Achieving Reproducibility in EEG-Based Machine Learning

Sean Kinahan*
skinahan@asu.edu
School of Computing and Augmented Intelligence, Arizona State University
Tempe, Arizona, USA

Pouria Saidi
School of Electrical, Computer, and Energy Engineering, Arizona State University
Tempe, Arizona, USA

Ayoub Daliri
College of Health Solutions, Arizona State University
Tempe, Arizona, USA

Julie Liss
College of Health Solutions, Arizona State University
Tempe, Arizona, USA

Visar Berisha†
College of Health Solutions, Arizona State University
Tempe, Arizona, USA

ABSTRACT

Despite the inherent complexity of electroencephalogram (EEG) data characterized by its high dimensionality, artifactual noise, and biological variability, many machine learning (ML) studies claim impressive performance in decoding or classifying EEG signals. Recently, several studies have highlighted that flawed data analysis is a prevalent issue in the literature, leading to irreproducible results and exaggerated claims. To address this issue, we propose a framework that addresses three primary obstacles in EEG ML research: data leakage, data scarcity, and flawed model selection. We introduce the EEG ML Model Card, a standardized and transparent EEG ML model documentation tool that aims to directly address these pitfalls and enhance reproducibility and trustworthiness in EEG ML research.

CCS CONCEPTS

• **Computing methodologies → Model verification and validation**.

KEYWORDS

EEG, Machine Learning, Reproducibility

ACM Reference Format:
Sean Kinahan, Pouria Saidi, Ayoub Daliri, Julie Liss, and Visar Berisha. 2024. Achieving Reproducibility in EEG-Based Machine Learning. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3630106.3658983

1 INTRODUCTION

Machine learning (ML) applications in healthcare have advanced rapidly in recent years under the assumption that ML algorithms are well suited to recognize patterns in complex biological signals. Electroencephalographic (EEG) data allows researchers to investigate task-related brain activity for many different purposes, including the design of computer-aided diagnostic systems for neurological disorders [13]. However, a trend of non-reproducible results plagues the literature [8, 15, 19, 22, 27, 30, 38], impeding the fair assessment of new approaches and algorithms.

The lack of reproducibility in EEG-based ML research is a symptom of broader issues within healthcare ML studies. This challenge underscores the critical need for adopting reproducible research practices across these scientific domains [3, 20]. For high-stakes applications like healthcare, the repercussions of deploying non-reproducible models are significant; they can deny necessary prescriptions, misinterpret X-rays, and overlook common health concerns, thereby harming patients and eroding public trust in ML technology [32, 39, 40].

The proliferation of these issues can be attributed, in part, to the rapid growth and appeal of ML in diverse fields. This attractiveness is coupled with a high degree of design flexibility and a range of analytical methods, which, as highlighted by Ioannidis' seminal work "Why Most Published Research Findings Are False," significantly increases the probability of reporting overoptimistic findings [6]. This problem is exacerbated in competitive fields where there is a rush to publish results, leading to the propagation of the Fallacy of AI Functionality [25]. Such overoptimism not only misleads scientific inquiry but also poses risks when these prematurely lauded AI models are deployed in real-world applications. In the context of EEG processing, which has critical applications ranging from rehabilitation and seizure detection to brain-computer interfaces (BCI), the stakes are exceptionally high. Overly optimistic results not only misguide further research and development efforts in these important areas but can also lead to deployment of the technology for questionable applications [37].

EEG data is complex, highly variable, and high-dimensional. The typical analysis pipeline includes several steps, including data preprocessing, channel and feature selection, and model selection, increasing analytical flexibility for algorithm developers. As a result, design and evaluation methods for EEG ML models vary considerably across research fields and are documented differently. It becomes challenging to compare ML approaches for EEG due to a lack of unified and transparent reporting standards. Together, these factors harm reproducibility and trustworthiness in EEG ML.

[*] Also with College of Health Solutions, Arizona State University.
[†] Also with School of Electrical, Computer, and Energy Engineering, Arizona State University.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658983

The purpose of this paper is twofold. Firstly, we identify the primary reproducibility risk factors and ML pitfalls facing EEG ML researchers: data leakage, data scarcity, and flawed or insufficient validation methods. We provide experimental examples to illustrate the impact of these pitfalls and provide our recommendations for avoiding them. Secondly, we aim to unify reporting standards in EEG ML by proposing a Model Card specifically tailored for EEG applications. The Model Card for EEG helps researchers to avoid methodological flaws in EEG ML analysis. If adopted at the individual level, the model card can advance transparent and trustworthy EEG ML research [21].

1.1 Electroencephalographic (EEG) Data

EEG signals have several characteristics that present a challenge for the design of effective ML systems based on EEG data:

- High dimensionality
  EEG is high-dimensional by nature, with multiple channels used in recording, and each sampled at hundreds of samples per second. As a result, developers often apply feature and channel selection to reduce dimensionality [19].
- Artifactual Noise
  Unwanted artifacts (e.g., eye blinks and movements) are present in EEG signals. Developers commonly use preprocessing techniques to remove these artifacts.
- Biological Variability
  Variability is high at the inter- and intra-subject level for EEG data. Myriad factors, including attention, fatigue, and cognitive state, can impact the EEG signal. This variability cannot be easily controlled using computational methods alone. To control for these sources of variability, developers must consider relevant noise sources during data acquisition.

These characteristics complicate the development process and directly impact the reproducibility of EEG-based ML studies. High dimensionality and artifactual noise necessitate complex preprocessing and feature selection methods, which can vary significantly between studies, leading to inconsistent results. Similarly, the inherent biological variability, both inter- and intra-subject, affects the consistency of EEG data collected across different studies or even within the same study over time. This variability can introduce significant fluctuations in the data, making it challenging to replicate findings. These characteristics collectively contribute to the primary reproducibility risk factors in EEG-based ML – data leakage, data scarcity, and flawed or insufficient validation methods:

- Data leakage
  Data leakage can be exacerbated by inconsistent preprocessing methods used to address high dimensionality and noise. This results in identification of spurious relationships between the independent and target variables. Data leakage is a primary source of errors and overoptimistic performance in ML models across varied research fields [8].
- Data scarcity
  EEG datasets tend to have few data samples compared to datasets typical of other ML sub-fields. Data scarcity is particularly problematic in the context of EEG's biological variability, as limited data samples may not adequately capture this variability, leading to overfitting. The effectiveness of ML methods for EEG analysis is hampered by limited data availability as repeated overuse of small datasets often leads to "overfitting to a dataset" [9].
- Flawed or insufficient validation methods
  Validation methods are critical for accurately estimating the performance of an EEG ML model. Improper or incomplete validation can introduce bias in model assessments, jeopardizing result validity [30].

2 REPRODUCIBILITY

A fundamental problem facing the ML research community is a lack of reproducibility, often called the "reproducibility crisis" [8]. ML-based science pitfalls such as data leakage and overfitting are common, often leading to unrealistic or exaggerated claims about model performance. To make matters worse, researchers often fail to identify these exaggerated results or the practices that lead to them prior to publication [8]. These issues are amplified in EEG analysis for the reasons described in the previous section. This section provides an overview of the sources that negatively impact reproducibility and our recommendations for ensuring trustworthiness and reproducibility in EEG-based ML.

2.1 Data Leakage

In the present study, we define data leakage as the unintentional inclusion of information in the training data that would not be available at the time of prediction, leading to overly optimistic performance estimates or incorrect model predictions. Kapoor and Narayanan [2022] identified a taxonomy of data leakage errors that can occur in ML science [8]. Based on this taxonomy, we highlight typical pitfalls in EEG studies with examples from the literature. We organize the discussion according to the three levels of errors identified in the original study.

2.1.1 L1 error: Lack of clean separation between training and test dataset. L1 errors include all data leakage caused by poor training and test data segregation. When this separation is not maintained, evaluating model performance fairly becomes infeasible. A recent study on a steady-state visual-evoked potential (SSVEP) based BCI system [22] exposed reproducibility challenges. Proper adherence to ML practices in data separation resulted in a significant accuracy decrease in the proposed system compared to the initially published results. Nakanishi et al. [2020] concluded that the original study exhibited leakage of test samples into the training set, leading to an overstatement of ML model performance in the original publication [10]. Another study identified data leakage caused by insufficient separation of training and test sets during data augmentation [8]. The lack of clean separation between the training and test samples is an example of L1 error. This example demonstrates how L1 errors can provide an unfair advantage to ML models due to overfitting.

To demonstrate the impact of data leakage on model accuracy, we present experimental demonstrations focusing on EEG motor imagery classifiers. The first experiment simulates an L1 error due to leakage during feature selection. The simulation design was motivated by the feature selection leakage issues that have been reported by Shim et al. [2021] [31] and Lee et al. [2023] [13]. For all EEG experiments, we utilized the open-access EEGBCI motor imagery dataset [29]. Our EEGBCI subset included 45 samples of

64-channel EEG recorded from a single subject performing hand and foot motor imagery tasks. We used an 80-20 split into train and test sets for every experimental trial. Unless otherwise stated, 20 trials were performed for each leakage experiment. Our classifier performance plots include mean test accuracy and 95% confidence intervals for each experimental condition. We performed basic EEG preprocessing steps, including bandpass filtering, epoching, and basic feature extraction. A total of 576 features were extracted, including per-channel maximum, minimum, and mean amplitudes, standard deviation, energy, skewness, and kurtosis. We then performed feature selection twice using a select K-best method with $K = 100$. In the first iteration, we provided only the training data to the feature selector during training. In the second iteration, we performed feature selection using the entire dataset. Lastly, we trained two logistic regression classifiers $\mathrm{L_A}$ and $\mathrm{L_B}$ on the training set. We trained classifier $\mathrm{L_A}$ using the properly selected features and classifier $\mathrm{L_B}$ using the contaminated features. We then evaluated the accuracy of these classifiers on the test set.

Classifier $\mathrm{L_B}$ outperformed classifier $\mathrm{L_A}$ because of data leakage during feature selection. Classifier $\mathrm{L_A}$ mean accuracy was 0.49, while classifier $\mathrm{L_B}$ attained a mean accuracy measure of 0.60. Fig. 1a illustrates the invalid performance advantage given to classifier $\mathrm{L_B}$ due to a clear L1 error since classifier $\mathrm{L_B}$ has been refined based on information from the test dataset. Although the training and test sets were separated during training, the feature selection process was not applied appropriately. The error magnitude in this experiment is comparable to the error reported by past studies of biased EEG feature selection [13, 31].

Brain activity is attenuated by the subject's skull and scalp surface during an EEG study before the electrodes detect it. As a result, EEG electrodes near one another tend to exhibit a high degree of redundancy in the data captured. Additionally, brain activity may be localized to specific cortical regions for a given task. EEG ML researchers often apply channel selection procedures during preprocessing to take advantage of these characteristics of EEG data while decreasing dataset dimensionality.

Channel selection can be beneficial in making a dataset more tractable for ML methods; however, improper application of channel selection methods results in data leakage. Using the same experimental setup as described previously, we demonstrate the impact of data leakage during channel selection in a binary EEG classification scenario. After preprocessing steps, including filtering and epoching, we perform channel selection using the Common Spatial Patterns (CSP) method with four components. We performed CSP channel selection twice. During the first iteration, we split the data correctly into training and test sets, and we applied the channel selection method only to the training set. In the second iteration, we supplied the entire dataset during channel selection. After applying the CSP transform, we trained and evaluated a logistic regression classifier using the training and test sets respectively. As before, we designate these classifiers as $\mathrm{L_A}$ (no leakage) and $\mathrm{L_B}$ (channel selection leakage).

As a result of the data leakage, classifier $\mathrm{L_B}$ significantly outperforms classifier $\mathrm{L_A}$. We note that although we separated the training and test sets during training, data leakage during the earlier channel selection step dramatically impacted the performance of these classifiers. The mean accuracy of classifier $\mathrm{L_A}$ was 0.46, while classifier $\mathrm{L_B}$ attained a mean accuracy of 0.94. Fig. 1b illustrates this comparison.

Data leakage errors are a pervasive issue in EEG analysis. Experienced ML researchers may consider L1 errors easily avoidable, yet they continually undermine the validity of published EEG ML results. A meta-analysis of 37 EEG epilepsy detection studies by Lemoine et al. [2023] noted that only eight studies did not present any data leakage [14]. Some of these studies evaluated performance by testing directly on the training data, a textbook L1 error. Additionally, all studies that conducted feature selection experienced data leakage during this stage [14].

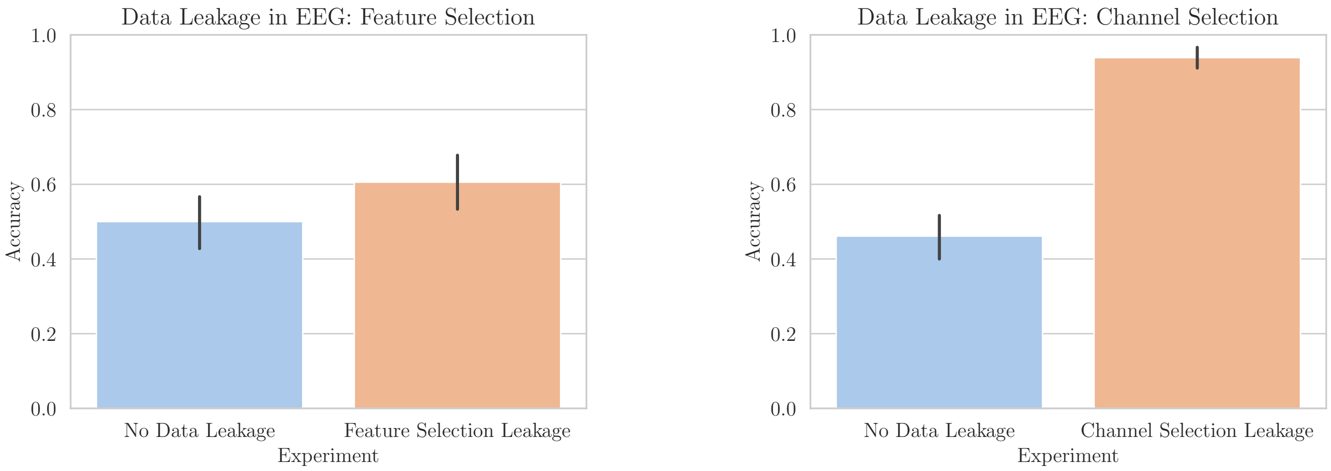
2.1.2 *L2 error: Model uses features that are not legitimate.*

L2 errors arise from using illegitimate features during classification. These errors arise when an ML model can access features that would not be available in practice. A study on biases in Event-Related Potential (ERP) BCI experiments modeled how BCI algorithms can leverage covariates not accounted for during experimental design to perform classification tasks [12]. Using EEG data from a visual priming experiment, La Fisca et al. [2022] demonstrated that psycho-linguistic and image covariates can significantly affect the regression process of a classifier [12]. In this context, covariate properties refer to uncontrolled stimulus characteristics that may be imbalanced across classification categories. For example, some psycho-linguistic covariates investigated by La Fisca et al. included the number of phonemes in an item's name, familiarity, and age of acquisition; image features included contrast, compactness, complexity, and homogeneity [12]. In the typical case, these covariate features are not modeled nor well-balanced across categories. Therefore, the classification algorithm can exploit dataset biases on these variables rather than the desired categorical effect. La Fisca et al. [2022] note that the biasing effect of these covariates increases with the complexity of the model [12]. When an EEG ML model can leverage covariate features to perform classification, an L2 error occurs.

To highlight the impact of L2 errors on accuracy, we present an experimental demonstration of a model that uses illegitimate features that can be used as a proxy for the outcome variable [8]. Using the EEGBCI motor imagery dataset [29], we performed bandpass filtering, epoching, and feature extraction. We then contaminated the feature set by introducing features that have a randomly weighted correlation with the true label. Sample labels were used to alter 2% of the original features. This process simulates the biasing effects of uncontrolled covariate properties, such as non-categorical image features in the EEG classification of visual stimuli [12]. We trained two logistic regression classifiers $\mathrm{L_A}$ and $\mathrm{L_B}$ on the training set. We trained classifier $\mathrm{L_A}$ using the original feature set and classifier $\mathrm{L_B}$ using the contaminated features. We finally evaluated the accuracy of these classifiers on the test set. This process was repeated 20 times, and we reported mean test accuracy of classifiers $\mathrm{L_A}$ and $\mathrm{L_B}$. As a result of the L2 error, classifier $\mathrm{L_B}$ outperforms classifier $\mathrm{L_A}$. The mean accuracy of classifier $\mathrm{L_A}$ was 0.62, while classifier $\mathrm{L_B}$ attained an accuracy measure of 0.70. Fig. 2a illustrates the impact of this error on the classification performance.

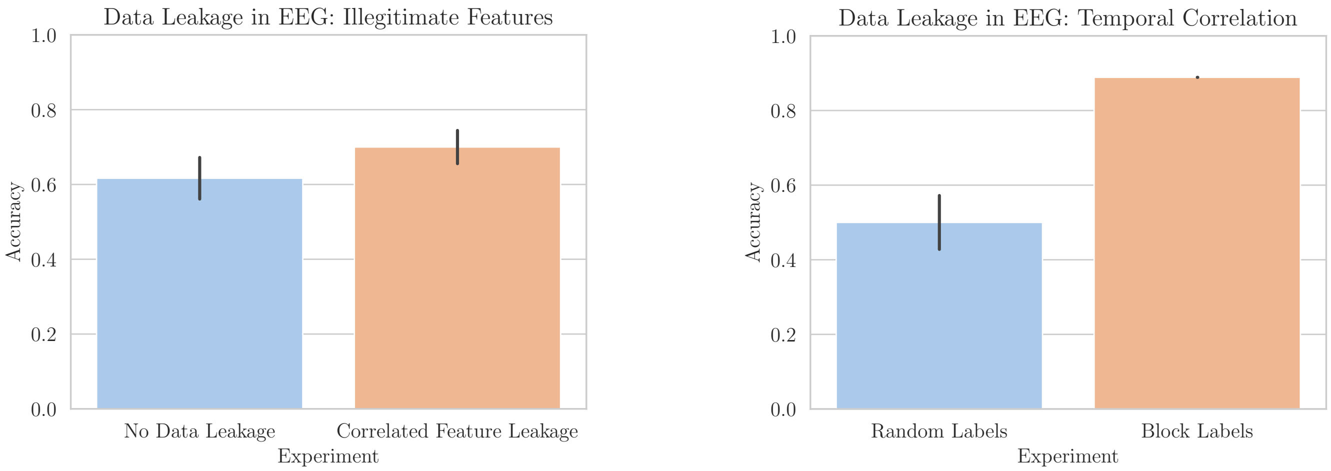
2.1.3 *L3 error: Test set is not drawn from the distribution of scientific interest.*

L3 errors occur when the test set does not originate from



(a) **(b)**

Figure 1: (a) Demonstration of data leakage during the feature selection process. For this comparison, we trained two logistic regression classifiers L_A and L_B after applying a Select K-Best feature selection method with $K = 100$. (b) Demonstration of data leakage during EEG channel selection. We trained two logistic regression classifiers L_A and L_B after applying a Common Spatial Patterns (CSP) channel selection method. Classifier L_A was trained after proper application of CSP, while classifier L_B was trained after CSP with intentional data leakage.



(a) **(b)**

Figure 2: (a) Demonstration of leakage due to the use of illegitimate classification features. For this comparison, we trained two logistic regression classifiers L_A and L_B to classify EEG motor imagery samples with and without correlated feature leakage. (b) Demonstration of leakage due to nonindependence of training and test set. For this comparison, we trained two logistic regression classifiers L_A and L_B to classify EEG motor imagery samples with two sets of randomly assigned labels. Classifier L_A was trained on samples with individually randomized class labels. Classifier L_B was trained on a feature set where it was highly likely for samples recorded close together in time to be assigned the same label.

the distribution of scientific interest. EEG signals capture the correlated activity of neurons, influenced by various factors including neuron connectivity [18]. Although the spatiotemporal structure of these oscillations is not fully known, Linkenkaer-Hansen et al. [2001] demonstrated the presence of long-range temporal correlations and power-law scaling behavior in the frequency range of 10 and 20 Hz [17]. These temporal correlations have a significant impact on the outcome of experiments that are designed in a block fashion. Li et al. [2021] provided evidence that ML models learn and classify arbitrary brain states based on these temporal correlations instead of the target features of the data and can result in unrealistic estimates of the ML model's performance [15]. Additionally, West et al. [2023] identified temporal correlations as a significant confound in EEG seizure prediction [38]. This nonindependence of training and test samples constitutes an L3 error [8].

We present an experiment demonstrating the impact of L3 errors on ML model performance. This experiment simulates temporal correlation leakage in EEG. We preprocessed the EEGBCI dataset [29] by bandpass filtering and epoching. Next, we assigned binary classification labels to the EEG epochs using two methods. The first labeling method assigned labels randomly for each sample. Our second method divided the samples into alternating fixed-length buckets so neighboring samples would likely be assigned the same class. Although the class labels are randomly assigned, this method emulates an EEG experiment with a block design for stimulus presentation. Lastly, we evaluated the performance of logistic regression classifiers $\mathrm{L_A}$ (no leakage) and $\mathrm{L_B}$ (leakage). This process was repeated 20 times, after which the mean accuracy of each classifier was reported. Classifier $\mathrm{L_B}$ outperforms classifier $\mathrm{L_A}$. The mean accuracy of classifier $\mathrm{L_A}$ was 0.50, while classifier $\mathrm{L_B}$ attained an accuracy measure of 0.88. Fig. 2b illustrates the impact of the L3 error on the classification performance.

Data leakage, occurring in experimental or preprocessing stages, detrimentally affects downstream ML model performance. Thus, vigilant monitoring of data leakage risks is crucial throughout the data analysis pipeline. Recommendations to mitigate such leakage in EEG ML research follow.

2.1.4 Avoiding Data Leakage. As discussed in Section 2.1.3, temporal correlations can undermine reproducibility when experimental stimuli are presented in a block fashion [15, 38]. This issue is especially problematic given the availability of neuroimaging datasets that are collected in a block fashion, and can lead to potential data contamination [15]. There are two main paths to remedy this problem. First, we recommend adopting a reporting checklist and highlighting efforts taken to avoid data contamination if the dataset was collected in a block fashion. Second, at the experimental design level it has been suggested to adopt other data collection methods, such as rapid event designs that allow stimuli randomization to avoid the block-level temporal correlations.

Researchers can avoid data leakage and improve the trustworthiness and reproducibility of EEG ML science by adopting rigorous standardized documentation practices. The Reporting Standards For Machine Learning Based Science (REFORMS) checklist was first proposed by Kapoor et al. [2023] to address invalid and irreproducible results in ML research [7]. This checklist is a practical guide for conducting transparent and reproducible ML-based research, addressing validity failures across diverse ML research fields. The REFORMS checklist is a valuable tool that covers all facets of a scientific study, from study design to limitations. We advocate for the widespread use of the REFORMS checklist in all EEG-ML research studies to aid in the promotion of reproducibility and prevention of validity issues such as data leakage.

The REFORMS checklist is based on three foundational goals for robust ML-based science. The first goal is establishing scientific claims. Scientific claims should be clearly articulated and linked to the ML task. Ensuring correct execution is the second goal, which involves verifying that the ML task is correctly performed by thoroughly documenting all design and performance aspects. The final goal is to enable independent verification of results, a critical aspect of transparency and credibility. Adoption of the REFORMS checklist can have extensive benefits for EEG ML researchers, referees, and journals.

2.2 Data Scarcity

Data scarcity is a primary reproducibility risk in EEG-ML studies. Deep learning methods show promise for enhancing EEG diagnostic tools, but their effectiveness relies on large training datasets [13]. Though EEG is cost-effective compared to other neuroimaging methods, data collection remains resource-intensive, resulting in smaller datasets, especially in motor imagery tasks [4]. Critically, EEG data scarcity has a negative downstream effect on model evaluation accuracy.

A lack of available data in EEG ML analysis contributes to the "selective inference" problem in statistics [33]. In selective inference, a dataset is mined to determine the strongest associations, for example, between EEG features and an outcome variable. Then, the same dataset is reused to assess the significance and effect sizes of the mined associations. When the effect of the selection process is not accounted for, it becomes challenging to assess the strength of the mined associations [33]. Therefore, reported claims tend to be overoptimistic due to a selective inference bias. Overoptimistic performance claims due to selective inference bias are extremely common across the ML neuroimaging literature [36].

Selective inference bias undermines reproducibility in EEG ML model evaluation. Small EEG datasets are prone to overfitting, in part due to extensive redundancies seen across sensor channels [19]. Dataset reuse in feature selection and subsequent model selection can increase the risk of overfitting [33]. Limited sample sizes increase the variability of performance estimates using ML methods [34]. This increased variability can cause over-optimistic reported accuracies, particularly when paired with insufficient or flawed validation methods [30].

When building an ML analysis pipeline, researchers use the same dataset repeatedly when making critical design decisions [1]. Dataset reuse in this manner leads to overoptimistic ML model performance estimates [33]. This issue is particularly pronounced for small sample sizes and high-dimensional data, including EEG [1, 4].

We present an experimental demonstration of selective inference bias in EEG motor imagery classification. Using the EEGBCI dataset [29], we performed basic preprocessing steps, including bandpass filtering, epoching, and feature selection. We trained and evaluated 100 logistic regression classifiers, where each classifier utilized a unique training and test data partition. The maximum test accuracy obtained using this method was 0.88. The lowest test accuracy obtained was 0.33. Meanwhile, the mean test accuracy was 0.62. These results are illustrated in Fig. 3. Without robust validation methods, EEG ML researchers may overestimate model performance due to this level of variability [30].

The high variability of EEG ML model performance highlights the joint issues of selective inference and publication biases [1, 8].

In a scenario where multiple research laboratories perform the same experiment with unique random train-test splits, only those who attain high accuracy are likely to publish their findings [36]. Selective inference and publication bias can jeopardize the reproducibility of EEG ML results and contribute to over-optimism across the research community [8].

2.2.1 *Working with Limited Data.*

A lack of available data is a critical issue affecting the reliability and reproducibility of EEG-ML science. Open access data sources, data augmentation methods, and transfer learning methods are three options for EEG-ML researchers to address data scarcity. This section will discuss how these approaches can enable EEG-ML researchers to produce reliable and reproducible scientific results.

Open access data is a boon to reproducible EEG ML research as it enables researchers to evaluate the benefits and drawbacks of varied ML approaches against substantially sized and freely available datasets. EEG ML researchers can utilize these open-access data sources to directly bolster the transparency and computational reproducibility of their results [8, 27]. However, many research questions will necessitate the use of smaller EEG datasets. In these scenarios, data augmentation and transfer learning methods can help mitigate the impact of the limited sample size.

Data augmentation for ML involves artificially expanding a dataset by applying various transformations to the existing data, enhancing model robustness and generalization. EEG data augmentation techniques include averaging trials, recombining time and frequency slices of trials, adding noise, sliding window cropping, and generating synthetic data [4]. EEG-ML researchers may benefit from data augmentation when analyzing datasets with few samples. However, data augmentation techniques also have some drawbacks. Augmentation techniques that have been shown to be effective for specific tasks or datasets often do not transfer well to other datasets or tasks [2]. Traditional augmentation methods for time-series data are based on making modifications to elements of the real dataset, which can often cause them to generate invalid or lower quality examples [5]. Finally, deep learning-based augmentation methods such as generative adversarial networks are highly complex, making it difficult to train and obtain results.

EEG-ML researchers may address data scarcity using transfer learning. In transfer learning, knowledge gained from training a model on one task is used to improve performance on a related but distinct task. This approach has garnered significant attention in self-supervised representation learning for EEG [23]. Representation learning automatically discovers and creates meaningful representations or features from raw data. Transforming highly complex EEG data into a compact, tractable, and informative feature representation is a primary challenge in EEG-ML. Self-supervised representation learning methods can help overcome the limitations imposed by smaller datasets by leveraging a pre-trained learned representation of EEG data to perform a downstream task.

There are data leakage risks in data augmentation and transfer learning. Lee et al. [2023] explore the risks of data leakage in data augmentation for EEG computer-aided diagnosis systems [13]. When cropped segments of EEG trials are placed into training and test sets, the same source EEG trial is utilized as training and test data, and L1 data leakage occurs. To prevent this leakage, EEG ML researchers must perform data augmentation only after proper train-test splitting such that these samples remain separated [13].

Feature leakage is a form of data leakage that can occur when applying transfer learning methods [28]. One example of feature leakage is when feature selection is informed by test performance during the transfer learning process. A researcher can reuse a test set numerous times during transfer learning for feature extraction and alter their methods until they achieve a satisfactory performance level [1, 28]. For example, a developer could adjust the number of frozen layers in a pre-trained feature extractor until test accuracy reaches the desired level. This is a variant of the selective inference problem discussed in Section 2.2; the test set has effectively become part of the validation set in this situation. Feature leakage is a particular risk when performing transfer learning on limited datasets, as with EEG transfer learning [28]. We recommend that researchers set aside a test set to evaluate model generalizability that is strictly isolated during the transfer learning process to mitigate feature leakage risks.

2.3 Flawed Validation Methods

In ML research, model validation significantly impacts the reproducibility of the reported results. We will focus our reproducibility discussion on cross-validation (CV). CV is a technique used to assess the performance of different ML models and their hyperparameters. Absent or flawed validation methods can undermine EEG ML reproducibility in two primary ways. Firstly, as discussed in Section 2.2, EEG ML analysis exhibits high performance variability due to limited sample sizes. When data is limited, EEG ML models are prone to overfitting. This scenario leads to overoptimistic performance estimates, particularly when researchers do not apply CV. CV methods can also introduce biased model assessments and compromise result validity when incorrectly applied due to data leakage risks. Even when applied correctly, CV measures of predictive accuracy in neuroimaging can be artificially inflated due to high variance in the prediction score [36].

In the field of neuropsychiatric disease prediction from neuroimaging data, biased estimates of predictive accuracy are a common error. Poldrack et al. [2020] identifies several factors for this issue, among them is the introduction of bias via misapplied CV methods [24].

Absent validation methods are a significant reproducibility risk factor in EEG ML studies. A study on the prediction performance of neuropsychiatric EEG biomarkers demonstrated how overfitting due to selection bias can cause overoptimistic results when CV is not applied [31]. In this study, feature selection was performed using real and simulated EEG data, and the prediction accuracy of the resulting features was evaluated both with and without CV methods. The non-CV results were found to be significantly higher than the CV results. This result shows that CV methods are essential to obtain robust estimates of EEG ML model performance. EEG ML researchers often fail to apply robust validation methods when estimating model performance. A meta-analysis by Roy et al. [2019] found that of 154 surveyed EEG deep learning studies, 42% did not apply CV [27].

CV methods are pivotal in reproducible model selection for EEG studies. As a technique to evaluate various ML models and their

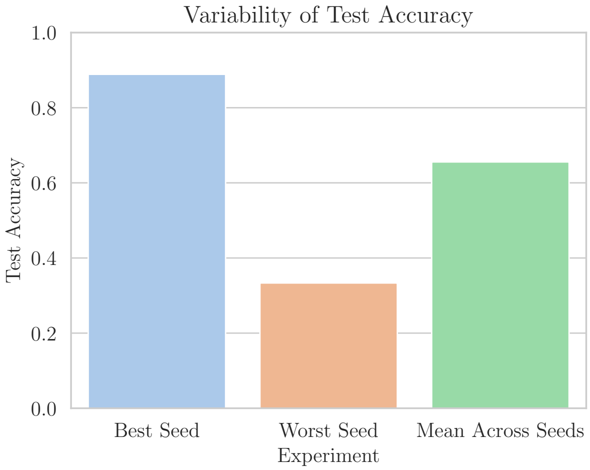


Figure 3: Measurement of test accuracy when EEG features are randomly split into training and test datasets. Logistic regression classifiers were trained on 100 unique train-test splits of the EEGBCI subset. We compare the group maximum, minimum, and mean test accuracies.

hyperparameters, CV allows researchers to balance model performance while minimizing overfitting and over-optimism. Despite these advantages, CV is not a panacea for reproducibility issues in EEG ML. Methodological flaws can jeopardize the validity of model assessments, even when CV is applied [8, 30, 38]. For example, when data is limited, K-Fold CV can yield overoptimistically biased performance estimates [34]. Inadequate data partitioning can undermine reproducibility by causing data leakage [13]. The next section discusses developing a correct CV strategy to avoid common reproducibility pitfalls.

2.3.1 Cross-validation in EEG-based ML. Incorrect application of CV methods can lead to over-optimistic estimates of ML model performance. This section provides recommendations to help define an appropriate CV strategy for EEG-based ML. We focus our discussion on data partitioning approaches and evaluation criteria. For additional information on neuroimaging predictive modeling best practices, researchers are encouraged to refer to the work by Poldrack et al. [2020] [24].

When applying ML techniques to EEG data, researchers are often interested in a classification task of some form. In this context, we recommend that researchers consider the downstream goals of the ML model being trained and evaluated. For example, an ML diagnostic system is only useful if the classification capabilities generalize well to unseen participants. To correctly evaluate the generalization performance of an ML model in this context, a leave-one-subject-out (LOSO) CV strategy is often recommended [11, 30]. Using LOSO validation, recordings from individual subjects are iteratively omitted from the training set. However, the left-out subject data is still used to evaluate the trained model during each iteration, and the final performance estimate is obtained by averaging over these iterations. By contrast, randomized K-Fold CV methods allow subject trials to appear in both training and test sets. LOSO CV performance estimates may exhibit lower bias than randomized CV methods [11, 30]. The LOSO CV scenario more closely resembles the real-world use case of such a model, wherein a previously unseen subject's data must be classified into one of the classes.

EEG ML researchers may be interested in obtaining a performance estimate using data from a single subject in non-diagnostic contexts. For example, a study may seek to determine the feasibility of a given ML task or whether model performance differs among groups of experimental subjects and healthy controls. Subject-wise CV methods may be appropriate for these scenarios [8]. For subject-wise CV, EEG data from a single subject is divided into training and evaluation sets. This process is applied iteratively, so different samples from the subject's data are used for training and evaluation at each step. A final performance estimate can be constructed by determining the mean across subjects. The development of a CV strategy must be informed by the specific characteristics of the EEG dataset and the ML task at hand. We recommend that researchers experiment with different approaches to assess their impact on model performance. As there is no universal CV solution, tailoring CV strategies to fit the EEG dataset is necessary to ensure that ML model results are reliable and reproducible. Transparent documentation of the chosen CV strategy can improve reproducibility in EEG ML. Researchers should document the CV strategy applied, including any hyperparameter tuning details, in the EEG ML model card (Section 3.1).

3 TOWARDS A SOLUTION: A FRAMEWORK FOR REPRODUCIBILITY IN EEG-BASED ML METHODS

In this section, we present our framework for achieving reproducibility in EEG-based ML studies. A first step towards ensuring that EEG-based ML studies are reproducible is establishing community-wide standards for transparency in reporting procedures and methods. In service of this goal, we present a Model Card tailored to EEG ML applications. We designed this model card to foster reproducible EEG ML research by mitigating data leakage and scarcity issues and promoting correct model validation practices.

3.1 Model Cards for EEG ML

One issue with the ML-based science literature is that model developers focus on improving "accuracy" during model training, with less attention paid to experimental design for data collection and potential consequences post deployment. We posit that model developers should consider a broader view of model design, one informed by the data collection methodology and other metadata describing the sample. Finding an approach for operationalizing this has been an ongoing concern of the ML community. For instance, Rostamzadeh et al. [2022] suggest a guideline to provide dataset insights for developers [26]. Our solution is optimized for EEG-based applications, as these guidelines do not currently exist.

We adopt the concept of Model Cards, first introduced by Mitchell et al. [2019], to facilitate standardization in reporting practices for ML technology [21]. Model cards are short documents accompanying released ML models, providing usage context, performance benchmarks, and other relevant information. They provide valuable contextual information regarding an ML model's use case, capabilities, and performance. Proper use of model cards can enhance reproducibility in EEG-based ML research by enabling researchers to precisely identify a public model's characteristics.

Model cards are a familiar tool for many ML researchers, having already seen widespread adoption among the ML community. A systematic analysis of over 32,000 model cards posted on the Hugging Face platform revealed that the 44% of models with corresponding model cards account for over 90% of total download traffic [16]. Additionally, the inclusion of detailed model card documentation for previously undocumented models had a positive influence on model utilization. The wide uptake of model card documentation suggests that the ML community recognizes the importance of model cards for facilitating model understanding and deployment [16]. Critically, model cards allow researchers to understand the limitations of an ML model. A greater understanding of ML model limitations can reduce the misuse of existing ML models or architectures [8, 21]. Additionally, thorough use of model card documentation can help researchers easily identify sources of data leakage or other reproducibility risks before publication. We propose this extension of model cards for EEG-based ML to aid transparency and reproducibility in this research area, addressing some potential pitfalls outlined herein. We additionally provide a sample EEG-ML model card to illustrate the usage of this extended framework.

Though the REFORMS checklist (Section 2.1.4) and our proposed Model Card for EEG aim to enhance ML research transparency, they adopt distinct roles in service of this objective. The REFORMS checklist is a field-agnostic guide emphasizing comprehensive reporting for all ML-based research. By contrast, our Model Card for EEG takes a focused approach explicitly tailored to EEG ML research. The REFORMS checklist can help researchers monitor reproducibility risks during experimental design and data analysis, while our model card facilitates public EEG ML model transparency. Lastly, the model card is geared towards concise model summaries, enabling rapid evaluation of public model capabilities.

A typical ML model card comprises sections related to model details, intended use, factors, metrics, evaluation data, training data, quantitative analyses, ethical considerations, and caveats and recommendations [21]. We focus on additional documentation for EEG-based ML models, assuming baseline information aligns with the original framework. To cover the entire EEG data analysis pipeline, we propose the addition of several new sections in the model card:

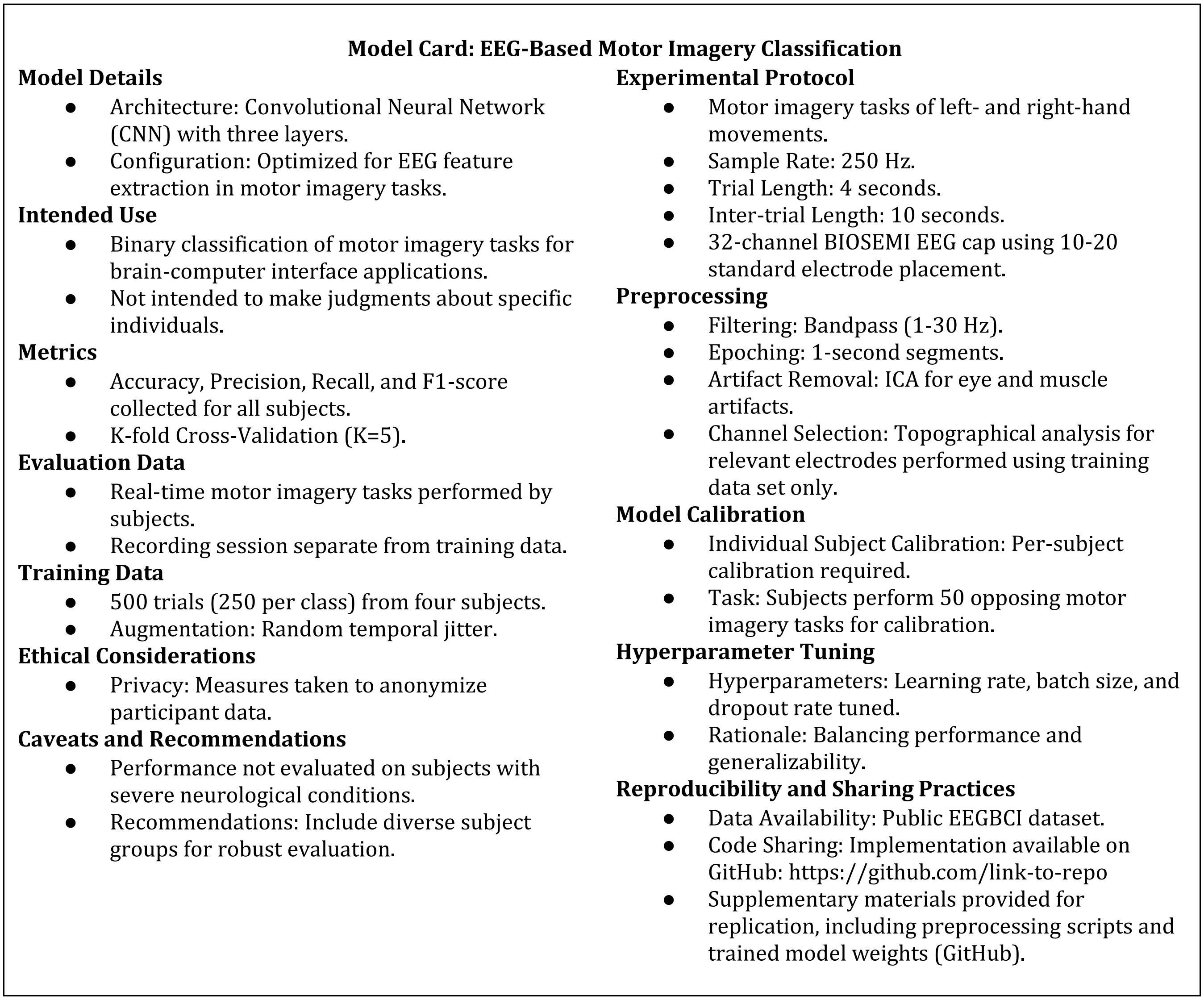
- Experimental protocol
  Summarize the EEG data collection process. This section includes information on the sample rate, behavioral task and stimuli details, trial length, and number of electrodes. Data collected alongside the EEG, such as electromyographic (EMG), audio, or other data types, can also be described in this section.
- Preprocessing
  Document all operations applied to EEG data after collection and before use in model training. Filter types and frequencies, epoching, artifact removal, channel rejection, and channel selection methods can all be included in the preprocessing section. Authors may justify preprocessing parameter selections and provide any additional information about the EEG data in this section.
- Hyperparameter tuning
  Specify the tuning process applied during model development. Hyperparameter tuning involves finding the best settings for an ML model, such as the learning rate or other control parameters. Hyperparameter tuning strategies applied to the model should be summarized here.
- Reproducibility and sharing practices
  Consider steps taken to ensure the reproducibility of the study. Reproducibility topics presented in this section include data availability, code sharing, and any additional resources to facilitate replication of a study's primary results.

We present a hypothetical EEG-based motor imagery classifier using the EEGBCI dataset [29]. Fig. 4 shows an example of the extended EEG ML model card.

The Model Card for EEG offers a systematic template for researchers to document essential model details. This standardized format helps to prevent reproducibility errors, such as data leakage and selective inference, by providing a consistent framework for model assessment. Researchers can efficiently compare EEG ML models and approaches using model cards, allowing for the identification and mitigation of potential risks. The streamlined presentation of crucial reproducibility details is the primary strength of the Model Card for EEG. Greater emphasis on standard reporting practices will promote collaboration and knowledge sharing within the EEG ML research community and foster reliable and transparent advancement of the field.

3.1.1 *Limitations.*

Although model cards are valuable for transparency and accountability in machine learning, they have limitations. They are limited in scope and may not capture all relevant information about complex models. Model cards provide a snapshot of a model's characteristics at a particular time and may not reflect changes or updates as a model is deployed in different contexts. Therefore, model developers must invest time to maintain public



Model Card: EEG-Based Motor Imagery Classification

Model Details

- Architecture: Convolutional Neural Network (CNN) with three layers.
- Configuration: Optimized for EEG feature extraction in motor imagery tasks.

Intended Use

- Binary classification of motor imagery tasks for brain-computer interface applications.
- Not intended to make judgments about specific individuals.

Metrics

- Accuracy, Precision, Recall, and F1-score collected for all subjects.
- K-fold Cross-Validation (K=5).

Evaluation Data

- Real-time motor imagery tasks performed by subjects.
- Recording session separate from training data.

Training Data

- 500 trials (250 per class) from four subjects.
- Augmentation: Random temporal jitter.

Ethical Considerations

- Privacy: Measures taken to anonymize participant data.

Caveats and Recommendations

- Performance not evaluated on subjects with severe neurological conditions.
- Recommendations: Include diverse subject groups for robust evaluation.

Experimental Protocol

- Motor imagery tasks of left- and right-hand movements.
- Sample Rate: 250 Hz.
- Trial Length: 4 seconds.
- Inter-trial Length: 10 seconds.
- 32-channel BIOSEMI EEG cap using 10-20 standard electrode placement.

Preprocessing

- Filtering: Bandpass (1-30 Hz).
- Epoching: 1-second segments.
- Artifact Removal: ICA for eye and muscle artifacts.
- Channel Selection: Topographical analysis for relevant electrodes performed using training data set only.

Model Calibration

- Individual Subject Calibration: Per-subject calibration required.
- Task: Subjects perform 50 opposing motor imagery tasks for calibration.

Hyperparameter Tuning

- Hyperparameters: Learning rate, batch size, and dropout rate tuned.
- Rationale: Balancing performance and generalizability.

Reproducibility and Sharing Practices

- Data Availability: Public EEGBCI dataset.
- Code Sharing: Implementation available on GitHub: https://github.com/link-to-repo
- Supplementary materials provided for replication, including preprocessing scripts and trained model weights (GitHub).

Figure 4: Sample EEG-ML Model Card for an EEGBCI Motor Imagery classification model.

model card information. Useful model cards are populated with attention to all aspects of a model, including thoughtful reflection on a model's limitations. However, researchers and developers tend to downplay model limitations. Liang et al. [2024] notes that the Environmental Impact, Evaluation, and Limitations model card sections are frequently omitted by developers [16]. Model cards may provide a false sense of transparency or accountability to organizations, as model cards are less interpretable to those without significant ML expertise [35]. Finally, there is no enforcement of a standard format or template for model cards. This limitation can make it difficult to evaluate competing models based on model cards alone.

Model cards are not a complete solution to reproducibility issues in EEG ML, but they are an important part of a more complete solution. Defining best practices for model card design and implementation will be necessary across ML research disciplines to support transparent, usable, and responsible ML [16]. Researchers should evaluate and apply additional transparency tools and approaches beyond model cards to improve reproducibility, including algorithmic auditing, adversarial testing, and inclusive user feedback mechanisms [21]. Despite their limitations, model cards remain a valuable tool for improving model transparency and reproducibility. Most importantly, model cards directly involve developers with reproducibility issues during model development, decreasing the likelihood of common reproducibility errors. This paper highlights the prevalence of ML pitfalls in the neural engineering community and acknowledges the reproducibility crisis in this field. We believe that proposing a model card tailored for EEG signals is the first step toward finding a standard and widely accepted guideline in this area of research to elevate transparency and achieve the safe deployment of these models.

4 CONCLUSION

EEG data analysis through ML methods is challenging but essential for decoding brain signals. Recent studies have claimed high performance, gaining widespread attention. Yet, reproducibility issues have arisen. This paper addresses three key reproducibility challenges in EEG-based ML analysis: data leakage, data scarcity, and flawed model selection. We introduce the Model Card for EEG, a documentation tool promoting transparency and standardization in EEG model reporting. Our framework aims to mitigate data leakage, leverage limited data effectively using strategies like data augmentation and transfer learning, and guide the development of appropriate cross-validation methods. By doing so, we seek to improve the reliability and credibility of EEG ML research, fostering trustworthiness and reproducibility in the field.

ACKNOWLEDGMENTS

This research was supported in part by grants from the National Institutes of Health, including NIH-NIDCD R01DC006859 to V. Berisha and J. Liss, and grants R01-DC019905 and R01-DC020162 to A. Daliri. The funders played no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

[1] Visar Berisha, Chelsea Krantsevich, Gabriela Stegmann, Shira Hahn, and Julie Liss. 2022. Are reported accuracies in the clinical speech machine learning literature overoptimistic?. In *Proc. Interspeech 2022*. 2453–2457. https://doi.org/10.21437/Interspeech.2022-691

[2] Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. *Transactions of the Association for Computational Linguistics* 11 (March 2023), 191–211. https://doi.org/10.1162/tacl_a_00542

[3] Stephane Doyen and Nicholas B. Dadario. 2022. 12 Plagues of AI in Healthcare: A Practical Guide to Current Issues With Using Machine Learning in a Medical Context. *Frontiers in Digital Health* 4 (May 2022). https://doi.org/10.3389/fdgth.2022.765406 Publisher: Frontiers.

[4] Olawunmi George, Roger Smith, Praveen Madiraju, Nasim Yahyasoltani, and Sheikh Iqbal Ahamed. 2022. Data augmentation strategies for EEG-based motor imagery decoding. *Heliyon* 8, 8 (Aug. 2022). https://doi.org/10.1016/j.heliyon.2022.e10240 Publisher: Elsevier.

[5] Guillermo Iglesias, Edgar Talavera, Ángel González-Prieto, Alberto Mozo, and Sandra Gómez-Canaval. 2023. Data Augmentation techniques in time series domain: a survey and taxonomy. *Neural Computing and Applications* 35, 14 (May 2023), 10123–10145. https://doi.org/10.1007/s00521-023-08459-3

[6] John P. A. Ioannidis. 2005. Why Most Published Research Findings Are False. *PLoS Medicine* 2, 8 (Aug. 2005), e124. https://doi.org/10.1371/journal.pmed.0020124

[7] Sayash Kapoor, Emily Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, Jessica Hullman, Michael A. Lones, Momin M. Malik, Priyanka Nanayakkara, Russell A. Poldrack, Inioluwa Deborah Raji, Michael Roberts, Matthew J. Salganik, Marta Serra-Garcia, Brandon M. Stewart, Gilles Vandewiele, and Arvind Narayanan. 2023. REFORMS: Reporting Standards for Machine Learning Based Science. https://doi.org/10.48550/arXiv.2308.07832 arXiv:2308.07832 [cs, stat].

[8] Sayash Kapoor and Arvind Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns (New York, N.Y.)* 4, 9 (Sept. 2023), 100804. https://doi.org/10.1016/j.patter.2023.100804

[9] Nikolaus Kriegeskorte, W. Kyle Simmons, Patrick S. F. Bellgowan, and Chris I. Baker. 2009. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience* 12, 5 (May 2009), 535–540. https://doi.org/10.1038/nn.2303

[10] G. R. Kiran Kumar and M. Ramasubba Reddy. 2020. Correction to "Designing a Sum of Squared Correlations Framework for Enhancing SSVEP Based BCIs". *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28, 4 (April 2020), 1044–1045. https://doi.org/10.1109/TNSRE.2020.2974271

[11] Sajeev Kunjan, T. S. Grummett, K. J. Pope, D. M. W. Powers, S. P. Fitzgibbon, T. Bastiampillai, M. Battersby, and T. W. Lewis. 2021. The Necessity of Leave One Subject Out (LOSO) Cross Validation for EEG Disease Diagnosis. In *Brain Informatics*, Mufti Mahmud, M. Shamim Kaiser, Stefano Vassanelli, Qionghai Dai, and Ning Zhong (Eds.). Springer International Publishing, Cham, 558–567. https://doi.org/10.1007/978-3-030-86993-9_50

[12] Luca La Fisca, Virginie Vandenbulcke, Erika Wauthia, Aurélie Miceli, Isabelle Simoes Loureiro, Laurence Ris, Laurent Lefebvre, Bernard Gosselin, and Cyril R. Pernet. 2022. Biases in BCI experiments: Do we really need to balance stimulus properties across categories? *Frontiers in Computational Neuroscience* 16 (Nov. 2022). https://doi.org/10.3389/fncom.2022.900571 Publisher: Frontiers.

[13] Hyung-Tak Lee, Hye-Ran Cheon, Seung-Hwan Lee, Miseon Shim, and Han-Jeong Hwang. 2023. Risk of data leakage in estimating the diagnostic performance of a deep-learning-based computer-aided system for psychiatric disorders. *Scientific Reports* 13, 1 (Oct. 2023), 16633. https://doi.org/10.1038/s41598-023-43542-8 Publisher: Nature Publishing Group.

[14] Émile Lemoine, Joel Neves Briard, Bastien Rioux, Oumayma Gharbi, Renata Podbielski, Bénédicte Nauche, Denahin Toffa, Mark Keezer, Frédéric Lesage, Dang K. Nguyen, and Elie Bou Assi. 2024. Computer-assisted analysis of routine EEG to identify hidden biomarkers of epilepsy: A systematic review. *Computational and Structural Biotechnology Journal* 24 (Dec. 2024), 66–86. https://doi.org/10.1016/j.csbj.2023.12.006 Publisher: Elsevier.

[15] Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. 2021. The Perils and Pitfalls of Block Design for EEG Classification Experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1 (Jan. 2021), 316–333. https://doi.org/10.1109/TPAMI.2020.2973153

[16] Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. 2024. What's documented in AI? Systematic Analysis of 32K AI Model Cards. https://doi.org/10.48550/arXiv.2402.05160 arXiv:2402.05160 [cs] version: 1.

[17] Klaus Linkenkaer-Hansen, Vadim V. Nikouline, J. Matias Palva, and Risto J. Ilmoniemi. 2001. Long-Range Temporal Correlations and Scaling Behavior in Human Brain Oscillations. *Journal of Neuroscience* 21, 4 (Feb. 2001), 1370–1377. https://doi.org/10.1523/JNEUROSCI.21-04-01370.2001 Publisher: Society for Neuroscience Section: ARTICLE.

[18] F. Lopes da Silva. 1991. Neural mechanisms underlying brain waves: from neural membranes to networks. *Electroencephalography and Clinical Neurophysiology* 79, 2 (Aug. 1991), 81–93. https://doi.org/10.1016/0013-4694(91)90044-5

[19] Jing Luo, Yaojie Wang, Rong Xu, Guangming Liu, Xiaofan Wang, and Yijing Gong. 2021. Channel Drop Out: A Simple Way to Prevent CNN from Overfitting in Motor Imagery Based BCI. In *Data Science*, Jianchao Zeng, Pinle Qin, Weipeng Jing, Xianhua Song, and Zeguang Lu (Eds.). Springer Nature, Singapore, 443–452. https://doi.org/10.1007/978-981-16-5940-9_34

[20] Matthew B. A. McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. 2021. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine* 13, 586 (March 2021), eabb1655. https://doi.org/10.1126/scitranslmed.abb1655

[21] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[22] Masaki Nakanishi, Minpeng Xu, Yijun Wang, Kuan-Jung Chiang, Jin Han, and Tzyy-Ping Jung. 2020. Questionable Classification Accuracy Reported in "Designing a Sum of Squared Correlations Framework for Enhancing SSVEP-Based BCIs". *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28, 4 (April 2020), 1042–1043. https://doi.org/10.1109/TNSRE.2020.2974272 Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering.

[23] Andi Partovi, Anthony N. Burkitt, and David Grayden. 2023. A Self-Supervised Task-Agnostic Embedding for EEG Signals. In *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*. 1–4. https://doi.org/10.1109/NER52421.2023.10123767

[24] Russell A. Poldrack, Grace Huckins, and Gael Varoquaux. 2020. Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA psychiatry* 77, 5 (May 2020), 534–540. https://doi.org/10.1001/jamapsychiatry.2019.3671

[25] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 959–972. https://doi.org/10.1145/3531146.3533158

[26] Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. 2022. Healthsheet: Development of a Transparency Artifact for Health Datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1943–1961. https://doi.org/10.1145/3531146.3533239

[27] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H. Falk, and Jocelyn Faubert. 2019. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering* 16, 5 (Aug. 2019), 051001. https://doi.org/10.1088/1741-2552/ab260c Publisher: IOP Publishing.

[28] Ravi K. Samala, Heang-Ping Chan, Lubomir Hadjiiski, and Mark A. Helvie. 2021. Risks of feature leakage and sample size dependencies in deep feature extraction

for breast mass classification. *Medical Physics* 48, 6 (June 2021), 2827–2837. https://doi.org/10.1002/mp.14678

[29] Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. 2004. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE transactions on bio-medical engineering* 51, 6 (June 2004), 1034–1043. https://doi.org/10.1109/TBME.2004.827072

[30] Sina Shafiezadeh, Gian Marco Duma, Giovanni Mento, Alberto Danieli, Lisa Antoniazzi, Fiorella Del Popolo Cristaldi, Paolo Bonanni, and Alberto Testolin. 2023. Methodological Issues in Evaluating Machine Learning Models for EEG Seizure Prediction: Good Cross-Validation Accuracy Does Not Guarantee Generalization to New Patients. *Applied Sciences* 13, 7 (Jan. 2023), 4262. https://doi.org/10.3390/app13074262 Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

[31] Miseon Shim, Seung-Hwan Lee, and Han-Jeong Hwang. 2021. Inflated prediction accuracy of neuropsychiatric biomarkers caused by data leakage in feature selection. *Scientific Reports* 11, 1 (April 2021), 7980. https://doi.org/10.1038/s41598-021-87157-3 Publisher: Nature Publishing Group.

[32] Maia Szalavitz. 2021. The Pain Was Unbearable. So Why Did Doctors Turn Her Away? *Wired* (Aug. 2021). https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/ Section: tags.

[33] Jonathan Taylor and Robert J. Tibshirani. 2015. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences* 112, 25 (June 2015), 7629–7634. https://doi.org/10.1073/pnas.1507583112 Publisher: Proceedings of the National Academy of Sciences.

[34] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J. Casson. 2019. Machine learning algorithm validation with a limited sample size. *PLOS ONE* 14, 11 (Nov. 2019), e0224365. https://doi.org/10.1371/journal.pone.0224365 Publisher: Public Library of Science.

[35] Cristina Vanberghen. 2023. AI Act: Model cards and 'The Emperor's New Clothes'? https://www.euractiv.com/section/artificial-intelligence/opinion/ai-act-model-cards-and-the-emperors-new-clothes/ Section: Artificial Intelligence.

[36] Gaël Varoquaux. 2018. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* 180, Pt A (Oct. 2018), 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

[37] Rutger J. Vlek, David Steines, Dyana Szibbo, Andrea Kübler, Mary-Jane Schneider, Pim Haselager, and Femke Nijboer. 2012. Ethical Issues in Brain–Computer Interface Research, Development, and Dissemination. *Journal of Neurologic Physical Therapy* 36, 2 (June 2012), 94. https://doi.org/10.1097/NPT.0b013e31825064cc

[38] Joseph West, Zahra Dasht Bozorgi, Jeffrey Herron, Howard J. Chizeck, Jordan D. Chambers, and Lyra Li. 2023. Machine learning seizure prediction: one problematic but accepted practice. *Journal of Neural Engineering* 20, 1 (Jan. 2023). https://doi.org/10.1088/1741-2552/acae09

[39] Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. 2021. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine* 181, 8 (Aug. 2021), 1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626

[40] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine* 15, 11 (Nov. 2018), e1002683. https://doi.org/10.1371/journal.pmed.1002683

Received 22 January 2024; revised 20 March 2024; accepted 30 March 2024