reliable systems that are better equipped to face real-world challenges.[30] The proposed *data-related ablation* paradigm can be successfully applied in several contexts, mainly when the phenomena explored are not clearly understood, from neuroscience to medicine, computer vision, natural language processing and more. In this specific case, we use one of the several applications: the classification of EEG signals.

Because the choice of what to perturb is closely related to the domain properties, there cannot exist a universal guideline for a *data-related ablation*. In practice, researchers are first tasked with identifying key characteristics or possible shortcuts in the specific domain, such as analyzing the role of specific frequency bands in EEG, texture cues in images, or language patterns in texts. Next, they should create

Int. J. Neur. Syst. 2026.36. Downloaded from www.worldscientific.com
by 79.54.49.253 on 03/02/26. Re-use and distribution is strictly not permitted, except for Open Access articles.

perturbations that selectively remove or modify characterizing features. By customizing the ablation to the data, researchers can better determine if the predictions of a model rely on real patterns or artifacts at the surface level, helping to ensure that the insights gained are valid and understandable for the specific application.

3. Proposed Method

The main focus of this work is to make *data-related ablation* a systematic evaluation method. In this section, we focus on describing the practical methodology for applying this approach to the ablation pipeline.

The procedure involves perturbing the input data while keeping the model architecture the same. In each ablation step, we selectively change or remove specific components of the data and retrain the model from scratch on the altered dataset. By comparing performance across these changes, we can directly measure how much the model depends on specific characteristics of the data. If the model maintains high accuracy after removing relevant features from the data, it suggests that the learned representations depend more on incidental patterns than on important, phenomenon-specific information (Fig. 1). In contrast, if performance drops with targeted changes, this indicates that the model needs and uses meaningful features. The removal of data-specific components can be done in one step or in different steps by separating the data into pieces. In this way, through *data-related ablation*, we could learn more about the phenomenon we observe and the model we use to interpret it.

A key requirement for this procedure is that changes must be carefully designed to avoid adding new confusing factors. They should keep the overall structure of the data while specifically targeting suspected sources of false correlations. Therefore, the ablation process is iterative: multiple changes are applied, assessed, and adjusted to isolate and examine different aspects of the relevance of the data. The strength of the evaluation depends on how well the perturbation functions are designed, which must be established with enough domain knowledge, being domain-specific.

3.1. *Problem formulation*

The data-related ablation procedure for a multiclass classification problem is formally described as follows. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N_d}$ be a dataset containing N_d tuples representing data samples and their corresponding labels, with c different labels. We define a set of functions $\mathcal{P} = I \cup \{P_j : \mathcal{D} \to \mathcal{D}\}_{j=1}^{N_p}$, where I is the identity function, and each P_j is a perturbation function that modifies a sample while maintaining its general characteristics. Note that $\mathcal{P}$ must contain I, which does not perform transformations, ensuring the evaluation of the unaltered dataset. Also, let $M : \mathcal{D} \to \mathbb{R}^c$ be a model architecture compatible with samples in $\mathcal{D}$ and initialized with randomized weights θ_0, and let $\mathcal{L} : \mathbb{R}^c \to [0, 1]$ represent the performance metric, where higher scores are considered to be better (e.g, accuracy or F_1-score).

Now, a training is performed for each perturbation $P \in \mathcal{P}$. At each iteration, a different version of $\mathcal{D}$ is created, namely $\mathcal{D}_P = \{(P(x_i), y_i)\}_{(x_i,y_i)\in\mathcal{D}}$. The model is therefore initialized again with θ_0 and subjected to full training on $\mathcal{D}_P$, leading to a suboptimal set of weights θ_P^*. Finally, the error Δ_P of architecture M parameterized by θ_P^* on the perturbed dataset $\mathcal{D}_P$ is defined as

$$\Delta_P = 1 - \mathcal{L}(\mathcal{M}, \theta_P^*, \mathcal{D}_P). \tag{1}$$

After all perturbations have been picked, the average error $\bar{\Delta}$ and the spread of the errors σ_Δ are computed as

$$\bar{\Delta} = \frac{\sum\limits_{P\in\mathcal{P}} \Delta_P}{|\mathcal{P}|}, \tag{2}$$

$$\sigma_\Delta = \sqrt{\frac{\sum\limits_{P\in\mathcal{P}} (\Delta_P - \bar{\Delta})^2}{|\mathcal{P}|}}. \tag{3}$$

A low spread σ_Δ suggests the model's insensitivity to the perturbations, that is, a potential reliance on dataset biases or shortcut learning; conversely, a significantly higher σ_Δ implies sensitivity to the perturbed features, suggesting the model learned more robustly. In principle, higher-order moments or information-theoretic metrics could be explored to increase sensitivity, but this would require an in-depth discussion that is beyond the scope of our work. Instead, our main contribution is

Int. J. Neur. Syst. 2026.36. Downloaded from www.worldscientific.com
by 79.54.49.253 on 03/02/26. Re-use and distribution is strictly not permitted, except for Open Access articles.

conceptual: establishing the data-related ablation paradigm, for which using commonly understood statistical measures makes the framework accessible and clear.

However, it should be noted that finding appropriate perturbations is domain-specific and, when the domain is not completely known, identifying possible sources of spurious correlations and separating them from useful components could be difficult. Moreover, biases may not originate solely from the data itself but can also stem from external factors. For example, specific architectural choices or regularization strategies could introduce biases that are not easily mitigated through data perturbations alone. Although some biases may be evident, others can be more subtle, necessitating an iterative and exploratory process to refine ablation strategies; with respect to *traditional ablation*, this inherently makes the approach more dependent on expert intuition and creativity, as there is no universal and systematic method to guarantee the removal of misleading factors.

4. Experimental Testbed

This section outlines the experimental framework used to evaluate the proposed ablation paradigm. We first describe the tasks considered in this work and the challenges they present. Next, we detail the benchmark datasets used, followed by the model architectures being compared and the preprocessing steps. Then, we introduce the specific perturbations applied to the data to test the robustness and interpretability of the model. Finally, we present the results obtained.

To test the proposed *data-related ablation* paradigm, we applied it to the real-world context of EEG classification. We are using this domain as a case study as it is complex and prone to overfitting,[14,34] defining one specific *data-related ablation* procedure. As every physical phenomenon has its proper specificities, which are reflected in the data it produces and, consequently, on the corresponding *data-related ablation* procedure, attempting to define a cross-domain (general) *data-related ablation* procedure is beyond the scope of this paper.

EEG collects the electrical activity of the brain, consisting of a highly variable, nonstationary, nonlinear and noisy signal[35,36] generated by various brain activities, most of which are not yet fully understood. It is widely used in applications that include brain-computer interfaces,[37] diagnosis of neurological disorders,[38] and monitoring of cognitive status,[39] making it an ideal test bed.

We tested seven models, with different complexities and designs: established EEG-specific architectures (EEGNet ,[23] EDPNet)[24]; simple baselines (a linear model, a two-layer multilayer perceptron called MLP) to gauge fundamental task requirements; a current state-of-the-art model for EEG ER (SATEER [14]); and a large-scale vision foundation model adapted for EEG (DINOv2).[25] These models were evaluated on two benchmarks represented by the DEAP dataset [13] for the ER task and the High Gamma dataset [27] for the ME task, both popular in their respective domains.[14,23,24] Data were pre-processed using a standard pipeline.[14,23,24] We applied two perturbations to the EEG data: a *frequency domain perturbation* and a *validation strategy perturbation*. The *frequency domain perturbation* involved retaining frequencies only <100 Hz or only ≥ 100 Hz ,[35,36] to assess whether the performance of the model relies on expected neural frequency bands (< 100 Hz) or unrelated frequency components (≥ 100 Hz). The EEG signal from healthy, awake subjects typically falls within the frequency range of 0.5–80 Hz. Broader frequency ranges, measured through highly sensitive probes, have been investigated by clinical neurophysiologists and researchers and shown to have context-specific clinical relevance occurring in pathological conditions.[40] We chose a 100 Hz cutoff to balance preserving relevant neural signals with reducing noise and computational load, a common practice in the field. The *validation strategy perturbation* compared the performance under trial-wise k-fold,[41,42] with $k = 10$, and Leave-One-Subject-Out (LOSO) cross-validations,[42] to test the generalization of the model and independence of the problem.

4.1. *Experimental details*

4.1.1. *Tasks*

The study utilizes two distinct EEG classification tasks to probe *data-related ablation paradigm* under different conditions: ER and ME. The ER task aims to classify the presumed emotional state of a subject (often along dimensions such as valence and

Int. J. Neur. Syst. 2026.36. Downloaded from www.worldscientific.com by 79.54.49.253 on 03/02/26. Re-use and distribution is strictly not permitted, except for Open Access articles.

arousal)[35,36] based on the EEG signals recorded while experiencing external stimuli, such as viewing images or videos.[13] ER involves analyzing distributed neural processes, many of which produce weak or diffuse EEG signatures.[34] This task challenges models to find reliable neural correlates for subjective internal states. In contrast, the ME task involves classifying specific physical movements performed by the subject (e.g. moving the left hand versus the right foot)[27] using the EEG recorded during the action.

A critical distinction between ER and ME lies in their associated labels. ER labels are inherently subjective, typically derived from participant self-reports (e.g. rating scales for valence and arousal after viewing stimuli) or inferred based on the presumed emotional impact of those stimuli (e.g. assuming a scary video induces fear).[13] Crucially, there is no direct external method to verify the ground truth of these internal feeling states; self-reports can be inaccurate, inconsistent, or influenced by factors beyond the stimulus, and stimuli may not reliably evoke the same emotion across individuals or even within the same individual over time. Consequently, the EEG signals in the ER are mapped to potentially noisy or imprecise representations of the subject's actual emotional experience. In contrast, ME labels are objective, since they correspond to specific, observable physical actions executed following explicit instructions. The ground truth, which corresponds to the intended or executed movement, is thus directly verifiable and unambiguous. This fundamental difference in label certainty significantly impacts the challenge posed to DL models: ER requires learning correlations with potentially unreliable indicators of internal states, while ME involves mapping signals to clearly defined, verifiable external events.

4.1.2. *Benchmark datasets*

We evaluated the performance and robustness of the model using two distinct and widely adopted EEG benchmarks, chosen to represent tasks with fundamentally different ground truth characteristics. Here, the two benchmarks are described in detail.

The first benchmark, DEAP (A Dataset for Emotion Analysis using EEG, physiological, and video signals),[13] was designed specifically to analyze affective states of humans. It contains recordings from 32 participants as they watched 40 carefully selected 1-min music video excerpts intended to elicit a range of emotions. For each participant, a uniform 32-channel scalp EEG and various peripheral physiological signals were recorded. After viewing each video, participants provided self-assessments on a continuous scale for Valence, Arousal, Dominance, and Liking.[35,36] In our study, we used the DEAP dataset for subjective ER and identification tasks.

The second benchmark, High Gamma,[27] originates from studies focused on decoding motor intentions and the execution of EEG. It includes recordings from 14 subjects performing blocks of cued ME tasks. Specifically, subjects performed one of four movements: left hand, right hand, both feet, or rest, typically for around 4 s per test after a visual cue. Contextually, uniform 128-channel EEG signals were recorded at 1000 hZ. The ground truth labels correspond directly to the instructed and executed physical movement. We employ this dataset for the objective ME classification task.

The two datasets used in this study, DEAP[13] and High Gamma,[27] can be downloaded from https://www.eecs.qmul.ac.uk/mmv/datasets/deap and https://gin.g-node.org/robintibor/high-gamma-dataset, respectively. The Python code used in this study is available at https://github.com/rom42pla/data_related_ablation.

4.1.3. *Evaluation metrics and performance validation*

To assess the performance of model classification, we focus on the F_1-score, which is defined as the harmonic mean of Precision and Recall. It provides a single metric that balances both and makes it especially useful for imbalanced datasets, and penalizes models that do well on the majority class but poorly on the minority class. In particular, we used the macro-averaged F_1-score, which calculates the score for each class independently and then takes the unweighted average, treating all classes equally regardless of their size. The F_1-score for a single class is given by

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (6)$$

The macro-averaged F_1 is then the average of the F_1-scores calculated for each class.

To complement the descriptive statistics of the F_1-score, we performed one-way analysis of variance (ANOVA) tests[43] to formally assess whether the observed differences in performance between folds and frequency ranges were statistically significant. ANOVA evaluates whether the variation between groups (like different folds) is larger than the variation within groups, producing a statistic F that quantifies the ratio of these variances. A higher F statistic indicates that the differences between groups are large relative to the variability within the groups. The p-value measures the probability that the observed differences could have arisen by chance; a small p-value (we considered below 0.001) suggests that the differences are unlikely to have arisen by chance, providing evidence that the performance differences are statistically meaningful. Conducting ANOVA allows us to confirm which models exhibit stable performance in all folds and which show significant variability, thus supporting the robustness of the reported score F_1.

4.1.4. *Implementation details*

All experiments were carried out on a workstation that had an NVIDIA RTX 4090 GPU, an Intel Core i9-14900K CPU, and 128 GB of RAM. We used PyTorch 2.6 for model training and evaluation.

For ensuring a fair comparison between different architectures, we provided consistent conditions across experiments, regarding both hyperparameter choices and optimization strategies. Regarding hyperparameters, we used the best combination provided by the authors in their original papers to fully exploit the specific capabilities of each architecture. Regarding optimization, we used the AdamW optimizer[44] with a weight decay of 0.01. We selected learning rates based on the model architecture: for non-Transformer-based models, we set a learning rate of 5×10^{-3}. For transformer-based models, we used a lower learning rate of 5×10^{-5} to ensure stable convergence. All models were trained with a batch size of 128 and a maximum of 30 epochs, without applying an early stopping policy.

4.1.5. *Preprocessing*

Before feeding the EEG waveforms to the network, we process them using a standard pipeline to ensure consistency and improve model learning. First, we normalize the raw signals to have a zero mean and unit variance and then scale their values to fit within the range $[-1, 1]$. This scaling helps to achieve a faster and more stable convergence.[45] After that, we segment each waveform into fixed-length nonoverlapping 2 s windows using a sliding window method, standardizing the sequence length. Each window is then transformed into a Mel spectrogram[46] using m Mel filter banks. This representation captures nonlinear, nonstationary brain activity more effectively than linear spectrograms.[14] It also provides a compact and computationally efficient representation.[47]

4.2. *Results*

The results of the experiments, evaluated using the F_1-score metric, which balances precision (the proportion of true positives among positive predictions) and recall (the proportion of true positives correctly identified), are presented in Tables 1 and 2. First, we present the results for the ER task on the DEAP dataset,[13] separately for the four different emotion labels (Table 1). The results show that the F_1-score for most models remains relatively stable across the two frequency ranges under both validation schemes. The performance of the models remains stable on the different emotion labels. For example, DINOv2[25] achieves a stable F_1-score of 99.7% and beyond for both frequency ranges under k-fold validation, which is peculiar considering that we filter out useful information related to brain activity. This work aims to debunk unrealistic results: in this case, 99% accuracy is clearly an unreliable result, but in some cases, also typical, biologically plausible values between 70% and 85% could be false and more difficult to discover since they resemble realistic values. Our analysis highlights that even with improved models, we must carefully interpret the results, grounded in a solid understanding of the validation methods used, as artificial inflation is still a major concern with poor validation practices.

To confirm the significance of the performance differences seen in the folds and frequency ranges, we conducted one-way ANOVA tests[43] for all models

Int. J. Neur. Syst. 2026.36. Downloaded from www.worldscientific.com
by 79.54.49.253 on 03/02/26. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 1. Classification results for the ER task on the DEAP[13] benchmark, with individual results for each of the four labels: Valence, Arousal, Dominance, and Liking.

Validation	Model	Valence F_1 (%) ↑		Arousal F_1 (%) ↑	
		<100 Hz	≥ 100 Hz	<100 Hz	≥ 100 Hz
k-fold	DINOv2[25]	99.7 ± 0.06	99.8 ± 0.10	99.7 ± 0.13	99.8 ± 0.12
	EDPNet[24]	69.5 ± 1.89	73.9 ± 0.84	64.6 ± 2.20	70.4 ± 1.56
	EEGNet[23]	65.3 ± 1.68	67.0 ± 1.20	67.7 ± 1.52	68.2 ± 0.79
	Linear	66.5 ± 1.37	66.8 ± 1.74	71.5 ± 0.84	70.9 ± 1.05
	MLP	72.7 ± 2.49	72.0 ± 2.64	76.4 ± 2.63	76.0 ± 2.50
	SATEER[14]	93.4 ± 1.04	91.7 ± 0.86	93.9 ± 0.75	92.0 ± 0.91
LOSO	DINOv2[25]	52.5 ± 14.86	51.2 ± 16.26	56.0 ± 20.10	57.3 ± 19.22
	EDPNet[24]	61.9 ± 20.09	58.6 ± 16.64	59.2 ± 12.70	55.1 ± 12.65
	EEGNet[23]	58.3 ± 13.26	50.8 ± 12.89	62.0 ± 19.38	53.3 ± 22.90
	Linear	56.3 ± 16.13	39.9 ± 24.27	61.5 ± 17.19	46.5 ± 25.58
	MLP	55.0 ± 18.93	51.9 ± 17.39	58.5 ± 19.15	54.6 ± 24.35
	SATEER[14]	53.8 ± 17.30	49.0 ± 22.35	58.4 ± 20.47	55.1 ± 20.54
Validation	Model	Dominance F_1 (%) ↑		Liking F_1 (%) ↑	
		<100 Hz	≥ 100 Hz	<100 Hz	≥ 100 Hz
k-fold	DINOv2[25]	99.8 ± 0.10	99.8 ± 0.12	99.8 ± 0.07	99.9 ± 0.08
	EDPNet[24]	68.7 ± 1.61	72.7 ± 1.27	78.9 ± 0.86	81.3 ± 1.10
	EEGNet[23]	68.0 ± 0.87	68.6 ± 0.76	80.4 ± 0.89	80.1 ± 0.78
	Linear	70.0 ± 0.89	70.5 ± 1.03	80.3 ± 1.02	80.7 ± 1.12
	MLP	76.0 ± 2.41	75.9 ± 2.49	81.9 ± 1.08	83.6 ± 0.81
	SATEER[14]	93.6 ± 0.74	92.3 ± 0.91	95.8 ± 0.76	93.4 ± 1.32
LOSO	DINOv2[25]	57.8 ± 17.14	52.0 ± 19.49	66.5 ± 12.99	68.8 ± 13.21
	EDPNet[24]	60.5 ± 18.76	61.5 ± 13.23	76.4 ± 10.05	71.0 ± 17.66
	EEGNet[23]	55.4 ± 12.90	49.2 ± 21.16	72.3 ± 10.23	67.0 ± 15.08
	Linear	56.6 ± 17.59	40.3 ± 24.55	76.5 ± 9.88	67.8 ± 17.49
	MLP	58.9 ± 20.88	53.4 ± 20.26	72.4 ± 11.04	74.7 ± 12.67
	SATEER[14]	60.0 ± 16.15	53.7 ± 20.82	72.4 ± 9.06	72.2 ± 13.44

Notes: Values are reported as mean ± standard deviation. The columns compare the performance achieved using only the physiological frequency band (< 100 Hz) and the high-frequency band (≥ 100 Hz).

and signal types. The results, summarized in Table 3, show that most models, such as EEGNet, Linear, and EDPNet, have large F statistics and extremely low p values ($p \ll 0.001$). This confirms that the differences across folds are significant and unlikely to occur by chance. In contrast, models such as DINOv2 display low F statistics and nonsignificant p values in some cases. This reflects the very low variance of their performance across folds. Thus, DINOv2 demonstrates highly stable performance, while the variability in other models is significant and noticeable. In general, these results support the strong performance patterns of the models, as reported by means of mean scores F_1.

Using the same data-related ablation procedure, we repeated the experiments for the ME task on the High Gamma dataset[27] (Table 2). The results show that the performance of the models decreases significantly when the frequency range is ≥ 100 Hz. For example, DINOv2's F_1-score drastically drops from 71.5% to 29.2% between the two frequency ranges. This confirms that when the phenomenon shows intense specific characteristics, such as in ME, the characteristics recognized by the models come from the activity of the brain cortex related to the task, where EEG has its maximum sensitivity.[34] In contrast, for weak and deep signals, such as those generated for ER, the features used by the model come from nonneuronal activity.

Int. J. Neur. Syst. 2026.36. Downloaded from www.worldscientific.com
by 79.54.49.253 on 03/02/26. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Neur. Syst. 2026.36. Downloaded from www.worldscientific.com
by 79.54.49.253 on 03/02/26. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 2. Classification results for ME on the High Gamma[27] benchmark.

Validation	Model	Motor execution F_1 (%) ↑ < 100 Hz	≥ 100 Hz
k-fold	DINOv2[25]	71.5 ± 1.66	29.2 ± 0.77
	EDPNet[24]	70.8 ± 1.44	28.5 ± 1.65
	EEGNet[23]	54.9 ± 1.34	26.5 ± 1.08
	Linear	59.5 ± 1.39	27.2 ± 0.49
	MLP	67.6 ± 1.82	28.9 ± 0.58
	SATEER[14]	64.5 ± 1.36	26.5 ± 0.58
LOSO	DINOv2[25]	48.1 ± 9.66	26.9 ± 1.06
	EDPNet[24]	50.2 ± 7.09	27.4 ± 1.62
	EEGNet[23]	48.5 ± 6.41	26.9 ± 1.07
	Linear	45.2 ± 8.12	27.4 ± 1.03
	MLP	45.6 ± 10.2	26.9 ± 0.85
	SATEER[14]	43.4 ± 9.94	25.2 ± 0.30

Notes: Values are reported as mean ± standard deviation. The columns compare the performance achieved using only the physiological frequency band (< 100 Hz) and the high-frequency band (≥ 100 Hz).

Table 3. One-way ANOVA results for all models across k-fold and LOSO validation splits on the DEAP[13] benchmark.

Validation	Model	ANOVA F statistic < 100 Hz	≥ 100 Hz	ANOVA p-value < 100 Hz	≥ 100 Hz
k-fold	DINOv2[25]	0.602	0.758	6.182×10^{-1}	5.252×10^{-1}
	EDPNet[24]	123.691	149.525	$p \ll 0.001$	$p \ll 0.001$
	EEGNet[23]	397.922	423.285	$p \ll 0.001$	$p \ll 0.001$
	Linear	132.111	214.701	$p \ll 0.001$	$p \ll 0.001$
	MLP	39.970	53.574	$p \ll 0.001$	$p \ll 0.001$
	SATEER[14]	22.681	7.686	$p \ll 0.001$	$p \ll 0.001$
LOSO	DINOv2[25]	3.216	4.498	2.668×10^{-2}	5.527×10^{-3}
	EDPNet[24]	5.848	4.592	1.089×10^{-3}	4.928×10^{-3}
	EEGNet[23]	6.620	4.755	$p \ll 0.001$	4.041×10^{-3}
	Linear	6.569	10.570	$p \ll 0.001$	$p \ll 0.001$
	MLP	4.371	3.647	6.446×10^{-3}	1.566×10^{-2}
	SATEER[14]	4.914	11.206	3.332×10^{-3}	$p \ll 0.001$

Notes: F-statistics and p-values indicate whether fold-to-fold performance differences are statistically significant. High F and low p-values correspond to meaningful variation across folds, while low F and high p-values reflect highly stable model performance. For p, values higher than 0.001 are fully reported. The analysis is divided into the two considered frequency bands to show how input bandwidth impacts the statistical stability of the predictions across different validation folds.

4.3. *Discussion on anomalous performance*

While most models show consistent trends across frequency ranges and validation methods, some unusual behaviors warrant attention. In particular, DINOv2 has notably high and stable F_1-scores under k-fold validation for the DEAP dataset, achieving almost perfect performance for all emotion labels. This stability is evident in the ANOVA results, where DINOv2 shows low F-statistics and high p-values, indicating little variation across folds.

Int. J. Neur. Syst. 2026.36. Downloaded from www.worldscientific.com
by 79.54.49.253 on 03/02/26. Re-use and distribution is strictly not permitted, except for Open Access articles.

This performance is in contrast to other models, which exhibit significant variability in folds and frequency ranges, as shown by large F statistics and very low p-values. These anomalies may be partly due to DINOv2's design and pre-training, which might enable the model to identify features that do not respond to the frequency changes used in our data-related ablation procedure. Although this leads to consistently strong performance for weak EEG signals in the ER task, it also suggests that DINOv2 may depend on misleading or nonneural patterns in the data, a trend that is less noticeable in other models.

Moreover, under LOSO validation, all models display much higher variability, with larger standard deviations in F_1-scores. This shows that subject-specific differences heavily influence performance when the model lacks access to data from the test subject during training. The difference between k-fold and LOSO emphasizes that the apparent stability of some models under k-fold validation may hide sensitivities to individual traits of the subject.

Finally, for the ME task on the High Gamma dataset, all models experience a sharp decrease in performance for frequencies ≥ 100 Hz. This reinforces the idea that the models are mainly responsive to neural activity in the lower frequency bands. DINOv2's unexpectedly high performance in ER tasks does not carry over to ME tasks, supporting the theory that its strong ER results may stem from capturing nonneural or dataset-specific characteristics instead of domain-related brain activity.

4.4. *Discussion*

The proposed *data-related ablation* paradigm shows that models for ER maintain high performance despite the removal of frequency components, suggesting the dependence on nonneuronal characteristics or artifacts over domain-specific neural activity. In contrast, models for ME decrease significantly when relying only on high-frequency components, indicating a dependence on lower-frequency bands crucial for motor signals. Disparities between k-fold and LOSO validation, with LOSO performance drops, highlight model limitations and overfitting to subject-specific artifacts.[41,42] These findings strongly suggest potential biases in the DEAP dataset or in the formulation of the ER task.[33] The insensitivity of ER to *data-related ablation* and substantial drops in LOSO performance indicate overfitting to subject-specific patterns or nongeneralizable neural correlates, which requires more rigorous exploration.[48]

Since the paradigm deliberately separates and removes some features, it simultaneously: validates the trustworthiness of the model, confirming that the predictions are genuinely derived from relevant signals; provides an explanation window by revealing which retained features drive the final decision. In this way, the proposed paradigm could contribute to explaining the domain to which it is applied and to better explain which parts of the DL models are most sensitive to data variability. However, we must recognize that our experiments were intentionally limited to certain spectral and validation-based changes to act as a proof-of-concept. As a result, this study does not claim to identify all potential sources of misleading correlations: there may still be biases related to spatial patterns, smaller temporal details, or other specific artifacts not captured therein. Our objective was to validate the methodology itself rather than to provide a universal, comprehensive audit of EEG models, which are intentionally left for future work.

5. Conclusion

In this work, we have proposed the introduction of a *data-related ablation* paradigm in the main ablation pipeline to reduce the possibility that a DL model could be trained on spurious features instead of phenomenon-specific features. Going forward, the word "*model*" should also include data related to the phenomenon they capture. When applied in the exemplary field of EEG, our paradigm reveals significant concerns about the validity of the DL model, particularly the ER tasks. Due to its complexity, EEG served as a testbed to demonstrate the importance of *data-related ablation*, and not to decrease DL based on EEG. These results have broad implications, and future work should assess generalizability across diverse phenomena and modalities, exploring granular ablations for specific applications[49–52] and other domains, such as fMRI, wearable sensors, and natural images. Furthermore, to add more depth and rigor to this approach, future experiments could shift from broad ablations, as we have done, to more specific and targeted

manipulations, like systematically removing frequency sub-bands or examining the impact of individual channel inputs. Other factors to investigate may be related to which parts of the network are more sensitive to *data-related ablation*.

In conclusion, this paper alerts the DL community that high performance alone is insufficient; models must also be reliable, robust, and grounded in the phenomena they capture. The proposed *data-related ablation* paradigm has far-reaching implications: it allows us to observe how models behave on different domains, to judge whether the data at hand are adequate for conducting a reliable analysis, and to uncover the underlying causes of potential flaws of certain models on particular domains or applications. Moreover, without *data-related ablation*, there is a serious risk of launching AI systems that reinforce biases or rely on false correlations. Making sure models depend on valid, causal features is not just a technical requirement; it is a fundamental ethical obligation, especially in high-stakes areas like healthcare. Therefore, we suggest that scientific journals and conferences require *data-related ablation* verification for DL models involving complex or unexplained phenomena. We urge a nuanced and rigorous evaluation approach that prioritizes explainability, generalizability, and validity. This will enable models that truly capture the underlying mechanisms, unlocking the full scientific potential of DL.[53]

Acknowledgments

This research was partially supported by:

- The European Union NextGenerationEU under the Italian Ministry of University and Research (MUR) National Innovation Ecosystem Grant No. ECS00000041 — VITALITY CUP E13C22001 060006;
- The project D^{3}4Health — Digital Driven Diagnostics, Prognostics and Therapeutics for Sustainable Health Care, Project PNC0000001 — CUP B53C22006090001, funded by the European Union — NextGenerationEU under the Italian National Plan for Complementary Investments to the NRRP.
- Strategic Departmental Plan on Artificial Intelligence, Department of Mathematics, Computer Science and Physics, University of Udine.
- The FVG Project "Supporting the diagnosis of rare diseases (MR) through artificial intelligence" (2023-26) (Project A with CUP: F53C22001770002 and Project B with CUP F53C22001780002).

ORCID

Romeo Lanzino https://orcid.org/0000-0003-2939-3007
Luigi Cinque https://orcid.org/0000-0001-9149-2175
Gian Luca Foresti https://orcid.org/0000-0002-8425-6892
Giuseppe Placidi https://orcid.org/0000-0002-4790-4029

References

1. M. I. Jordan and T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, *Science* **349** (6245) (2015) 255–260.
2. Y. Xue, Y. Lin and F. Neri, Architecture knowledge distillation for evolutionary generative adversarial network, *Int. J. Neural Syst.* **35**(04) (2025) 2550013.
3. Z. Xie, J. Lian and D. Wang, Enhanced graph attention network by integrating transformer for epileptic EEG identification, *Int. J. Neural Syst.* **35** (08) (2025) 2550037.
4. A. Hassanpour, M. Moradikia, H. Adeli, S. R. Khayami and P. Shamsinejadbabaki, A novel end-to-end deep learning scheme for classifying multi-class motor imagery electroencephalography signals, *Expert Syst.* **36**(6) (2019) e12494.
5. H. S. Nogay and H. Adeli, Detection of epileptic seizure using pretrained deep convolutional neural network and transfer learning, *Eur. Neurol.* **83** (2021) 602–614.
6. R. Yuvaraj, M. Murugappan, U. R. Acharya, H. Adeli, N. M. Ibrahim and E. Mesquita, Brain functional connectivity patterns for emotional state classification in Parkinson's disease patients without dementia, *Behav. Brain Res.* **298** (2016) 248–260.
7. D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel and D. Hassabis, Mastering the game of go without human knowledge, *Nature* **550** (2017) 354–359.
8. R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge and F. A. Wichmann, Shortcut learning in deep neural networks, *Nat. Mach. Intell.* **2** (2020) 665–673.

Int. J. Neur. Syst. 2026.36. Downloaded from www.worldscientific.com by 79.54.49.253 on 03/02/26. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Neur. Syst. 2026.36. Downloaded from www.worldscientific.com
by 79.54.49.253 on 03/02/26. Re-use and distribution is strictly not permitted, except for Open Access articles.

9. A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, F. Hormozdiari, N. Houlsby, S. Hou, G. Jerfel, A. Karthikesalingam, M. Lucic, Y. Ma, C. McLean, D. Mincu, A. Mitani, A. Montanari, Z. Nado, V. Natarajan, C. Nielson, T. F. Osborne, R. Raman, K. Ramasamy, R. Sayres, J. Schrouff, M. Seneviratne, S. Sequeira, H. Suresh, V. Veitch, M. Vladymyrov, X. Wang, K. Webster, S. Yadlowsky, T. Yun, X. Zhai and D. Sculley, Underspecification presents challenges for credibility in modern machine learning, *J. Mach. Learn. Res.* **23** (2022) 1–61.
10. B. Recht, R. Roelofs, L. Schmidt and V. Shankar, Do imagenet classifiers generalize to imagenet?, in *Proc. 36th Int. Conf. Machine Learning*, eds. K. Chaudhuri and R. Salakhutdinov, Proceedings of Machine Learning Research, Vol. 97 (Proceedings of Machine Learning Research, 2019), pp. 5389–5400.
11. Y. LeCun, Y. Bengio and G. Hinton, Deep learning, *Nature* **521** (2015) 436–444.
12. C. Ahuja and D. Sethia, SS-emerge — self-supervised enhancement for multidimension emotion recognition using GNNS for EEG, *Sci. Rep.* **15** (2025) 14254.
13. S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt and I. Patras, Deap: A database for emotion analysis using physiological signals, *IEEE Trans. Affect. Comput.* **3** (1) (2012) 18–31.
14. R. Lanzino, D. Avola, F. Fontana, L. Cinque, F. Scarcello and G. L. Foresti, Sateer: Subject-aware transformer for EEG-based emotion recognition, *Int. J. Neural Syst.* **35**(02) (2025) 2550002.
15. X. Ping and W. Huang, Interactive EEG emotion recognition with incremental Gaussian processes, *Int. J. Neural Syst.* **35**(09) (2025) 2550041.
16. A. Olamat, P. Ozel and S. Atasever, Deep learning methods for multi-channel EEG-based emotion recognition, *Int. J. Neural Syst.* **32**(05) (2022) 2250021.
17. B. Frenay and M. Verleysen, Classification in the presence of label noise: A survey, *IEEE Trans. Neural Netw. Learn. Syst.* **25**(5) (2014) 845–869.
18. G. Placidi, L. Cinque, G. L. Foresti, F. Galassi, F. Mignosi, M. Nappi and M. Polsinelli, A context-dependent CNN-based framework for multiple sclerosis segmentation in MRI, *Int. J. Neural Syst.* **35**(03) (2025) 2550006.
19. J. White and S. D. Power, K-fold cross-validation can significantly over-estimate true classification accuracy in common EEG-based passive BCI experimental designs: An empirical investigation, *Sensors* **23** (2023) 6077.
20. K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. Computer Vision and pattern Recognition* (IEEE, 2016), pp. 770–778.
21. D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, A. Plaza, P. Gamba, J. A. Benediktsson and J. Chanussot, SpectralGPT: Spectral remote sensing foundation model, *IEEE Trans. Pattern Anal. Mach. Intell.* **46**(8) (2024) 5227–5244.
22. S. Ren, K. He, R. Girshick and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6) (2017) 1137–1149.
23. 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, *J. Neural Eng.* **15** (2018) 056013.
24. C. Han, C. Liu, J. Wang, Y. Wang, C. Cai and D. Qian, A spatial-spectral and temporal dual prototype network for motor imagery brain-computer interface, *Knowl. Based Syst.* **315** (2025) 113315.
25. M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin and P. Bojanowski, DINOv2: Learning robust visual features without supervision, *Trans. Mach. Learn. Res.* **1** (2024) 1–32.
26. D. Hendrycks and T. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations, in *Proc. Int. Conf. Learning Representations*, New Orleans (USA) (2019), pp. 1–16.
27. R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard and T. Ball, Deep learning with convolutional neural networks for EEG decoding and visualization, *Hum. Brain Mapp.* **38** (2017) 5391–5420.
28. C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nat. Mach. Intell.* **1** (2019) 206–215.
29. M. T. Ribeiro, S. Singh and C. Guestrin, "Why should i trust you?": Explaining the predictions of any classifier, in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY, 2016), pp. 1135–1144.
30. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, Vol. 31 (Curran Associates, 2017), pp. 6000–6010.
31. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran,

Int. J. Neur. Syst. 2026.36. Downloaded from www.worldscientific.com by 79.54.49.253 on 03/02/26. Re-use and distribution is strictly not permitted, except for Open Access articles.

D. Wierstra, S. Legg and D. Hassabis, Human-level control through deep reinforcement learning, *Nature* **518** (2015) 529–533.
32. J. Byrd and Z. Lipton, What is the effect of importance weighting in deep learning?, in *Proc. 36th Int. Conf. Machine Learning*, eds. K. Chaudhuri and R. Salakhutdinov, Proceedings of Machine Learning Research, Vol. 97 (Proceedings of Machine Learning Research, 2019), pp. 872–881.
33. A. Torralba and A. A. Efros, Unbiased look at dataset bias, in *Proc. IEEE Conf. Computer Vision Pattern Recognition* (IEEE, 2011), pp. 1521–1528.
34. D. Lozzi, E. Mattei, R. Ciuffini, A. Di Matteo, A. Marrelli, R. Ornello, M. Polsinelli, C. Rosignoli, S. Sacco and G. Placidi, The challenge of driving BCI with emotional signals collected by EEG, in *Proc. Graz Brain-Computer Interface Conf.* (Verlag der Technischen Universitat Graz, 2024), pp. 366–371.
35. S. M. Alarco and M. J. Fonseca, Emotions recognition using EEG signals: A survey, *IEEE Trans. Affect. Comput.* **10**(3) (2019) 374–393.
36. E. P. Torres, E. A. Torres, M. Hernandez-Alvarez and S. G. Yoo, EEG-based BCI emotion recognition: A survey, *Sensors* **20** (2020) 5083.
37. Y. Li, C. Guan, H. Li and Z. Chin, A self-training semi-supervised SVM algorithm and its application in an EEG-based brain computer interface speller system, *Pattern Recognit. Lett.* **29**(9) (2008) 1285–1294.
38. R. Ramos-Aguilar, J. A. Olvera-Lpez, I. Olmos-Pineda and S. Snchez-Urrieta, Feature extraction from EEG spectrograms for epileptic seizure detection, *Pattern Recognit. Lett.* **133** (2020) 202–209.
39. R. G. Hefron, B. J. Borghetti, J. C. Christensen and C. M. S. Kabban, Deep long short-term memory structures model temporal dependencies improving cognitive workload estimation, *Pattern Recognit. Lett.* **94** (2017) 96–104.
40. A. Bonaccini Calia, E. Masvidal-Codina, T. M. Smith, N. Schäfer, D. Rathore, E. Rodríguez-Lucas, X. Illa, J. M. De la Cruz, E. Del Corro, E. Prats-Alfonso, D. Viana, J. Bousquet, C. Hébert, J. Martínez-Aguilar, J. R. Sperling, M. Drummond, A. Halder, A. Dodd, K. Barr, S. Savage, J. Fornell, J. Sort, C. Guger, R. Villa, K. Kostarelos, R. C. Wykes, A. Guimerà-Brunet and J. A. Garrido, Full-bandwidth electrophysiology of seizures and epileptiform activity enabled by flexible graphene microtransistor depth neural probes, *Nat. Nanotechnol.* **17** (2022) 301–309.
41. M. Stone, Cross-validatory choice and assessment of statistical predictions, *J. R. Stat. Soc. B* **36**(2) (1974) 111–147.
42. T.-T. Wong, Performance evaluation of classification algorithms by K-fold and leave-one-out cross validation, *Pattern Recognit.* **48**(9) (2015) 2839–2846.
43. L. Sthle and S. Wold, Analysis of variance (Anova), *Chemometr. Intell. Lab. Syst.* **6**(4) (1989) 259–272.
44. I. Loshchilov and F. Hutter, Decoupled weight decay regularization, in *7th Int. Conf. Learning Representations*, New Orleans, LA (2019), pp. 1–18.
45. Y. LeCun, L. Bottou, G. Orr and K. Muller, Efficient backprop, in *Neural Networks: Tricks of the Trade*, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 1998), p. 546.
46. S. S. Stevens, J. Volkmann and E. B. Newman, A scale for the measurement of the psychological magnitude pitch, *J. Acoust. Soc. Am.* **8**(3) (1937) 185–190.
47. B. García-Martínez, A. Martínez-Rodrigo, R. Alcaraz and A. Fernández-Caballero, A review on nonlinear methods using electroencephalographic recordings for emotion recognition, *IEEE Trans. Affect. Comput.* **12** (3) (2021) 801–820.
48. C. Zhang, S. Bengio, M. Hardt, B. Recht and O. Vinyals, Understanding deep learning (still) requires rethinking generalization, *Commun. ACM* **64** (2021) 107–115.
49. A. Petracca, M. Carrieri, D. Avola, S. Basso Moro, S. Brigadoi, S. Lancia, M. Spezialetti, M. Ferrari, V. Quaresima and G. Placidi, A virtual ball task driven by forearm movements for neuro-rehabilitation, *Proc. Int. Conf. Virtual Rehabilitation* (IEEE, 2015), pp. 162–163.
50. W.-L. Zheng, H.-T. Guo and B.-L. Lu, Revealing critical channels and frequency bands for emotion recognition from EEG with deep belief network, in *Proc. Int. IEEE/EMBS Conf. Neural Engineering (NER)* (IEEE, 2015), pp. 154–157.
51. M. Spezialetti, L. Cinque, J. M. R. S. Tavares and G. Placidi, Towards EEG-based BCI driven by emotions for addressing BCI-illiteracy: A meta-analytic review, *Behav. Inf. Technol.* **37** (2018) 855–871.
52. E. Marox, H. Rodriguez, G. Yanez, J. Bernal, M. Rodriguez, T. Fernandez, J. Silva, A. Reyes and V. Guerrero, Broad band spectral measurements of EEG during emotional tasks, *Int. J. Neurosci.* **108** (3–4) (2001) 251–279.
53. L. Messeri and M. J. Crockett, Artificial intelligence and illusions of understanding in scientific research, *Nature* **627** (2024) 49–58.