The More, the Better? Evaluating the Role of EEG Preprocessing for Deep Learning Applications

Federico Del Pup, Andrea Zanola, Louis Fabrice Tshimanga, Alessandra Bertoldo, and Manfredo Atzori

***Abstract*— The last decade has witnessed a notable surge in deep learning applications for electroencephalography (EEG) data analysis, showing promising improvements over conventional statistical techniques. However, deep learning models can underperform if trained with bad processed data. Preprocessing is crucial for EEG data analysis, yet there is no consensus on the optimal strategies in deep learning scenarios, leading to uncertainty about the extent of preprocessing required for optimal results. This study is the first to thoroughly investigate the effects of EEG preprocessing in deep learning applications, drafting guidelines for future research. It evaluates the effects of varying preprocessing levels, from raw and minimally filtered data to complex pipelines with automated artifact removal algorithms. Six classification tasks (eye blinking, motor imagery, Parkinson's, Alzheimer's disease, sleep deprivation, and first episode psychosis) and four established EEG architectures were considered for the evaluation. The analysis of 4800 trained models revealed statistical differences between preprocessing pipelines at the intra-task level for each model and at the inter-task level for the largest model. Models trained on raw data consistently performed poorly, always ranking last in average scores. In addition, models seem to benefit more from minimal pipelines without artifact handling methods. These findings suggest that EEG artifacts may affect the performance and generalizability of deep neural networks.**

***Index Terms*— EEG, preprocessing, deep learning, motor imagery, disease classification.**

Received 29 July 2024; revised 14 January 2025 and 20 February 2025; accepted 27 February 2025. Date of publication 3 March 2025; date of current version 10 March 2025. This work was supported in part by European Union's Horizon Europe Research and Innovation Programme through HEREDITARY under Grant 101137074; and in part by the Supporting Talent in Research@University of Padua (STARS@UNIPD) Funding Program of the University of Padova, Italy, through the Project Multimodal Exploration of Data for Medical Analytics and Knowledge Extraction (MEDMAX). *(Federico Del Pup and Andrea Zanola are co-first authors.) (Corresponding author: Federico Del Pup.)*

Federico Del Pup and Alessandra Bertoldo are with the Department of Information Engineering, University of Padua, 35131 Padua, Italy, and also with Padova Neuroscience Center, University of Padua, 35129 Padua, Italy (e-mail: federico.delpup@phd.unipd.it; alessandra.bertoldo@unipd.it).

Andrea Zanola and Louis Fabrice Tshimanga are with Padova Neuroscience Center, University of Padova, 35129 Padua, Italy, and also with the Department of Neuroscience, University of Padua, 35121 Padua, Italy (e-mail: andrea.zanola@phd.unipd.it; louisfabrice.tshimanga@unipd.it).

Manfredo Atzori is with Padova Neuroscience Center, University of Padova, 35129 Padua, Italy, also with the Department of Neuroscience, University of Padua, 35121 Padua, Italy, and also with the Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO Valais-Wallis), 3960 Sierre, Switzerland (e-mail: manfredo.atzori@unipd.it).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2025.3547616, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2025.3547616

I. Introduction

ELECTROENCEPHALOGRAPHY (EEG) is a non-invasive functional imaging technique widely used to examine neural activity. It is effectively employed in a multitude of domains, including Brain-Computer Interfaces (BCI) [1], Emotion Recognition [2], Sleep Staging [3], and the study of various diseases such as epilepsy [4], Parkinson's [5], and Alzheimer's [6]. Due to their acquisition modality, EEG records have a low signal-to-noise ratio (SNR) caused by a combination of biological and non-biological artifacts (e.g., blinking, subject movement, line noise, bad electrode placement) that can degrade the quality of the signal itself. Furthermore, neural activity is characterized by highly non-linear patterns. Conventional statistical techniques and even classical machine learning approaches may be insufficient for capturing and modeling such complexity.

To enhance the analysis of EEG data, Deep Learning (DL) has emerged as a new powerful resource in the domain [7]. As a subset of machine learning, deep learning focuses on training artificial neural networks, biologically inspired systems composed of multiple interconnected layers that perform both linear and nonlinear operations on the input data, typically a portion of a multichannel EEG recording. Such models have proven capable of discovering complex relationships by learning representations of data with multiple levels of abstraction [8]. However, the increased learning capability is achieved at the cost of a higher computation and number of parameters to optimize. As a result, the demand for larger EEG repositories has increased significantly, leading to the release of numerous open-source datasets [9]. If, on the one hand, the growing number of repositories has facilitated the investigation of EEG-DL strategies, on the other hand, the time required to inspect and preprocess such data has increased remarkably.

EEG data preprocessing is a complex and time-consuming phase comprising several steps. Some of them, such as independent component rejection, bad channels detection, or segments removal, still require expert supervision to be optimally performed. Consequently, fully manual or semi-automated pipelines are unsuitable when dealing with large repositories with thousands of subjects. Automatic pipelines, on the other

hand, represent the optimal compromise between computational requirements and quality of the results. For example, software libraries such as HAPPE [10] and Automagic [11] can preprocess an entire dataset with a rich and customizable pipeline. Others, like *BIDSAlign* [12], further expand the preprocessing functionalities and allow to harmonise and combine multiple repositories, facilitating the investigation of novel DL strategies such as self-supervised learning [13].

EEG data preprocessing is extremely important and, if not done correctly, can lead to a significant drop in model performance, as thoroughly described in Section III. However, despite the large body of work published in the last decade, there is still no real consensus on the extent to which EEG data should be preprocessed for DL applications. According to Roy et al. [14], who reviewed 154 works that applied deep learning to EEG data, only a small percentage of researchers (23%) used an extensive preprocessing pipeline that included artifact handling, while the majority (47%) opted for minimal preprocessing. Furthermore, a small but still relevant percentage (15%) fed raw data directly into the model, sometimes achieving similar or even better results compared to a preprocessing scenario [15], [16]. Such numbers reflect the general idea that features could potentially be learned even with raw or minimally preprocessed data, reducing the need for domain-specific processing.

The use of minimally preprocessed data can be justified if one assumes that noise and artifacts defined in one scenario could become important features in another one, leading to improvements in model accuracy, as stated in the recent analysis of Cui et al. on model interpretability [17]. In contrast, studies like [18] found that a substantial preprocessing was required to prepare data for their specific task, leaving open the question of whether and to what extent EEG data preprocessing is necessary. The first attempt to give a formal answer to this question was that of Kingphai et al. [19], who investigated the effect of EEG preprocessing for the assessment of cognitive workload via DL approaches. They found an improvement in model's performance as richer pipelines were applied to raw data. However, their analysis focused only on a specific EEG task, limiting its generalizability. Therefore, a more comprehensive investigation of the impact of preprocessing on DL applications to EEG data could be of great interest to the research community seeking to fully exploit the capabilities of DL models in this domain.

Contributions: This work aims at thoroughly investigating the role of EEG preprocessing in DL applications, considering six classification tasks (both clinical and non-clinical) and four established EEG-DL architectures. First, it investigates the role of raw data and whether they can be effectively given to DL models without losing predictive power. Then, it compares pipelines with different levels of complexity, from a minimal filtering to richer preprocessing with established artifact handling automated methods. Finally, it can also be useful to researchers who want to understand the role of EEG preprocessing on specific research areas.

Paper structure: The outline of this paper is as follows. Section II describes in detail the experimental setup. Section III presents the results, which are discussed in Section IV. Finally, a conclusion is drawn in Section V.

II. Methods

This section details the methodological design of the presented analysis. It follows the checklist provided in [14], which covers all the information necessary to ensure results' reproducibility, namely: dataset selection, data preprocessing, models architecture, data partitioning, training hyperparameters, performance evaluation, and statistical analysis.

A. Dataset Selection

The inclusion criteria used for the dataset selection are:

- *open access*: all dataset must be publicly available at their respective OpenNeuro's [20] webpage.
- *raw data*: data should be raw, in order to allow applying progressively more preprocessing steps.
- *number of channels*: the minimum was set to 19 channels plus A1-A2 for contralateral referencing, following American Clinical Neurophysiology Society (ACNS) guidelines [21].

Additionally, the datasets were selected in order to represent several EEG tasks:

- 'Eye': physiological classification of eyes open and eyes closed recordings.
- 'MMI': motor movement imagery, a widely studied BCI application [22].
- 'Parkinson', 'FEP', 'Alzheimer': two and three classes pathology classification focused on relevant medical use-cases, i.e. Parkinson's, First Episode Psychosis and Alzheimer's diseases.
- 'Sleep': normal sleep vs sleep deprivation recognition.

The list of datasets and tasks used is summarised in Table II. The next paragraphs are dedicated to a brief description of the datasets used.

1) Eye - ds004148: This dataset [23], contains resting (eyes closed, eyes open) and cognitive state (subtraction, music, memory) recordings for 60 healthy subjects (age 20.0 $\pm$ 1.9 years). All the subjects have three sessions in which all the tasks are performed. Only eyes closed and eyes open recordings from session one (duration 5.0$\pm$0.0 minutes) have been considered.

2) MMI - ds004362: This dataset [24], contains recordings for rest and both imagery and physical movements for 109 healthy subjects (age 39$\pm$11 years). All subjects have one session (duration 1.9$\pm$0.4 minutes) with 14 tasks: two are rest, and the other twelve are three repetitions of four tasks.

- Task 1: open and close left or right fist;
- Task 2: imagine opening and closing left or right fist;
- Task 3: open and close both fists or feet;
- Task 4: imagine opening and closing both fists or feet;

This dataset is used to perform the 'MMI' task, thus only Task 2 is considered. Subjects with ID 88, 92 and 100, have been removed from the analysis since they have a different sampling rate and inconsistent trial length with respect to the others, as done in other works [25], [26].

3) Parkinson's - ds002778: This dataset [27], contains rest eyes open EEG recordings (duration 3.3 $\pm$ 0.3 minutes) from 15 Parkinson's patients and 16 age matched healthy controls (respectively 63.3$\pm$8.2 years, 63.5$\pm$9.7 years.) Healthy subjects have only one session, instead Parkinson's patients have two: one named 'ses-off' containing recordings where patients discontinued medication use at least 12 hours before

the session; one named 'ses-on', where patients are under medication. This dataset is used to perform the 'control' vs 'parkinson-off' task (named 'Parkinson'), thus only 'ses-off' was considered for the Parkinson's patients.

4) Parkinson's - ds003490: This dataset [28], contains rest eyes open/closed and auditory oddball EEG recordings (duration 9.9 ± 1.2 minutes) from 25 Parkinson's patients and 25 age matched healthy controls (respectively 69.7±8.7 years, 69.3 ± 9.6 years). Healthy subjects have only one session, while Parkinson's patients have two: one in which patients discontinued medication use at least 15 hours before the session and the other where patients are under medication. This dataset is used to perform the 'control' vs 'parkinson-off' task (named 'Parkinson'), thus only one session was considered for the Parkinson's patients, together with data from dataset ds002778.

5) Alzheimer's - ds004504: This dataset [29], contains EEG resting state eyes closed recordings (duration 13.4±2.3 minutes) from 36 Alzheimer's patients, 23 subjects with diagnosed frontotemporal dementia and 29 age matched healthy controls (respectively 66.4±7.9 years, 63.7 ± 8.2 years and 67.9 ± 5.4 years). This dataset is used to perform the 'control' vs 'alzheimer' vs 'frontotemporal dementia' task (named 'Alzheimer').

6) Sleep - ds004902: This dataset [30], contains EEG resting state eyes open recordings from 71 healthy subjects (age 20.0 ± 1.4 years). All the subjects have two sessions (duration 4.9 ± 0.4 minutes), one for sleep deprivation and the other for normal sleep state. Not all subjects have eyes-closed recordings, so only eyes-open recordings were used. This dataset is used to perform the 'normal' vs 'sleep deprived' task (named 'Sleep').

7) FEP - ds003947: This dataset [31], contains EEG resting state eyes open recordings (duration 5.2 ± 0.2 minutes) from 31 first episode schizophrenia (FESz) subjects and 30 age matched healthy controls (respectively 23.4±4.7 years, 24.2± 5.1 years). This dataset is used to perform the 'control' vs 'first episode psychosis' task (named 'FEP').

B. Data Preprocessing

All recordings from the selected datasets have been preprocessed using four distinct pipelines that incorporate common EEG preprocessing steps used by both the medical [10], [11] and the deep learning [14] community. These pipelines were implemented in *BIDSAlign*[1] [12], a customizable MATLAB-based library built on top of EEGLAB [32]. They were designed to gradually add preprocessing steps based on their complexity, computational requirements, and impact on the data. As detailed in the BIDSAlign paper [12], these steps are also featured in other well-known automated pipelines such as HAPPE [10] and Automagic [11] (refer to Section D of the Supplementary Material for more details). The algorithms used by BIDSAlign, for each of the mentioned steps, are the same used in these established pipelines, which are currently considered state-of-the-art. The four pipelines, named 'Raw', 'Filt', 'ICA' and 'ICA+ASR', are summarized in Table I, with a detailed description of each preprocessing step provided below, and the effect on the EEG bands defined in [12].

[1]https://github.com/MedMaxLab/BIDSAlign

TABLE I
PREPROCESSING PIPELINES CONSIDERED IN THIS WORK

preprocessing step	Raw	Filt	ICA	ICA+ASR
non-EEG channels removal	✓	✓	✓	✓
time segments removal			✓	✓
baseline removal		✓	✓	✓
resampling	✓	✓	✓	✓
filtering		✓	✓	✓
independent component analysis			✓	✓
automatic component rejection			✓	✓
bad-channel removal				✓
bad-time windows correction (ASR)				✓
spherical interpolation				✓
re-reference		✓	✓	✓
template alignment	✓	✓	✓	✓

- *non-EEG channels removal*: all non-EEG data such as electrocardiogram (ECG) or electrooculogram (EOG) included as extra channels are removed at the beginning.
- *time segments removal*: the first and last 8 seconds are removed. This avoids potential initial and final divergences in the recording due to technical issues like switching on-off the EEG amplifier. It affects mainly the (δ) band.
- *baseline removal*: DC-component is removed by subtracting the channel-wise mean from each data point. It affects only the zero frequency, not a specific band.
- *resampling*: Records were resampled to 250 Hz to reduce the computational load in while preserving relevant spectral properties. This choice follows the conservative rule of multiplying the maximum frequency of interest (45 Hz, lower gamma band limit) by four. 250 Hz corresponds to the smallest multiple of the native sampling rate for most selected datasets. It attenuates frequencies, above 125Hz.
- *filtering*: EEG records were filtered with a pass-band Hamming windowed sinc FIR filter between 1Hz and 45Hz. This avoids the use of a notch filter for line artifacts and allows for the removal of low frequency artifacts. Moreover, high-pass filtering at 1Hz is recommended to obtain good quality ICA decompositions [33]. It attenuates EEG bands outside the passband edges.
- *independent component analysis*: using the '*runica*' algorithm [34], posing no limit to the number of components.
- *automatic component rejection*: independent components were rejected automatically using IClabel [35], with thresholds for the brain class set between 0 and 0.1, while for the other six artifacts it was set between 0.9 to 1. This is a good trade-off between discarding lot of components, and keeping what could really be brain signal. Channel frequency changes follow the removed ICs' spectrum, weighted by mixing matrix coefficients.
- *bad-channel removal*: flat or extremely noisy channels are removed with the EEGLAB's '*clean_rawdata*' plugin, using default parameters. EEG band changes depend on the behavior of the removed channels and their artifacts.
- *bad-time windows removal*: this step was performed with Artefact Subspace Reconstruction (ASR) algorithm [36], using default parameters; bad-time windows are corrected and not removed. EEG band spectrum may shift significantly with extensive window removal/correction.

- *spherical interpolation*: previously automated removed channels are interpolated with the spherical method [37].
- *re-reference*: records are re-referenced to the common average.
- *template alignment*: to the International Federation of Clinical Neurophysiology (IFCN) 10-10 standard containing 61 channel. This step allows the alignment of EEG recording acquired with different electrodes and channel systems to the same expected input of a DL architecture. Consequently, it makes possible the merging of the ds002778 and ds003490 datasets for the Parkinson's task.

For dataset ds004362 (motor imagery task), whose records have a sampling rate of 160 Hz, resampling to 250Hz was not performed to avoid upsampling. In addition, a dedicated trial extraction step was done in place of time segment removal to preserve as many trials as possible.

To conclude, BIDSAlign preprocessed data went through two additional preparatory steps for the deep learning part. First, EEG recordings were further downsampled to 125 Hz to reduce computational load by decreasing the input sample size. In addition, this step standardizes all EEG data to a common sampling rate, which is a multiple of the native sampling rates for most of the selected datasets, and helps to minimize variations in model size across all considered tasks. A direct resampling from 250 Hz to 125 Hz was avoided to adhere to the conservative rule previously described, as such resampling could potentially compromise the quality of certain preprocessing steps, particularly the independent component analysis. Second, each EEG record was normalized by applying the z-score operator along the EEG channel dimension, i.e., by transforming each channels' signal so to get zero mean and unit variance. Data normalization have proven to be beneficial for EEG-DL applications [38].

C. Implementation Details

The rest of the experiment was conducted within a Python environment. In particular, models were trained using *SelfEEG*[2] [39], a comprehensive library for EEG-DL applications built on top of *Pytorch* [40]; statistical analysis relied on the *Scipy* framework [41], and figures were generated with *Seaborn* [42] coupled with *statannotations* [43]. Experiments were conducted on two NVIDIA Tesla V100 GPU (CUDA 12.1) devices to train multiple models in parallel. The maximum GPU memory allocation was 6.18 GB. Additional computational details can be found in the Supplementary Material.

1) EEG Architectures: Different models are expected to give different results. While this aspect is desired in other works that aim to compare novel strategies against baseline methods, here it can be considered as an external factor that could affect the results of the statistical analysis. For this reason, four different established EEG-DL architectures have been used, namely: EEGNet [44], ShallowConvNet [45], DeepConvNet [45], and FBCNet [46]. Models have been chosen considering several factors, such as their overall usage among the research community and their computational requirements (e.g., number of layers and parameters, training time, GPU occupancy). All models consist of a series of layers, including among others convolutions with horizontal or vertical kernels that compute operations respectively on the temporal (EEG samples) and on the spatial (EEG channels) dimension. Horizontal kernels allow extracting temporal patterns associated with artifacts or oscillatory activities, while vertical kernels find linear combination of channels reflecting brain activations. Furthermore, FBCNet incorporates a filter bank layer, a concept proven successful in the BCI domain. This layer is particularly suitable for this study, as certain artifact handling steps, such as the IC rejection, can highly impact the signal's power spectral density within specific frequency bands (e.g., delta [0.1-4] Hz, lower-gamma [30-45] Hz). A detailed description of each model can be found in the relative papers.

2) Data Partition: EEG data are known to be characterized by a high inter-subject variability [47]. The training could end up producing a model whose embeddings are dominated by subject-specific characteristics that could lead to misleading results: extremely good on subjects already seen but rather poor on unseen ones (see subsection A.3 of the Supplementary Material). However, most of the real-world applications expect models to be used on unseen subjects, which is why recent works stressed the importance of inter-subject evaluations, such as the Leave-N-Subjects-Out (LNSO), or the Leave-One-Subject-Out (LOSO) when N=1 [48]. The LNSO is a cross-validation method that partitions the dataset at the subject level, assigning all samples from a subject to either the training or validation set. This procedure provides a subject-independent performance estimate for new subjects, better reflecting real-world applications.

Further extending the LNSO procedure in a training-validation-test scenario, better suited for DL applications, this work proposed a Nested Leave-N-Subjects-Out (N-LNSO) cross-validation, which is schematized in Figure 1. In short, the N-LNSO combines two nested LNSO. The first LNSO creates a set of outer folds, which assign the subjects to use as test set, while the second LNSO repeatedly partitions the remaining subjects into training and validation sets, determining the number of inner folds. A model is trained for each of the inner folds and evaluated on the corresponding outer fold's test set. The procedure is repeated for each outer fold, yielding a total of $N_{\text{outer}} \times N_{\text{inner}}$ training instances. By doing so, each subject is guaranteed to participate at least once in every split (training, validation, and test). It is important to stress that, while a classical nested K-Fold is used to nest the hyperparameters optimization procedure under the model selection procedure, the goal of the N-LNSO is to better estimate the model performance distribution against unseen subjects and to provide an unbiased measure of centrality (with the test sets) for the subsequent statistical analysis [49].

For this work, the number of outer folds is $N_{outer} = 10$ and inner folds is $N_{inner} = 5$. Considering also the number of preprocessing pipelines $N_{pipeline} = 4$, models $N_{model} = 4$ and tasks $N_{task} = 6$, the total number of training was:

$$N_{pipeline} \times N_{model} \times N_{task} \times N_{outer} \times N_{inner} = 4800 \quad (1)$$

Splits were stratified in order to preserve class ratios.

Once assigned to a specific set, the EEG records were divided into non-overlapping windows of 4 seconds each. This duration was selected to match the length of the MMI trial (4.1 seconds), facilitating a more accurate comparison of

[2]https://github.com/MedMaxLab/selfEEG

TABLE II
DATASET AND TASKS SELECTED FOR THIS WORK

dataset ID	original reference	original number of channels	original sampling rate [Hz]	number of subjects	number of windows raw/filt (ica/ica+asr)*	task description	task acronym
ds004148	FCZ	64	500	60	9000 (8520)	Eyes Open vs Eyes closed	Eye
ds004362	A1-A2	64	160	109**	4748 (4748)	Imagine open/close left vs right fist	MMI
ds002778	CMS-DRL	41	512	31	8932 (8608)	Control vs Parkinson's off-medication	Parkinson
ds003490	CPZ	64	500	50			
ds004504	A1-A2	19	500	88	17604 (17252)	Control vs Alzheimer vs FT-dementia	Alzheimer
ds004902	FCZ	61	500	71	10482 (9914)	Normal vs Sleep deprived	Sleep
ds003947	TP9	61	1000	61	4718 (4474)	Control vs First Episode Psychosis (FEP)	FEP

*ICA and ICA+ASR pipelines have fewer samples due to the segment removal step, **three subjects were excluded from the analysis

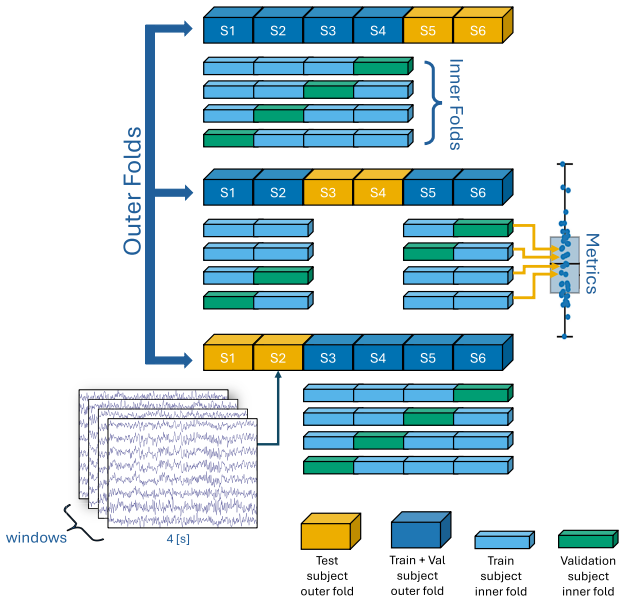

Fig. 1. Schematization of the Nested Leave-N-Subjects-Out cross-validation. Each outer fold identifies the subjects to use as test set, while the remaining subjects are repeatedly partitioned into training and validation sets based on the number of inner folds. All subjects' 4-second time windows are assigned to the same split set. A model is trained for each inner fold, using the validation set for early stopping purposes, and evaluated on the corresponding outer fold's test set. This is repeated for each outer fold, yielding a total of $N_{\text{outer}} \times N_{\text{inner}}$ training instances. This approach provides an unbiased estimate of the centrality measure to be used for the statistical analysis.

results at the inter-task level. The last portion of the recording was discarded if it was not long enough to generate another sample. Consequently, considering that the resampling and template alignment steps were included in each preprocessing pipeline, the input dimension was 61×500 (61×512 for the motor imagery). Additional information for each dataset (task) and its total number of samples are reported in Table II.

3) Training Hyperparameters: The whole experiment was conducted using a custom seed (83136297) to improve the reproducibility of the results. Each model was initialized using Pytorch's default settings, which are specific for each type of layer. Model architectures matched those proposed by the original authors, except for the max norm constraints, which were not used in this work. Models were trained with the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, no weight decay) [50], a batch size of 64, cross-entropy loss, and an exponential learning rate scheduler ($\gamma = 0.995$). The maximum number of epochs was set to 100, coupled with an early stopping with patience of 25 epochs (monitoring on the validation loss). Since different models are characterized by different learning curves, which can also vary depending on the investigated task, a custom learning rate was chosen for each possible couple (task, model), keeping it equal among all the pipelines. Learning rates were selected by evaluating the performance of a subset of models on the validation set in terms of consistency (median and overall rank of different metrics), searching on a discrete grid of 13 possible values: $1.0 \cdot 10^{-3}$, $7.5 \cdot 10^{-4}$, $5.0 \cdot 10^{-4}$, $2.5 \cdot 10^{-4}$, $1.0 \cdot 10^{-5}$, $7.5 \cdot 10^{-5}$, $5.0 \cdot 10^{-5}$, $2.5 \cdot 10^{-5}$, $1.0 \cdot 10^{-6}$. The used learning rate are listed in Table III.

TABLE III
LEARNING RATE GRID

Task	EEGNet	ShallowNet	DeepConvNet	FBCNet
Eye	$5.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$	$7.5 \cdot 10^{-4}$	$7.5 \cdot 10^{-4}$
MMI	$1.0 \cdot 10^{-3}$	$7.5 \cdot 10^{-4}$	$7.5 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$
Parkinson	$1.0 \cdot 10^{-4}$	$2.5 \cdot 10^{-4}$	$2.5 \cdot 10^{-4}$	$2.5 \cdot 10^{-4}$
Alzheimer	$7.5 \cdot 10^{-4}$	$5.0 \cdot 10^{-5}$	$7.5 \cdot 10^{-4}$	$7.5 \cdot 10^{-5}$
Sleep	$1.0 \cdot 10^{-3}$	$5.0 \cdot 10^{-5}$	$2.5 \cdot 10^{-4}$	$1.0 \cdot 10^{-4}$
FEP	$1.0 \cdot 10^{-4}$	$7.5 \cdot 10^{-5}$	$1.0 \cdot 10^{-3}$	$1.0 \cdot 10^{-5}$

4) Performance Evaluation: The performance of each model was evaluated using the balanced accuracy, which is preferred over the unbalanced one in biomedical applications as it takes into account differences in class ratio. By definition, balanced accuracy is the macro average of recall scores per class, so for a multi-class classification problem, with C classes:

$$\text{Accuracy}_{\text{Balanced}} = \frac{1}{C} \sum_{i=1}^{C} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{2}$$

where TP_i and FN_i stands respectively for the true positive and the false negative for class i. The statistical analysis against other metrics, such as recall, precision, and F1-score, is reported in section B of the Supplementary Material.

D. Statistical Analysis

The main objective of this work is to assess whether varying amounts of data preprocessing can yield different results in terms of metric under evaluation (balanced accuracy).

The procedure of repeatedly partitioning the dataset into different training, validation and test sets using the Nested Leave-N-Subjects-Out strategy previously described, gives as result an ensemble of trained models. In particular, each of

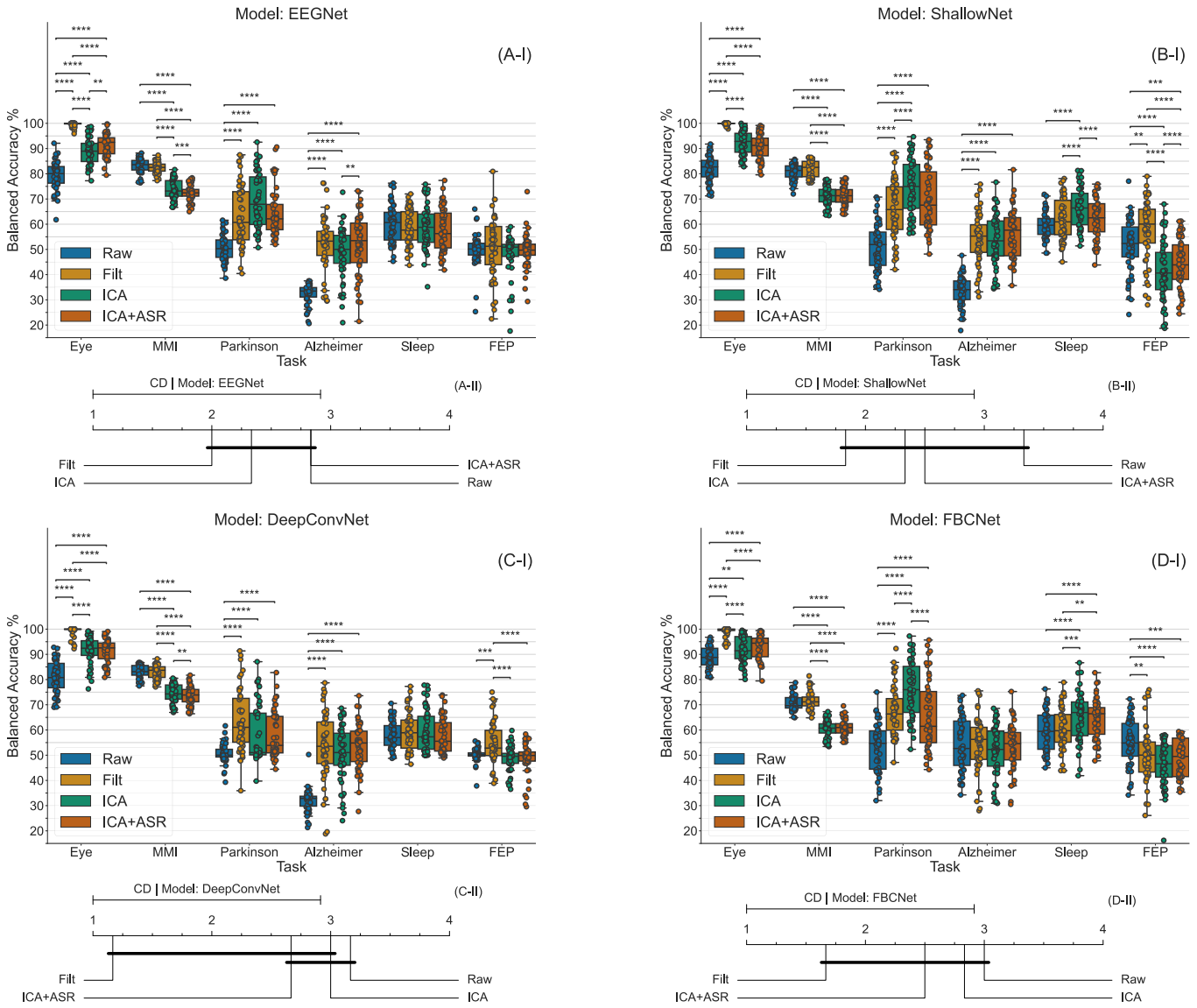

Fig. 2. Panels A to D (I) show the balanced accuracies for the 6 tasks across the 4 preprocessing pipelines. Results are marginalized by deep learning model: EEGNet in Panel A, ShallowNet in Panel B, DeepConvNet in Panel C, and FBCnet in Panel D. Asterisks indicate significant results from Wilcoxon's signed rank tests performed within the same task and Holm-corrected for multiple comparisons: (****) $p < \mathbf{0.0001}$; (***) $p < \mathbf{0.001}$; (**) $p < \mathbf{0.01}$; (*) $p < \mathbf{0.05}$. Panels A to D (II) show the Critical Distance (CD) diagram for the four pipelines, marginalized by model. All classifiers are compared against each other using Nemenyi's test. Groups of pipelines that are not significantly different ($p < \mathbf{0.05}$) are connected.

the six tasks has 200 trained models, evaluated in groups of 50 (one N-LNSO) for each of the four pipelines.

Considering one task (intra-task level), the distributions of the results for the four pipelines are compared pairwise with a Wilcoxon's signed rank test [51]. Since classification results derived from the same dataset preprocessed with different pipelines are dependendent, the Wilcoxon's test is preferred over the Mann-Whiteny U test [52]. The p-values are corrected for multiple comparison using the Holm's [53] procedure within the same task, and visualized together with the box-plots in Figure 2 (Panels A-D I).

Median balanced accuracy is used as measure of centrality for the distribution of the 50 models' performances, as it is preferable over the sample mean when distributions are not normal or skewed. The Friedmann's non-parametric test [54] is used to evaluate the differences on the medians between the four pipelines; in particular, it ranks the preprocessing pipeline for each dataset separately, the best performing pipeline getting rank 1, the second best rank 2, and so on. The core idea is to use different tasks, associated to one or more datasets, as evidence to demonstrate that pipelines have an impact in the classification result (inter-task level). This statistical test is well suited for such an analysis as described in [55] and [56] and it was previously applied for analogous purposes in [57]. The Friedmann's test does not rely on the one-way ANOVA's assumptions, which is its parametric counterpart. Therefore, it does not assume that samples are drawn from normal distributions (which does not hold in Figure 2, Panel B-I, task Parkinson) and does not assume sphericity (which does not hold in Figure 2, Panel C-I, task Alzheimer) [55].

In detail, let r_i^j be the rank of the j-th of k pipelines on the i-th of N task; the Friedman's test compares the average ranks

of the k pipelines R_j. Under the null-hypothesis H_0, which states that all the pipelines are equivalent, so their ranks R_j should be equal; thus the Friedman statistic

$$\chi_F^2 = \frac{12N}{k(k+1)} \sum_{j=1}^{k} R_j^2 - 3N(k+1), \quad R_j = \frac{1}{N} \sum_{i=1}^{N} r_i^j \quad (3)$$

is distributed according to a χ^2 distribution, with $k-1$ degrees of freedom, when N and k are big enough. For $k = 4$ (pipelines) and $N = 6$ (tasks), the Friedman's statistic at critical value $\alpha = 0.05$ is $\chi^2_{approx} = 7.815$ within the χ^2 approximation, while the tabulated value is $\chi^2_{tabular} = 7.6$. If H_0 is rejected, a post-hoc test is needed to understand which pipelines are significantly different. The Nemenyi's test [58] is used when all pipelines are compared to each other [55]. The performances of two pipelines, i and j, are significantly different if the corresponding average ranks R_i and R_j differ by at least the critical difference (CD):

$$|R_i - R_j| \geq CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \quad (4)$$

where critical values q_α are based on the Studentized range statistic divided by $\sqrt{2}$. A table with critical values can be found in [55]; for $k = 4$ (pipelines) and $N = 6$ (tasks), the critical distance is $CD = 1.915$. Results are visualized using the critical difference diagram, in Figure 2 (Panels A-D II).

This procedure is repeated using the four different deep neural architectures described in the previous subsection.

III. Results

Differences were found between the investigated pipelines at both the intra-task and inter-task levels. The balanced accuracies of all the 4800 trained models are reported in Figure 2. Results are presented in four panels, one for each deep learning architecture, respectively: A-EEGNet, B-ShallowNet, C-DeepConvNet, and D-FBCNet. An alternative visualization, with results marginalized by task is provided in Section E of the Supplementary Material. Considering the intra-task level of analysis, only strong significant results ($p < 0.001$) from the Wilcoxon tests are discussed in the next paragraphs.

A. EEGNet

Figure 2, Panels A I-II, presents EEGNet results. Wilcoxon tests reveal that Raw performs significantly worse than the other three pipelines in the Eye, Parkinson, and Alzheimer tasks; Raw and Filt outperform the others in the MMI task; and there are no significant differences in the Sleep and FEP tasks. The Friedman statistic $\chi_F^2 = 1.8$ shows no significant results ($p = 0.61$, under χ^2 assumption). Average ranks for the four pipelines are: $R_{\text{Raw}} = 2.83$, $R_{\text{Filt}} = 2$, $R_{\text{ICA}} = 2.33$ and $R_{\text{ICA+ASR}} = 2.83$, with Filt being the best. The largest average rank difference is $R_{\text{Filt-Raw}} = R_{\text{Filt - ICA+ASR}} = 0.83$, between Filt and Raw, and Filt and ICA+ASR.

B. ShallowNet

Figure 2, Panels B I-II, presents the ShallowNet results. Wilcoxon tests reveal that Raw is significantly worse than the other three pipelines in the Eye, Parkinson and Alzheimer task; Raw and Filt outperform the others in the MMI and FEP tasks; ICA outperforms the others in the Sleep task. The Friedman' statistic $\chi_F^2 = 4.2$ shows no significant results ($p = 0.24$, under χ^2 assumption). Average ranks for the four pipelines are: $R_{\text{Raw}} = 3.33$, $R_{\text{Filt}} = 1.83$, $R_{\text{ICA}} = 2.33$ and $R_{\text{ICA+ASR}} = 2.5$, with Filt being the best. The largest average rank difference is $R_{\text{Filt-Raw}} = 1.5$, between Filt and Raw.

C. DeepConvNet

Figure 2, Panels C I-II, presents the DeepConvNet results. Wilcoxon tests reveal that Raw is significantly worse than the other three pipelines in the Eye, Parkinson and Alzheimer task; Raw and Filt outperform the others in the MMI and FEP tasks; Filt outperforms the others in the FEP task; and there are no significant in the Sleep task. The Friedman' statistic $\chi_F^2 = 9.0$ shows statistically significant results at $\alpha = 0.05$ ($p = 0.029$, under χ^2 assumption). Average ranks for the four pipelines are: $R_{\text{Raw}} = 3.17$, $R_{\text{Filt}} = 1.17$, $R_{\text{ICA}} = 3.0$ and $R_{\text{ICA+ASR}} = 2.67$, with Filt being the best. The largest average rank difference is $R_{\text{Filt-Raw}} = 2$, between Filt and Raw.

D. FBCNet

Figure 2, Panels D I-II, presents the results for FBCNet. Wilcoxon tests reveal that Raw is significantly worse than the other three pipelines in the Eye and Parkinson tasks; Raw and Filt outperform the others in the MMI task; ICA outperforms the others in the Parkinson task, and is better than Raw and Filt in the Sleep task; Raw is better than ICA and ICA+ASR in the FEP task; there are no significant differences in the Alzheimer task. The Friedman' statistic $\chi_F^2 = 3.8$ shows no significant results ($p = 0.28$, under χ^2 assumption). Average ranks for the four pipelines are: $R_{\text{Raw}} = 3$, $R_{\text{Filt}} = 1.67$, $R_{\text{ICA}} = 2.83$ and $R_{\text{ICA+ASR}} = 2.5$, with Filt being the best. The largest average rank difference is $R_{\text{Filt-Raw}} = 1.33$, between Filt and Raw.

IV. Discussion

Researchers must deal with many degrees of freedom when working with EEG data, and preprocessing is undoubtedly a key one. If in the medical field the motto "the more, the better" is usually followed, with a large research community devoted to constantly improving the efficacy of automatic pipelines, when it comes to deep learning, things are not so clear-cut. While previous research work has demonstrated that DL can work with raw data for certain applications, the presented analysis shows how better results could be if data are at least filtered. The same applies for a more intense preprocessing. While the investigation of specific diseases (e.g., Parkinson's, Alzheimer's) can benefit from the addition of automatic artifact handling techniques, the results indicate that a minimal filtering represents the optimal trade-off between computational requirements and achieved accuracy.

Concerning the methodological aspects of this study, it is important to further discuss the relevance of data splitting, as extra care is needed when measuring and comparing the performance of EEG-DL models in relation to data preprocessing. A properly designed DL application requires separate training, validation, and independent test sets; improper data splitting can lead to overly optimistic results. For instance, splitting at the window level allows EEG windows from the same subject to appear in both training and test sets, introducing

bias. Future windows may end up in the training set, leading to scenarios where future data predict past events. Additionally, consecutive windows, even if separated in time, can remain highly correlated, further skewing results. Subject-specific information could be present throughout the recording, making windows easily predictable and overestimating test metrics (see section A.3 of the Supplementary Material). Splitting at the subject level helps producing more unbiased performance estimates, although how subjects are assigned to different sets still impacts model accuracy. N-LNSO allows for repeated measures of test accuracy across different sets, enhancing reliability. When repeated measures are available, t-tests are often used to assess group differences. However, Dietterich's work [59] shows that t-tests on k-results from k-fold cross-validation yield inflated p-values due to underestimated variance. Nadeau and Bengio [60] introduced a correction factor based on the ratio of samples between training and test sets, but their method assumes a constant ratio and ignores the validation set, which is often not suitable in many DL applications. In contrast, the Friedman test evaluates relative ranks of classifier performances rather than numerical differences. The N-LNSO strategy formalized in this study provides an unbiased estimate of median test accuracy, while the Friedman test yields realistic p-values for performance differences.

Concerning the results obtained, it is clear that EEG preprocessing significantly impacts model performance. For example, when performing the Parkinson's task with EEGNet (Figure 2 Panel A-I), the median balanced accuracy using raw data is 50.3%, while using the ICA pipeline it raises to 67.9%. Again, when performing the Eye task with the same model, the median value using raw data is 80.1%, while using the Filt pipeline it reaches 100%. The Wilcoxon signed rank test shows significant results across all models, supporting this finding. However, the goal and novelty of this study is to go beyond pairwise comparisons. Considering an architecture-specific perspective, Friedman's test does not always give significant results. Figure 3 demonstrates that the Filt pipeline gives the best results in more than half of the splits over the six tasks. The critical difference diagrams in Figure 2 further confirm this, with Filt consistently ranked first on average. Yet, Friedman's test only shows significance for DeepConvNet. Consequently, it remains unclear whether the outperformance of Filt over the Raw pipeline stems from the pipeline's inherent superiority, DeepConvNet's sensitivity to EEG preprocessing, or other factors. This question can be further investigated in future works with some insights from this study. Based on the performances of the Filt pipeline, it is plausible that other architectures will show significance on Friedman's test when the number of tasks is increased, and hence the critical difference is reduced. This would confirm the superiority of the Filt pipeline, regardless of the task, and would put the problem back to the low-power of Friedman's test.

The results of this study challenges the conventional reliance on complex preprocessing methods, suggesting that both neural activity and EEG artifacts can be used by EEG-DL models to solve the task at hand, as noted in [17]. However, the predictive power of these components is not necessarily equal in every scenario. For instance, in the eye task, both blinking artifacts and drops of alpha oscillations in the occipital area serve as discriminative features for the eyes-open condition.

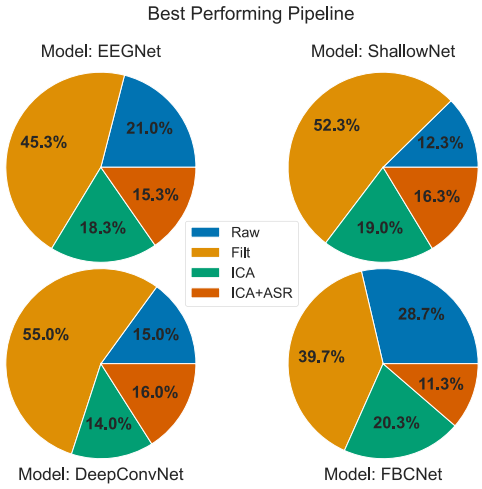

Fig. 3. Pie-charts reporting the number of times (in %) a pipeline is better than the others over the 300 training instances (6 tasks × 50 N-LNSO partitions). Results are marginalized by model.

At the same time, eye blinking, which typically appears as an artifact in the prefrontal channels, is a much easier characteristic to identify and exploit. This characteristic offers greater predictive power when it is not removed by artifact handling methods used in more complex pipelines, such as ICA and ICA+ASR. While the model is interpretable in both cases, the use of artifacts to solve the task can lead to less robust models. In contrast, clinical practice suggests that the characterization of neurological disorders such as Parkinson's or Alzheimer's relates more to power spectral density variations in specific frequency bands (e.g., alpha, theta) than to pathology-related artifacts. This may justify, at least in medical applications, the use of more complex pipelines (ICA, ICA+ASR) capable of improving the SNR. Unlike blinking artifacts, spectral differences are harder to quantify. They depend on raw data quality, noise overlap—where ICA may fail to separate noise from brain components—and inter-subject variability. This might explain the generally lower and more variable performance of clinical tasks (Parkinson's, Alzheimer's) compared to non-clinical tasks (MI and Eye).

The connection between preprocessing and performance of EEG-DL models is crucial and highlights how advancements in one area should consider the implications for the other. Future developments in EEG data preprocessing must acknowledge the differing needs of clinical and non-clinical approaches. Real-world non-clinical applications, like BCI experiments using MI or eye tasks in distracted driving, require real-time data collection and analysis. Fortunately, the Filt pipeline not only ranked first in similar tasks but also has the lowest computational load and is the only one applicable in real-time. However, despite its minimalist nature, this pipeline can be further enhanced, for example by increasing the efficiency of the filter, reducing its delay, or even integrating this step directly into DL models. The chosen EEG-DL models feature a convolutional temporal layer that processes the multi-channel input similarly to digital filters. If properly initialized and controlled during the training phase, this layer can be used to study frequency ranges or artifact-specific patterns, reducing the preprocessing computational load and improving model interpretability. Moving to

clinical applications, which have fewer real-time constraints, the accuracies in Figure 2 indicate that pipelines incorporating artifact handling methods should not be disregarded, but rather reviewed to address their limitations and enable active support in medical practice. Although integrating artifact handling methods as network layers is challenging, it should be noted that some algorithms, like ICLabel, are deep learning-based. Analyzing and improving these algorithms in light of this study's results could significantly enhance data quality and the robustness of deep learning models in clinical practice. This can be achieved, for example, by modifying the original architecture and leveraging open-source data alongside self-supervised training approaches.

The optimization of preprocessing pipelines, if directed at extracting crucial information for solving a task, can enhance model performance and reduce the variability of results shown in Figure 2. However, this process must consider the potential effects on the quality of features learned by the model and the feature space available during training. Studying these aspects is valuable for optimizing preprocessing pipelines for EEG-DL applications or for improving model design. Here are some important points and examples. The features learned by EEG-DL architectures depend on model design and type of preprocessing. Considering once again the eye task, minimalist pipelines (Filt) might produce models whose embeddings focus on ocular artifacts, while complex pipelines (e.g., ICA+ASR) can move the attention to the brain components of the signal. Sensitivity analyses, such as gradient-based methods, are commonly used to identify features that can influence decisions, and can guide the researcher in understanding which types of features are learned during training and selecting the best model [17]. The same can be said about the impact of preprocessing on the features learned. Narrow bandpass filters restrict temporal kernels from extracting band-specific information. Bad-channel interpolation methods can hinder spatial convolutional layers—vertical kernels with a height equal to the number of channels—from extrapolating linear combinations of inputs, as some channels are already linear combinations of others. Squared activation with pooling allows to extract the signal power globally or locally [45]. While preprocessing effects on learned features are easy to inspect, its impact on feature space is not. Shaped by preprocessing, model initialization, and training trajectory, exploring this space requires analyzing numerous plausible EEG windows under continuous training, a largely unexplored area.

In conclusion, the following study not only helps to clarify the role of preprocessing on the performance of EEG-DL models but also lays the groundwork for future research. This pertains not only to the methodological aspects for conducting comparative analyses of confounding elements in EEG-DL experiments, but also to how preprocessing and model design must be optimized to work in synergy and improve the reliability of these systems in real-world applications.

V. Conclusion

In this work, the role of preprocessing in EEG-DL applications was investigated. Four distinct preprocessing of varying complexity were examined (Raw, Filt, ICA, ICA+ASR), ranging from raw and minimally filtered data to complex pipelines with automated artifact handling algorithms. The analysis was conducted on six representative clinical and non-clinical use cases, using four established EEG-DL architectures. Performance metrics from 4800 training instances were collected with an unbiased cross-validation strategy (Nested Leave-N-Subjects-Out) and used to identify statistical differences at both the intra-task and inter-task levels. Significant results among the preprocessing pipelines emerged at the intra-task level across all models, while inter-task differences were found only on the largest investigated model, DeepConvNet. However, the analysis revealed that models trained using raw data consistently performed poorly, ranking last on average, while models trained on minimally preprocessed data with no artifact handling consistently ranked first on average. The Filt pipeline obtains higher performance in most cases than reacher pipelines. However, the choice of preprocessing should be performed carefully in relation to the analyzed task.

Acknowledgment

Contributions: Federico Del Pup: Conceptualization, data preprocessing, model training, and writing—original draft; Andrea Zanola: Conceptualization, data preprocessing, statistical analysis, and writing—original draft; and Manfredo Atzori: Project administration and supervision. All authors reviewed and edited the manuscript.

Code and data availability: code for results and figures is available at https://github.com/MedMaxLab/eegprepro. Supporting data are openly accessible on OpenNeuro. The authors declare no relevant competing interests.

References

[1] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of EEG-based brain–computer interface paradigms," *J. Neural Eng.*, vol. 16, no. 1, Nov. 2018, Art. no. 011001.

[2] X. Li et al., "EEG based emotion recognition: A tutorial and review," *ACM Comput. Surv.*, vol. 55, no. 4, pp. 1–57, 2022.

[3] K. Aboalayon, M. Faezipour, W. Almuhammadi, and S. Moslehpour, "Sleep stage classification using EEG signal analysis: A comprehensive survey and new investigation," *Entropy*, vol. 18, no. 9, p. 272, Aug. 2016.

[4] K. Rasheed et al., "Machine learning for predicting epileptic seizures using EEG signals: A review," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 139–155, 2020.

[5] A. M. Maitin, J. P. Romero Muñoz, and Á. J. García-Tejedor, "Survey of machine learning techniques in the analysis of EEG signals for Parkinson's disease: A systematic review," *Appl. Sci.*, vol. 12, no. 14, p. 6967, Jul. 2022.

[6] Y. Zhao and L. He, "Deep learning in the EEG diagnosis of Alzheimer's disease," in *Proc. ACCV*. Cham, Switzerland: Springer, 2015, pp. 340–353.

[7] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *J. Neural Eng.*, vol. 16, no. 3, Jun. 2019, Art. no. 031001.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] C. P. D. Silva, S. Tedesco, and B. O'Flynn, "EEG datasets for healthcare: A scoping review," *IEEE Access*, vol. 12, pp. 39186–39203, 2024.

[10] L. J. Gabard-Durnam, A. S. Mendez Leal, C. L. Wilkinson, and A. R. Levin, "The Harvard automated processing pipeline for electroencephalography (HAPPE): Standardized processing software for developmental and high-artifact data," *Frontiers Neurosci.*, vol. 12, Feb. 2018, Art. no. 316496.

[11] A. Pedroni, A. Bahreini, and N. Langer, "Automagic: Standardized preprocessing of big EEG data," *NeuroImage*, vol. 200, pp. 460–473, Oct. 2019.

[12] A. Zanola, F. Del Pup, C. Porcaro, and M. Atzori, "BIDSAlign: A library for automatic merging and preprocessing of multiple EEG repositories," *J. Neural Eng.*, vol. 21, no. 4, Aug. 2024, Art. no. 046050.

[13] M. H. Rafiei, L. V. Gauthier, H. Adeli, and D. Takabi, "Self-supervised learning for electroencephalography," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 1457–1471, Feb. 2024.

[14] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: A systematic review," *J. Neural Eng.*, vol. 16, no. 5, Oct. 2019, Art. no. 051001.

[15] N. K. N. Aznan, S. Bonner, J. Connolly, N. Al Moubayed, and T. Breckon, "On the classification of SSVEP-based dry-EEG signals via convolutional neural networks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2018, pp. 3726–3731.

[16] M. A. Almogbel, A. H. Dang, and W. Kameyama, "EEG-signals based cognitive workload detection of vehicle driver using deep learning," in *Proc. 20th Int. Conf. Adv. Commun. Technol. (ICACT)*, Chuncheon, (South) Korea, Feb. 2018, pp. 256–259.

[17] J. Cui, L. Yuan, Z. Wang, R. Li, and T. Jiang, "Towards best practice of interpreting deep learning models for EEG-based brain computer interfaces," *Frontiers Comput. Neurosci.*, vol. 17, Aug. 2023, Art. no. 1232925.

[18] R. Hefron, B. Borghetti, C. Schubert Kabban, J. Christensen, and J. Estepp, "Cross-participant EEG-based assessment of cognitive workload using multi-path convolutional recurrent neural networks," *Sensors*, vol. 18, no. 5, p. 1339, Apr. 2018.

[19] K. Kingphai and Y. Moshfeghi, "On EEG preprocessing role in deep learning effectiveness for mental workload classification," in *Proc. H-WORKLOAD*. Cham, Switzerland: Springer, 2021, pp. 81–98.

[20] C. J. Markiewicz et al., "The OpenNeuro resource for sharing of neuroscience data," *eLife*, vol. 10, Oct. 2021, Art. no. e71774.

[21] S. R. Sinha et al., "American clinical neurophysiology society guideline 1: Minimum technical requirements for performing clinical electroencephalography," *J. Clin. Neurophysiol.*, vol. 33, no. 4, pp. 303–307, 2016.

[22] H. Altaheri et al., "Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review," *Neural Comput. Appl.*, vol. 35, no. 20, pp. 14681–14722, Jul. 2023.

[23] Y. Wang, W. Duan, D. Dong, L. Ding, and X. Lei, "A test-retest resting and cognitive state EEG dataset," OpenNeuro, 2022, doi: 10.18112/openneuro.ds004148.v1.0.1.

[24] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "EEG motor movement/imagery dataset," OpenNeuro, 2022.

[25] C. M. Köllõd, A. Adolf, K. Iván, G. Márton, and I. Ulbert, "Deep comparisons of neural networks from the EEGNet family," *Electronics*, vol. 12, no. 12, p. 2743, Jun. 2023.

[26] G. Zoumpourlis and I. Patras, "Motor imagery decoding using ensemble curriculum learning and collaborative training," in *Proc. 12th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2024, pp. 1–8.

[27] A. P. Rockhill, N. Jackson, J. George, A. Aron, and N. C. Swann, "UC San Diego resting state EEG data from patients with Parkinson's disease," OpenNeuro, 2021.

[28] J. F. Cavanagh, "EEG: 3-stim auditory oddball and rest in Parkinson's," OpenNeuro, 2021.

[29] A. Miltiadous et al., "A dataset of EEG recordings from: Alzheimer's disease, frontotemporal dementia and healthy subjects," *OpenNeuro*, vol. 1, p. 88, Feb. 2023.

[30] C. Xiang, X. Fan, D. Bai, K. Lv, and X. Lei, "A resting-state EEG dataset for sleep deprivation," *Sci. Data*, vol. 11, no. 1, p. 427, Apr. 2024.

[31] D. Salisbury, D. Seebold, and B. Coffman, "EEG: First episode psychosis vs. control resting task 2," OpenNeuro, 2022.

[32] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[33] M. Klug and K. Gramann, "Identifying key factors for improving ICA-based decomposition of EEG data in mobile and stationary experiments," *Eur. J. Neurosci.*, vol. 54, no. 12, pp. 8406–8420, 2021.

[34] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.

[35] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "ICLabel: An automated electroencephalographic independent component classifier, dataset, and website," *NeuroImage*, vol. 198, pp. 181–197, Sep. 2019.

[36] T. R. Mullen et al., "Real-time neuroimaging and cognitive monitoring using wearable dry EEG," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 11, pp. 2553–2567, Nov. 2015.

[37] S. S. Kang, T. J. Lano, and S. R. Sponheim, "Distortions in EEG interregional phase synchrony by spherical spline interpolation: Causes and remedies," *Neuropsychiatric Electrophysiol.*, vol. 1, no. 1, pp. 1–17, Dec. 2015.

[38] A. Apicella, F. Isgrò, A. Pollastro, and R. Prevete, "On the effects of data normalization for domain adaptation on EEG data," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106205.

[39] F. Del Pup, A. Zanola, L. F. Tshimanga, P. E. Mazzon, and M. Atzori, "SelfEEG: A Python library for self-supervised learning in electroencephalography," *J. Open Source Softw.*, vol. 9, no. 95, p. 6224, Mar. 2024.

[40] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, p. 12. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[41] P. Virtanen, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, Feb. 2020.

[42] M. Waskom, "Seaborn: Statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021.

[43] F. Charlier et al., "Statannotations," Zenodo, Version 0.7.1, Nov. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.14258156

[44] V. Lawhern, A. Solon, N. Waytowich, S. M. Gordon, C. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, p. 056013, 2018.

[45] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, 2017.

[46] R. Mane, N. Robinson, A. P. Vinod, S. Lee, and C. Guan, "A multi-view CNN with novel variance layer for motor imagery brain computer interface," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 2950–2953.

[47] J. Y. Cheng, H. Goh, K. Dogrusoz, O. Tuzel, and E. Azemi, "Subject-aware contrastive learning for biosignals," 2020, *arXiv:2007.04871*.

[48] S. Kunjan et al., "The necessity of leave one subject out (LOSO) cross validation for EEG disease diagnosis," in *Proc. 14th Int. Conf. Brain Inform.* Cham, Switzerland: Springer, 2021, pp. 558–567.

[49] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, Jul. 2010.

[50] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, p. 13.

[51] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.

[52] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, Mar. 1947.

[53] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandin. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.

[54] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Stat. Assoc.*, vol. 32, no. 200, pp. 675–701, Dec. 1937.

[55] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.

[56] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," 2018, *arXiv:1811.12808*.

[57] E. Alshdaifat, D. Alshdaifat, A. Alsarhan, F. Hussein, and S. M. F. S. El-Salhi, "The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance," *Data*, vol. 6, no. 2, p. 11, Jan. 2021.

[58] P. B. Nemenyi, *Distribution-Free Multiple Comparisons*. Princeton, NJ, USA: Princeton Univ., 1963.

[59] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.

[60] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Mach. Learn.*, vol. 52, no. 3, pp. 239–281, Sep. 2003.