Biomedical Signal Processing and Control 120 (2026) 110020

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

What is the across-subject performance of your Artificial Intelligence method for EEG analysis?

Debora Gil, Aura Hernández-Sabaté *

Computer Vision Center, C/ de les Sitges, Edifici O, Cerdanyola del Vallès, 08193, Spain
Computer Science Department, Universitat Autònoma de Barcelona, C/ de les Sitges, Edifici Q, Cerdanyola del Vallès, 08193, Spain

ARTICLE INFO

Keywords:
Electroencephalography
Machine learning
Deep learning
Cross-subject reproducibility
Statistical validation
Generalized mixed-effects models
Model personalization
Seizure detection
Hierarchical Validation
Personalized EEG Models

ABSTRACT

Reproducibility is a critical challenge in the application of Artificial Intelligence (AI) to electroencephalography (EEG) data for neurological function recognition. Data leakage in cross-validation of methods leads to over-optimistic performance metrics and uncertain across-subject reproducibility of results.

This study introduces a hierarchical statistical validation framework based on Generalized Mixed-Effects Models (GMMs) to rigorously assess the reproducibility and generalization capabilities of machine learning (ML) and deep learning (DL) models. To show its potential six neural network architectures for the fusion of multi-channel EEG data were evaluated on the Children's Hospital Boston and the Massachusetts Institute of Technology, CHB-MIT Scalp EEG Database using a cross-validation at three different stratified validation levels: window, state, and subject.

Results show that window-level metrics are over-optimistic relative to leakage-free partitions: Sensitivity (Recall+) and Specificity (Recall-) of 0.96 ± 0.00 and 1.00 ± 0.00, respectively, at the window level drop to 0.77 ± 0.22 and 0.90 ± 0.17, respectively, at subject level. The framework also enables identification of statistically distinct subpopulations, underscoring the impact of neurophysiological heterogeneity on model performance.

1. Introduction

Neurological function can be affected by a variety of conditions, including epilepsy, traumatic brain injuries, neurodegenerative diseases, psychiatric disorders and psychological or physiological factors (e.g. cognitive workload, divided attention, or stress). These conditions manifest as changes in brain electrical activity measurable with electroencephalography (EEG). Since the interpretation of EEG data is complex and expertise-dependent, Artificial Intelligence (AI) methods can play a crucial role. By leveraging machine learning (ML) algorithms and pattern recognition techniques, AI can assist in the analysis of EEG data, helping to identify subtle changes in neurological function linked to neurological or psychological states (like seizures or cognitive workload).

In the last two decades, there has been an increase in the development of machine learning and deep learning (DL) systems for the detection and prediction of neurological function from EEG data. Since EEG signals are naturally time series, temporal features analysis of these signals are crucial for accurate neurological function classification. In this fashion, most approaches classify the different behaviors of electrical activity using, as input, temporal windows of few seconds extracted from each state (e.g. epileptic seizure, working memory demand). Despite advancements in DL for EEG analysis, limited attention has been paid to rigorous reproducibility assessments across diverse neurological contexts [1,2]. In response, we propose a statistical framework using Generalized Mixed-Effects Models (GMMs) [3] to assess the generalization of models and, thus, to verify their success in predicting new unseen cases.

1.1. State-of-the-art

A crucial step in automatic neurological function classification is the proper validation of the technique used to ensure the ability of generalization of the model across unseen subjects, sessions, states or tasks [4–7].

Cross-validation techniques are widely used to evaluate model performance and can be categorized based on the experimental goal: k-fold (or repeated k-fold) and leave-one-out (or leave-p-out) methods. However, the explanations of the experimental design in the majority of the revised works are fuzzy and confusing in the terminology used to describe data splitting [1]. Still, a critical consideration is that the data

* Corresponding author at: Computer Science Department, Universitat Autònoma de Barcelona, C/ de les Sitges, Edifici Q, Cerdanyola del Vallès, 08193, Spain.
E-mail addresses: debora@cvc.uab.cat, debora.gil@uab.cat (D. Gil), aura@cvc.uab.cat, aura.hernandez@uab.cat (A. Hernández-Sabaté).

https://doi.org/10.1016/j.bspc.2026.110020
Received 24 July 2025; Received in revised form 3 February 2026; Accepted 1 March 2026
Available online 5 March 2026
1746-8094/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

splitting must align with the classification problem being addressed and avoid data leakage to ensure maximum reproducibility [5]. The inherent variability across subjects, states, and sessions necessitates treating these factors as multi-dimensional variables, directly linked to the study's research objectives and the classification challenges. Specifically, these strategies impact the level of model personalization and the evaluation of its generalization capabilities [8].

A k-fold split of all windows considers the window as sampling unit, regardless of the subject or state. Since all windows are considered, the train and test sets are not independent and share windows from the same patient and state intervals. This validation scheme, which is the most common in the literature [9–15], achieves the highest scores at the cost of the lowest generalization power and reproducibility of results in new unseen data, even do not translate into the clinical setting [16].

Leave-one-out cross-validation techniques use the "one-sample-out" as sampling unit. For example, in case of a leave-one-state-out, state intervals are the sampling units used to train and test data. Models are trained using windows from some state intervals belonging to a group of subjects and tested on new unseen intervals from the same set of patients [17,18]. The same procedure is computed in the case of leave-one-session-out [19] where the model is trained in one context and is tested in another one. In this way, the capability to learn a specific cohort of individuals is validated. Although it is not guaranteed a high performance in unseen individuals, this is a valid option from the point of view of the clinical application.

Finally, leave-one-subject-out assesses models using subjects different from the ones used for training. This is the most generalist and realistic assessment of across-subject generalization, and it is still the most challenging since it typically yields the lowest metrics [20–25].

Although the different choices can be found in the literature, most authors often only work with one of the options and specifically do not investigate the level of generalization. Since the metrics evaluating the performance of models are highly sensitive to data splitting, they cannot be fairly compared. The few works approaching different levels of validation show a decrease in performance between cross-validation with a split at window level and a leave-one-patient-out scheme. This also holds even for most recent DL models, as the comparison of several SoA approaches (including a transformer-based [25], a CNN squeeze-and-excitation modules, EEGNET [26], Wei-CNN [27], and CNN-LSTM [28]) using a leave-one-patient-out scheme on CHB-MIT cases shows. This is further supported by the results of the Neureka™ 2020 Epilepsy Challenge [29], for example, with a highest sensitivity on the evaluation independent set of 20%.

As far as we know, there is a lack of systematic studies of the actual across-subject generalization of models. Also, it is not clear what the best way is to incorporate into validation inter- and intra-subject variability. In this context, a recent study [8] emphasizes the need for subject-based cross validations to avoid over-optimistic metrics obtained in sample-based partitions in the validation of ML/DL for Neurological diseases detection (Alzheimer, Parkinson). Authors identify lack of statistical analysis as a main limitation of their study, being potential lack of normality in data a main reason for not using hypothesis tests. The use of GLM and the proposed subpopulation analysis relaxes the impact of the normality requirement and allows the analysis of non-normal distributions.

This paper contributes to a hierarchical statistical validation framework for the evaluation of performance reproducibility of AI models for the analysis of EEG signals. The proposed framework is tested on deep learning architectures as exemplars to demonstrate how validation choices affect reported performance and reproducibility. In particular we propose to use Generalized Mixed-Effects Models (GMMs) [3] to assess the models at different levels of generalization, taking into account inter- and intra- subject variability by means of random effects that allow to measure the reproducibility of quality scores. To do so, we structure temporal windows in a hierarchical population structure of states and patients and define different GMMs to upscale inference on performance across hierarchy levels. To demonstrate the validity and utility of our hierarchical statistical validation framework in a well-established benchmark task, our framework is applied to the assessment of the performance for epilepsy detection of several DL methods for the integration of multi-channel EEG data

The following sections are organized as follows: Section 2 describes our proposal for any validation protocol. Section 3 describes the experimental design and Section 4 reports the results on a use case. Finally, Section 5 summarizes the main findings.

2. Statistical evaluation of performance reproducibility

The target population of a statistical inference from samples is precisely the population from which they come their experimental sampling. In EEG, sampling choices (window, state, subject) determine both cross-validation splits and the generalization level of the reported metrics. Independence between training and test sets is therefore essential for ensuring the highest reproducibility of results.

In the context of EEG analysis, ML/DL methods take as input temporal windows cropped from recordings of subjects that include different neurological activity intervals (e.g. ictal/interictal states for epilepsy, or baseline and different levels of workload for memory workload). These windows can be stratified (grouped) at three different levels:

- individual samples as temporal window of a few seconds;
- abnormal states as state intervals corresponding to normal vs abnormal conditions, such as a workload level, ictal and interictal states, anxiety, apnea, narcolepsy, etc;
- subjects as whole individuals.

Each level defines an experimental sampling for the computation of metrics and data splitting in k-fold schemes. Each split allows us to validate ML/DL methods at different levels of personalization, but also defines a hierarchical population structure that allows different levels of inference on methods' performance.

Fig. 1 illustrates the three experimental samplings for signals splitting within the hierarchical population structure defined by these experimental samplings: the window level, H_0, with N_{H_0} samples, the state level, H_1, with N_{H_1} samples and the subject level, H_2, with N_{H_2} samples. The fact that the experimental sampling at level H_i can be clustered into $N_{H_{i+1}}$ independent samples of the higher level H_{i+1} defines a dependence across them that should be taken into account in statistical models.

A validation at window level has the least reproducibility power, unless the variability across the groups of a higher order level is taken into account. A validation at state level can assess reproducibility of results on unseen (for example, ictal and interictal, or workload) intervals for models personalized for the set of patients in H_2, provided that inter-state variability is modeled and incorporated into the analysis of metrics. Finally, the higher level can assess the performance of population models on unseen subjects provided that inter-subject variability is incorporated into the validation.

The performance of methods with inference at different levels of the hierarchical population structure can be analyzed using GMMs [3]. GMMs have a flexible formulation that allows an analysis of factors having an impact on performance with inference of results at different levels of generalization. On the one hand, these regression models allow to estimate the variance of a metric across methods at the level of the experimental sampling by including the method as a fixed effect. On the other one, they can include a random effect to estimate across-sample variability and, thus, allow the evaluation of the generalization of results for new unseen cases of the same population level.

GMMs are defined as soon as the dependent variable to be analyzed and the independent variables/factors candidates to have a correlation or influence on the dependent variable values are identified. In our

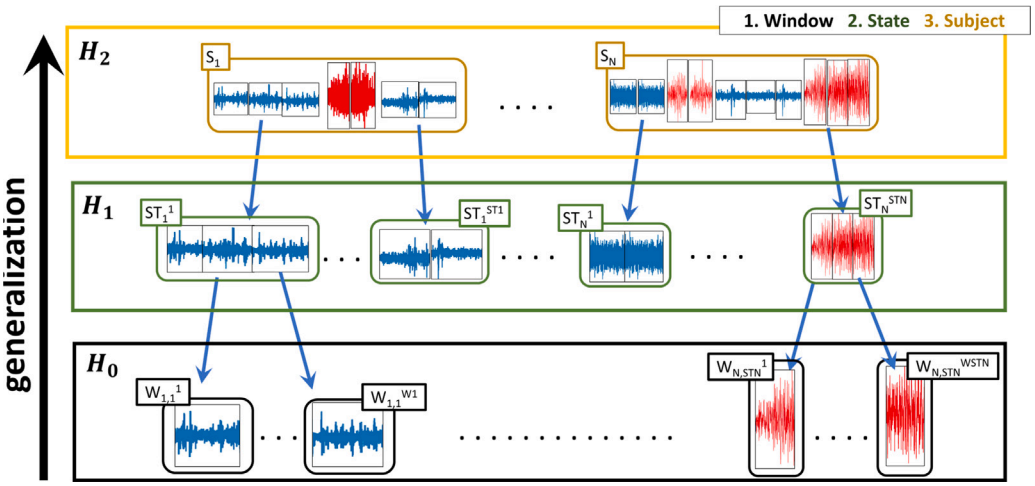

Fig. 1. Hierarchical population structure defined by EEG levels: yellow for subject, green for state and black for window.

case, the goal is to compare the performance of a set of methods and assess their across-subject generalization/variability. Thus, the dependent variable is any performance metric analyzed as function of the method used. Since the methods are fixed, they are formulated as a fixed effect in the GMM equation. Given that they are evaluated in a set of individuals sampled from a population, in order to model and assess differences in performance across them and allow GMM for inference to the whole population (i.e. subjects not belonging to the evaluation set), the subject is added as a random effect to GMM formulation. Next, we formulate GMMs for the different levels of generalization.

Let the subscript h_i denote the value of a quality metric, Q, computed for the h_i-th sample of the H_i population level and m the mth method. Then, the formulation for a model analyzing Q across N_{Mth} methods at the H_i population level is given by:

$$Q_{h_i m} = \mu + \beta_m + \epsilon_{h_i m} \tag{1}$$

for μ the intercept of the model, β_m, $m = 1, \ldots, N_{Mth}$ being the method fixed factor, $Q_{h_i m}$ the metric computed for the h_i-th sample using the mth method and $\epsilon_{h_i m} \sim N(0, \sigma_\epsilon^2)$ the error of the GMM regression model for $Q_{h_i m}$. The intercept μ represents the average metric for the reference method and β_m the deviations (increase or decrease) in the mth method metric. The model (Eq. (1)) compares Q across the methods β_m assuming that the observations $Q_{h_i m}$ are independent and cannot be grouped into higher order samplings. Under this assumption inference is somehow a "personalized" one since it is limited to the level of the sampling included in the study. This implies that, in case h_i samples can be clustered in N_{H_k} samples of a higher order level, H_k, the inference would be limited to the N_{H_k} samples used for validation. Since the model (Eq. (1)) is not evaluating performance on H_k population, it is not guaranteed reproducibility in new unseen samples of the level H_k.

This population inference can be modeled by including a random effect for the clustered samples (e.g. subjects):

$$Q_{h_i m h_k} = \mu + \beta_m + u_{h_k} + \epsilon_{h_i m h_k} \tag{2}$$

with $Q_{h_i m h_k}$ the metric of the h_i-th sample from the h_k-th cluster computed using the mth method, $\epsilon_{h_i m h_k} \sim N(0, \sigma_\epsilon^2)$ the error of the GMM regression model and the random effect $u_{h_k} \sim N(0, \sigma_u^2)$, $h_k = 1, \ldots, N_{H_k}$, modeling a sampling of size N_{H_k} of the population at H_k level. The parameters σ_ϵ^2 and σ_u^2 model, respectively, within-cluster and across-cluster variability. Therefore an inference based simultaneously on both variabilities has a "population" inference at the higher level H_k, meaning that results will be reproducible for new samplings at level k.

Random effects also compensate correlations that arise in case of repeated measurements which are usual in case of limited datasets (like in medical applications) forcing that methods are tested on the same set of samples. This inter-sample variability can be decoupled from the main fix effect target of the inference by incorporating a random effect to (Eq. (1)) as follows:

$$Q_{h_i m} = \mu + \beta_m + u_{h_i} + \epsilon_{h_i m} \tag{3}$$

with the random effect $u_{h_i} \sim N(0, \sigma_u^2)$, $h_i = 1, \ldots, N_{H_i}$, representing the across-sample (poblational) variability.

To mitigate the impact of non-normality in residuals (ϵ. in Eqs. (1)–(3)), several transformations (including logarithmic, Gamma and Poison) can be applied to the dependent variable Q. In such cases, models are adjusted in the transformed domain and estimates on Q are back-transformed to the original scale for their interpretation.

The analysis of the distribution of random effects also allows us detecting subpopulations with a distinct behavior. Under the assumption that all individuals come from the same population, the residuals u_k should follow a normal distribution $N(0, \sigma_u^2)$. Otherwise, random effects follow a mixture of gaussians, each modeling one of the underlying subpopulations. Subpopulations can be identified using a clustering (like K-means) of the estimated random effects with the number of clusters given by the minimum K such that the random effects of models adjusted for each population cluster are all normal. Since each subpopulation has a different distribution of the metric Q, their identification can be useful to design the level of specialization of models. Finally, single cases with extreme values are statistical outliers prone to bias GMM estimations and, thus, should be excluded. Statistical outliers can be detected using interquartile ranges (IQR method) or z-values of the random effects distribution.

In the following, we formulate the different GMMs associated to each of the experimental levels of EEG data and analyze the degree of reproducibility and generalization that each level is able to infer on the metric Q.

At **window level**, windows cropped from all subjects in the dataset are considered independent and, thus, data is split regardless of the subject or interval windows were sampled from. Given that usual quality metrics can only be computed on a group of windows, the implicit sampling used for the formulation of GMMs is the K-fold split. In this way the model (Eq. (1)) is formulated with Q_{h_i} computed for each fold, $h_i = 1, \ldots, K$, being K the number of folds. This GMM validates the reproducibility of methods on new K-fold splittings of the same set of windows. In order to leverage inference to unseen windows, K-folds should be clustered into higher order samplings of the hierarchical population structure. However, a K-fold split is not designed to leverage inference to a population level higher in the hierarchy given that each fold includes samples from several clusters of higher levels and, thus, K-fold metrics cannot be estimated using models with random effects.

At **state level**, the training set excludes windows belonging to a random selection of state intervals in a design that allows to evaluate the performance of models personalized for a given set of subjects at two different levels. In a state-wise split, quality metrics can be computed for each interval of the test set and their reproducibility on new intervals from the same set of subjects can be inferred using the personalized model (Eq. (1)). Besides, since metrics computed for each interval can be grouped according to the subject, the poblational GMM given by (2) can also be formulated. This GMM leverages inference to methods personalized for any set of subjects (different than the one used in the experiment) in the population.

At **subject level**, test windows belong to subjects excluded from the training and, thus, allows the assessment of generalist population methods using the random effect model given by (3) with metrics computed for each test subject. In this case, random effects would account for inter-subject variability and reproducibility would be assessed for new unseen subjects evaluated using models trained on the current population.

3. Application to epileptic seizures detection using DL methods

To show how the validation protocol described in Section 2 can be applied we compare the generalization power of several neural network architectures applied to a use case that involves, windows, states and subjects, for the detection of epileptic seizures in EEG Multi-Channel Signals.

3.1. Architectures for EEG analysis

Approaches for the analysis of EEG signals differ on the type of EEG features they use and also where and how they combine the information from EEG signals. The most popular architectures are Long Short-Term Memory (LSTM) models and Convolutional Neural Networks (CNN). All of them, take as input a 2D matrix given by EEG stacked channels. This matrix will be noted by $X^{C\times T}$, for C the number of EEG channels and T the temporal dimension.

LSTM networks are a type of recurrent neural network (RNN) specifically designed to model sequential data by capturing long-range temporal dependencies. In the context of EEG analysis, where signals are inherently time series data, LSTMs are particularly well-suited, since EEG signals show complex temporal dependencies, both within short intervals (e.g., windows of some seconds) and across longer periods (e.g., states or sessions). Several models [9,10,30–33] use LSTM to project channels taking into account their learned temporal features. These models process the spatial 2D matrix given by the stacking of windows for all channels. In this way, channels are projected taking into account the projection at previous times, so that the classification of the window can take into account the evolution of the EEG channels inside the window time interval.

CNN-based models for EEG analysis share three essential units: channel fusion, convolutional feature extraction and classification layers. The key difference lies in the placement and design of the fusion unit. When placed before the convolutional unit, the channels are fused at the raw input level (**early fusion**), while post-convolution fusion combines feature representations at the mid-level (**mid fusion**). Early fusion models [20,26,34–36] extract convolutional temporal features from the projected EEG channels, while mid fusion models [11,37,38] project channel-specific temporal representations channels obtained independently by applying convolutional operations.

The design of the fusion unit plays a critical role in combining multi-channel EEG data for feature extraction. Three approaches are considered:

1. **Concatenation (CAT).** The dimensionality of the EEG channels is flattened to create a representation vector by concatenation of the representation space of all channels. In the case of early fusion, the representation space is given by the value of the signal recorded by the EEG electrodes, so that the resulting projection is given by: $X^{C\times T} \mapsto X^{C*T}$. In the case of mid fusion, channels are represented by the values of the N_K neurons of the last convolutional block and, thus, the projection is given by: $X^{N_K\times C\times L'} \mapsto X^{N_K\times (C*L')}$
2. **Trainable Weighted Average (W-AVG).** A trainable linear layer is used to compute an average pooling with weights learned for each channel. For an early fusion approach, the projected signal is a 1D vector, X^T, with the tth component, $x^{early}_{W-AVG}(t)$, given by:

$$x^{early}_{W-AVG}(t) = \sum_{c=0}^{C} w_c x^{in}(c,t) \tag{4}$$

being $x^{in}(c,t)$ the (c,t) component of the input matrix $X^{C\times T}$. For a fusion after the convolutional unit, the projection of channels is a 2D tensor, $X^{N_K\times L'}$, with the (n_k,l') component, $x^{mid}_{W-AVG}(n_k,l')$, computed from the 3D feature tensor components $x^{mid}(n_k,c,l')$ as:

$$x^{mid}_{W-AVG}(n_k,l') = \sum_{c=0}^{C} w_c x^{mid}(n_k,c,l') \tag{5}$$

For both cases, the C weights, w_c, are the weights of the linear layer learned during the training process. These weights are shared for all the dimensions representing a channel.
3. **Multiple Trainable Weighted Average (MW-AVG).** Since in case of a mid fusion, each channel is represented by an N_K dimensional space, another option is to learn different weights for each dimension and compute $x^{N_K\times L'}$ as:

$$x^{mid}_{MW-AVG}(n_k,l') = \sum_{Nn=0}^{N_K}\sum_{c=0}^{C} w_c^{n_k} x^{mid}(n_k,c,l') \tag{6}$$

for $w_c^{n_k}$ the components of a 2D matrix of weights that is learned during training. This multidimensional trainable pooling is implemented using a 2D convolutional layer of kernel size $(C,1)$, N_K input channels and N_K output neurons.

Six methods have been carefully selected to enable a thorough comparison of temporal and spatial modeling approaches (CNN vs. LSTM), explore the diversity in fusion strategies (early vs. mid fusion), and assess the impact of trainable components (e.g., fixed concatenation vs. trainable weights) on model reproducibility and generalization to unseen data. These methods span a spectrum of complexity and capability in channel fusion, providing a comprehensive framework to evaluate how different fusion strategies influence the robustness and generalizability of models in EEG signal analysis.

Taking into account the methods found in the literature that can be reproduced the final selection includes following methods:

M1: Early CAT fusion [39] as baseline for early-stage integration of spatial information before feature extraction begins. This architecture has one linear layer to concatenate all the signals, three sequential blocks of convolution and max-pooling layers of 16, 32 and 64 units, respectively, and two fully connected dense layers of 128 and 2 units, respectively. ReLu is the activation function.

M2: Mid CAT fusion [39], for the evaluation of how spatial feature integration after initial processing impacts model performance. This architecture has three sequential blocks of convolution and max-pooling layers of 16, 32 and 64 units, respectively, one linear layer to concatenate all the signals and two fully connected dense layers of 128 and 2 units, respectively. ReLU is the activation function.

M3: Early W-AVG fusion [20] to assess the benefits of incorporating a trainable weights of channel importance. This architecture consists of four blocks of convolution and max-pooling layers, and a dense classification layer at the end. The first convolution operation is broken down into two convolution layers of 20 units each, one over time followed by another over electrodes. The following convolutional layers have 40 and 80 units respectively. ELU is used as the activation function.

M4: Mid W-AVG fusion [11] to evaluate the benefits of dynamically learning channel importance at channel-specific feature representation level. This architecture consists of three core blocks to handle input samples, followed by two dense layers of 50 and 2 units respectively, and a softmax classifier. Convolution layers have 16, 32 and 64 units, respectively. Each pooling layer goes only in temporal dimension without overlap, where the first two are set as max pooling with kernel size 2, and the third is set as average pooling with kernel size 7.

M5: Mid MW-AVG fusion [39] to assess the benefits of advanced, trainable pooling mechanisms. This architecture has three sequential blocks of convolution and max-pooling layers of 16, 32 and 64 units, respectively, another convolutional layer with kernel of size the number of channels, and two fully connected dense layers of 128 and 2 units, respectively. In this model there is no activation function to ensure the linear combination of channels.

M6: LSTM [10] to evaluate benchmark convolutional architectures against sequential model explicitly designed to capture EEG temporal dynamics. This architecture consists of 256 memory units with ReLU as the activation function, followed by two fully connected dense layers. The first dense layer has 128 units and a non-linear activation function (ReLU), while the final layer has the sigmoid activation function.

It is important to notice that all the proposed fusion strategies operate on multi-channel inputs and learn a joint representation of the multi-channel signals. We agree that these strategies do not explicitly model inter-channel dependencies, like coherence, correlation, or connectivity measures used for the diagnosis of disorders like ADHD or emotional dysfunctions. However, they still integrate information from multiple EEG channels into a shared feature space, in contrast to single-channel analyses, whose features are computed independently from each electrode without combining information across channels. Therefore, they should be considered a multi-channel analysis approach based on an implicit Integration of EEG multi-signal information.

3.2. Dataset

For all the experiments, we have used the public CHB-MIT Scalp EEG Database [40] that contains 24 cases (labeled chb_i, for i=1, …,24) recorded using a 21 electrode EEG device sampled at 256 Hz [41]. Each case contains several .edf files from a single subject. There are some files that do not contain any seizure so that are labeled as normal records while the remaining files are labeled as seizure files. Recordings have been annotated with the starting and ending second of seizures. The number and duration of seizures varies significantly across patients, from 3 to 27 seizures lasting between 6 and 752 s being a total of 198 seizures with an average of 586,414 s. This represents a total of 11,015 s of seizures for a total of 951.93 h of recording, being a total of 185.51 h of interictal state recording in the seizures files.

EEG signals went through the following pre-processing steps before being the input to models. Since we address the discrimination between ictal and interictal states, we exclude the immediate pre-ictal and post-ictal periods, which are known to show transient dynamics and recovery processes that differ from the more stable ictal/interictal activity that we aim to model [42,43]. Given that these states are not annotated in the CHB-MIT database, and based on prior EEG and SEEG works, we discarded a symmetric 30 s before and after each annotated seizure to exclude transient peri-ictal activity while discarding only a small fraction of the available data. The remaining intervals were down-sampled to 128 Hz [44] and cut into 1 s windows [9,17,45]. No additional denoising was applied since a basic Butterworth filter was used during acquisition [10]. To mitigate class imbalance, ictal windows were extracted with 80% of overlap, while interictal ones were non-overlapping and drawn from files labeled by "seizure". This yielded a total of 645,999 interictal and 56,192 ictal windows. Alternative preprocessing (e.g., different band-limits, artifact rejection, window lengths) could influence absolute metrics; however, our hierarchical validation and GMM analysis remain applicable and would quantify their impact on reproducibility

Table 1 reports patient-specific information, summing up 951.93 h of recording and 181 seizures with a duration of 11,015 s. As a general information, the dataset was recorded from 23 patients with ages 3–22 years, being recording chb21 obtained from the same patient chb01, but 1.5 years later; there is no information about sex and age for chb24 because it was added to the dataset later [40].

3.3. Data splits for hierarchical cross-validation

To assess the level of generalization of models, we used a cross-validation scheme with the following variations for each level of the hierarchical data stratification used.

For the analysis at **window** level, we performed a 15-fold cross-validation. In each fold, approximately 6.7% of the data was held out for testing. Training was class-balanced by random undersampling of interictal windows. Test sets preserved the actual ratio between ictal (20%) and non-ictal (80%) windows.

For the analysis at **state** level, we have used a cross-validation over seizure intervals, that is, a leave-one-state-out procedure. We trained 198 independent models, each time holding out one entire seizure interval (all its windows) for testing. The training set comprised windows from all remaining seizure intervals plus interictal windows; training was class-balanced by randomly undersampling interictal windows. The test set consisted of all windows from the held-out seizure interval, with interictal windows drawn from the central portion of recordings labeled with "normal" to maintain a 20%/80% ictal/non-ictal ratio.

For the analysis at **subject** level, we have used a cross-validation over subjects (each subject serves as the test unit once), that is a leave-one-patient-out procedure. We trained 24 different models, each time holding out one subject for testing and training on the remaining 23 subjects. Training sets were class-balanced by undersampling interictal windows; the test set preserved the subject's original class distribution.

3.4. Metrics

We report recall, precision, and F1-score for each class (with seizure as the positive class), plus overall accuracy. Reporting per-class metrics addresses class imbalance; specifically, precision and F1 for the positive class quantify how the far more numerous interictal windows affect performance relative to the scarcer ictal windows.

We report descriptive statistics for all metrics as mean ± standard deviation ($\mu \pm \sigma$). For each metric, best performers are indicated in bold. Variability is computed as $\mu \pm \sigma$ across the folds at the window level and across subjects at the subject level.

For each experimental level (window folds, state intervals and subjects), class-specific recall were analyzed independently using the GMMs described in Section 2. For method comparisons, the GMMs compare five of the methods to the Early CAT as the reference. Model normality assumptions were assessed via residual analysis and normality tests to determine the best scale transformation for each metric. By considering separate GMMs for each class, conclusions about reproducibility of the models are not affected by the class imbalance inherent to the CHB-MIT dataset.

Table 1
Information of patients of the CHB-MIT Scalp EEG Dataset.

Patient	Gender-Age	Seizures	Seizures: Total duration (sec)	Only seizures records: Total duration (h)	All records: Total duration (h)
01	F-11	7	442	6.65	40.55
02	M-11	3	172	2.27	35.27
03	F-14	7	402	7	38
04	M-22	4	378	10.66	156.07
05	F-7	5	558	5	39
06	F-1.5	10	153	25.89	66.74
07	F-14.5	3	325	9.04	67.05
08	M-3.5	5	919	5	20.01
09	F-10	4	276	9.58	67.87
10	M-3	7	447	14.02	50.02
11	F-12	3	806	2.79	34.79
12	F-2	27	989	9.68	20.69
13	F-3	10	440	7	11
14	F-9	8	169	7	26
15	M-16	20	1992	14.01	39.01
16	F-7	8	69	5	17
17	F-12	3	293	3.01	20.01
18	F-18	6	317	5.63	34.63
19	F-19	3	236	2.93	28.93
20	F-6	8	294	5.57	27.6
21	F-13	4	199	3.83	32.83
22	F-9	3	204	3	31
23	F-6	7	424	8.96	26.56
24	–	16	511	12	21.3
Total		181	11 015	185.51	951.93

For each GMM, we report model parameters, p-values for fixed effects (with significant values in bold), and 95% confidence intervals (CI) for their mean values. Notice that these intervals correspond, indeed, to the 95% CIs of each method's average recall for seizure and non-seizure windows as estimated by the GMMs. In case of change of scale or transformation of data required to satisfy model normality assumptions, CIs were back-transformed to the original scale. For models with random effects, we also provide CI for the random-effect estimates.

Random-effect distributions were analyzed to detect subpopulations introducing a bias in normality. Visual assessment was also performed using dot plots: each subject (y-axis) is shown with a horizontal line representing the CI of its conditional mode (deviation from the population mean). Effects are plotted after subtracting the intercept, so intervals are centered at 0 (vertical reference line) to facilitate checking the normality assumption. Statistical outliers were identified using Grubbs' test [46], Shapiro–Wilkinson tests confirmed significant deviations from normality and K-means was used to identify subpopulations. Unless stated otherwise, a p-value < 0.05 was considered statistically significant.

All the models were implemented in Python 3.9 environment and PyTorch 2.10 framework. Experiments were carried out on a desktop computer with Windows 10 and an NVIDIA GeForce RTX 2070 Super graphic card. Statistical analyses were conducted using R version 4.3.2.

4. Results

Table 2 reports window-level descriptive statistics. Standard deviations are uniformly small, reflecting lack of independence between training and test sets. Average metrics are very high across methods, with M1, M2, M5 and M6 being the top performers with similar ranges (average recall > 0.96). In contrast, the trainable weighted early fusion (M3) is the worst performer across metrics, especially for those computed for seizure detection.

Table 3 summarizes the GMMs (Eq. (1)) for Recall+ and Recall- at window level. Both models were fit on the original metric scale. GMM significant factors (methods) are aligned with descriptive statistics in Table 2. For Recall+, there are not significant differences across M1, M2, M4 and M5 with M3 performing significantly worst and M6 significantly better. For Recall-, feature concatenation (M1 and M2) and LSTM (M6) are significantly better than the remaining methods, being (as expected) the early fusion of M3 the one with lower CI ranges.

Table 4 reports-state level descriptive statistics. Concatenation (M1,M2) and multi-weighted (M5) fusions achieve the best ranges. However, compared to window-level results, mean performance decreases (especially for the positive class), and standard deviations increase substantially. This greater variability is consistent with the heterogeneity of epileptic patterns reported in prior works [47,48], which current methods do not fully capture. Accordingly, these effects should be further analyzed using the GMM with subject as random effect given by (Eq. (2)) to identify subpopulations and discard outliers.

Two GMMs given by (2) were adjusted for Recall+ and Recall- on a logarithmic scale to account for non-normality of models' residuals. Fig. 2 plots 95% CI for the random effects estimated for Recall- (left plot) and Recall+ (right plot). We observe a large across-subject variability in estimated effects ranges associated to data heteroscedasticity. For both metrics, subjects chb_{17}, chb_{02}, chb_{11}, chb_{22} have the largest CI ranges (around 0.1 of estimated variance) and subjects chb_{12}, chb_{15} the smallest ones (0.045 of estimated variance approximately). Both effects also show substantial deviation from centrality. In particular, random effects for the GMM estimating Recall+ are symmetric with chb_{14} having extreme effects. A Grubbs test detects this case as an outlier (p-val = 0.03796), while the remaining cases do not significantly deviate from normality according to a Shapiro–Wilkinson test (p-val=0.2243). Random effects for the GMM estimating Recall- are clearly biased from centrality, clear indicator of non-normality, and have case chb_{08} with extreme values. This case is detected by a Grubbs test as an outlier (p-val = 0.02407) and it will be excluded from the further analysis. A Shapiro–Wilkinson test on the remaining population rejects the normality hypothesis with p-val<0.01. To further investigate and identify subpopulations, a K-means clustering with K=2 was applied to Recall- random effects. K-means clustered random effects into two subpopulations: Subpopulation1 with {chb_{02}, chb_{03}, chb_{04}, chb_{07}, chb_{09}, chb_{10}, chb_{11}, chb_{12}, chb_{14}-chb_{23},} and Subpopulation2 with cases {chb_{06}, chb_{13}, chb_{05}, chb_{08} and chb_{01}}. The random effects of two GMMs in logarithmic scale adjusted for each subpopulation did not significantly deviate from normality according to a Shapiro–Wilkinson (p-val >0.3).

Table 5 reports a summary of the GMMs in logarithmic scale estimated for Recall+ (excluding the outlier chb_{14}) and for each Recall-

Table 2
Comparison of the methods. Descriptive Statistics (top metrics in bold) at Window Level.

Models	Recall+	Recall−	Precision+	Precision−	F1+	F1−	Accuracy
M1	0.96 ± 0.00	1.00 ± 0.00	0.95 ± 0.00	1.00 ± 0.00	0.96 ± 0.00	1.00 ± 0.00	0.99 ± 0.00
M2	0.96 ± 0.00	**1.00 ± 0.00**	**0.95 ± 0.00**	**1.00 ± 0.00**	**0.96 ± 0.00**	**1.00 ± 0.00**	**0.99 ± 0.00**
M3	0.88 ± 0.01	0.94 ± 0.01	0.55 ± 0.04	0.99 ± 0.00	0.68 ± 0.02	0.96 ± 0.00	0.93 ± 0.01
M4	0.96 ± 0.00	0.98 ± 0.00	0.81 ± 0.03	1.00 ± 0.00	0.88 ± 0.01	0.99 ± 0.00	0.98 ± 0.00
M5	0.96 ± 0.00	0.99 ± 0.00	0.94 ± 0.01	0.97 ± 0.00	0.95 ± 0.00	1.00 ± 0.00	0.99 ± 0.00
M6	**0.97 ± 0.00**	**1.00 ± 0.00**	0.94 ± 0.01	1.00 ± 0.00	0.96 ± 0.00	1.00 ± 0.00	0.99 ± 0.00

Table 3
Comparison of the methods. Regression Models (significant effects shaded) at Window Level.

	Recall +			Recall −		
Models	coefficient	p_val	95% CI	coefficient	p_val	95% CI
M1	9.60e-01	–	(0.96,0.96)	1.00e+00	–	(1.00,1.00)
M2	−5.75e-07	1.00	(0.96,0.96)	3.04e-07	1.00	(1.00,1.00)
M3	−7.57e-02	**0.00**	(0.88,0.89)	−6.21e-02	**0.00**	(0.94,0.94)
M4	−2.80e-07	1.00	(0.96,0.96)	−2.00e-02	**0.00**	(0.98,0.98)
M5	−1.45e-07	1.00	(0.96,0.96)	−1.00e-02	**0.00**	(0.99,0.99)
M6	1.00e-02	**0.00**	(0.97,0.97)	1.87e-07	1.00	(1.00,1.00)

Table 4
Comparison of the methods. Descriptive Statistics (top metrics in bold) at State Level.

Models	Recall+	Recall−	Precision+	Precision−	F1+	F1−	Accuracy
M1	**0.86 ± 0.22**	**0.98 ± 0.04**	0.91 ± 0.19	**0.97 ± 0.05**	0.87 ± 0.19	**0.97 ± 0.03**	**0.96 ± 0.05**
M2	**0.87 ± 0.21**	**0.98 ± 0.03**	**0.92 ± 0.17**	**0.97 ± 0.04**	**0.88 ± 0.17**	**0.97 ± 0.03**	**0.96 ± 0.04**
M3	0.75 ± 0.24	0.91 ± 0.12	0.73 ± 0.26	0.94 ± 0.05	0.71 ± 0.23	0.92 ± 0.08	0.88 ± 0.11
M4	0.82 ± 0.20	0.93 ± 0.11	0.79 ± 0.20	0.96 ± 0.04	0.78 ± 0.18	0.94 ± 0.08	0.90 ± 0.10
M5	0.85 ± 0.22	**0.98 ± 0.04**	0.90 ± 0.19	**0.97 ± 0.05**	0.86 ± 0.19	**0.97 ± 0.03**	0.95 ± 0.05
M6	0.75 ± 0.21	0.95 ± 0.06	0.82 ± 0.18	0.94 ± 0.05	0.77 ± 0.17	0.94 ± 0.04	0.91 ± 0.06

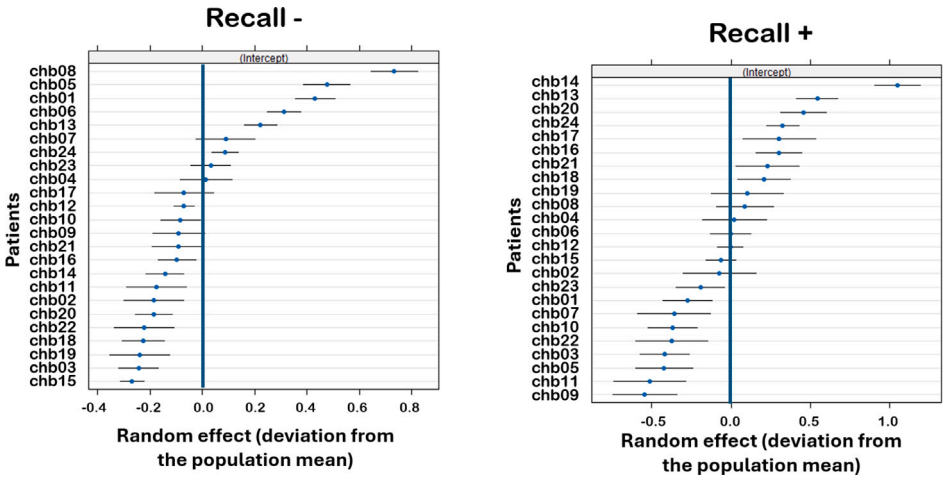

Fig. 2. Comparison of the methods at State Level. Visual analysis of random effects normality estimated for Recall- (left) and Recall+ (right) GMMs. For each plot, *y*-axis represents each subject and *x*-axis are confidence intervals for the estimated deviation (conditional modes) of each subject from the estimated metric average.

subpopulations at state level. For all metrics, concatenation (M1, M2) and multi-weighted average (M5) perform significantly better than the other fusion methods. For all models, Recall- CI ranges for Subpopulation2 are lower than ranges for Subpopulation1 with the largest decrease for W-AVG fusions models (M3,M4). This identifies Subpopulation2 cases as having potentially unusual heterogeneous interictal patterns with a variability comparable to the one observed in ictal episodes (Recall+ ranges are similar for M1,M2 and M5). For Subpopulation1, all CI ranges are comparable to the ones obtained at window level, with similar high average and small variance, which indicates highly homogeneous non-ictal patterns.

Table 6 reports subject-level descriptive statistics. Concatenation at input level (M1) and multi-weighted (M5) mid-fusion achieve the highest metrics. However, all population-generalist models show a substantial performance drop, with the lowest means and highest standard deviations, especially for seizure detection. The high precision in detection of normal states in contrast to low seizure detection reflects the strong class imbalance. As before, the large variability motivates a random-effects GMM analysis (Eq. (3)) to probe subject-level heterogeneity.

In order to identify subpopulations, 2 GMMs were adjusted for Recall+ and Recall-. Both models were adjusted using the Gamma distribution with log link function to compensate data heteroscedasticity and non-normality. Fig. 3 shows the random effects of the 2 GMMs estimated for Recall- (left plot) and Recall+ (right plot).

Table 5
Comparison of the methods. Regression Models (significant effects shaded) at State Level.

	Recall +			Recall −					
				Subpopulation 1			Subpopulation 2		
Models	coefficient	p_val	95% CI	coefficient	p_val	95% CI	coefficient	p_val	95% CI
M1	−1.75e+00	–	(0.89,0.96)	−2.23e+00	–	(0.99,1.00)	−1.82e+00	–	(0.90,0.98)
M2	−2.71e-02	0.62	(0.90,0.96)	6.69e-03	0.78	(0.99,1.00)	−4.95e-02	0.55	(0.91,0.98)
M3	**4.54e-01**	**<0.01**	**(0.78,0.88)**	**2.97e-01**	**<0.01**	**(0.95,0.96)**	**6.55e-01**	**<0.01**	**(0.71,0.87)**
M4	**2.41e-01**	**<0.01**	**(0.84,0.92)**	**2.07e-01**	**<0.01**	**(0.96,0.98)**	**5.88e-01**	**<0.01**	**(0.74,0.88)**
M5	2.79e-02	0.61	(0.89,0.95)	1.13e-03	0.96	(0.99,1.00)	6.23e-02	0.45	(0.89,0.97)
M6	**4.61e-01**	**<0.01**	**(0.77,0.88)**	**1.67e-01**	**<0.01**	**(0.97,0.98)**	**3.23e-01**	**<0.01**	**(0.82,0.93)**

Table 6
Comparison of the methods. Descriptive Statistics (top metrics in bold) at Subject Level.

Models	Recall+	Recall−	Precision+	Precision−	F1+	F1−	Accuracy
M1	**0.77 ± 0.22**	**0.90 ± 0.17**	0.43 ± 0.36	**0.99 ± 0.01**	**0.41 ± 0.31**	**0.93 ± 0.13**	**0.90 ± 0.17**
M2	0.68 ± 0.31	**0.91 ± 0.17**	**0.45 ± 0.36**	**0.99 ± 0.01**	**0.42 ± 0.31**	**0.94 ± 0.13**	**0.90 ± 0.17**
M3	0.69 ± 0.26	0.87 ± 0.15	0.19 ± 0.15	**0.99 ± 0.01**	0.27 ± 0.19	0.92 ± 0.11	0.87 ± 0.15
M4	0.69 ± 0.26	0.83 ± 0.17	0.18 ± 0.18	**0.99 ± 0.01**	0.24 ± 0.20	0.89 ± 0.15	0.82 ± 0.19
M5	**0.70 ± 0.32**	0.89 ± 0.21	0.41 ± 0.34	**0.99 ± 0.01**	**0.41 ± 0.32**	0.92 ± 0.16	0.89 ± 0.20
M6	0.60 ± 0.31	**0.91 ± 0.13**	0.21 ± 0.16	**0.99 ± 0.01**	0.27 ± 0.19	**0.95 ± 0.07**	**0.91 ± 0.11**

In this case, Recall- effects present homoscedasticity with equal CI ranges for random effects with 0.25 variance approximately. Recall+ effects have a heteroscedasticity pattern with increasing ranges from 0.2 for chb_{09} to over 0.2 for chb_{14}. Regarding centrality associated to normality, random effects for the GMM estimating Recall+ are symmetric and also have chb_{14} having extreme effects. Unlike effects at state level, a Grubbs test does not detect this case as an outlier (p-value =0.9761) and the effects for the whole population significantly deviate from normality according to a Shapiro–Wilkinson test (p-val=0.01). To identify subpopulations, a K-means clustering with K=2 was applied to Recall+ random effects. K-means clustered random effects into two subpopulations: Subpopulation1 with {chb_{01}-chb_{07}, chb_{09}-chb_{11}, chb_{17}-chb_{19} and chb_{22}-chb_{24} } and Subpopulation2 with cases {chb_{08}, chb_{12}-chb_{16} and chb_{20}-chb_{21}}. The random effects of two GMMs using the Gamma distribution adjusted for each subpopulation did not significantly deviate from normality according to a Shapiro–Wilkinson (p-val>0.05).

Recall- random effects distribution is also asymmetric and has chb_{09} subject with extreme values. A Grubbs test (p-val = 0.02331) confirms chb_{09} as an outlier case and, thus, it will be excluded from the further analysis. A Shapiro–Wilkinson test on the remaining population rejects the normality hypothesis with p-val<0.001. To identify subpopulations, a K-means clustering with K=2 was applied to Recall- random effects. The k-means clustered random effects into two subpopulations: Subpopulation1 with {chb_{01}, chb_{02}, chb_{03}, chb_{04}, chb_{07}, chb_{08} and chb_{15}-chb_{23},} and Subpopulation2 with cases {chb_{05}, chb_{06} and chb_{10}-chb_{14}}. The random effects of two GMMs in Gamma scale adjusted for each subpopulation did not significantly deviate from normality according to a Shapiro–Wilkinson (p-val=0.9).

Table 7 reports the models estimated for each Recall- (on top) and Recall+ (on bottom) subpopulations. For Recall-, CI ranges for Subpopulation2 are lower than ranges for Subpopulation1 for all models with no significant differences among performance on Subpopulation2. This identifies Subpopulation2 cases as having potentially unusual heterogeneous interictal patterns. It is worth noting that, although non-significant (p-val=0.06), the LSTM method performs better than the remaining ones with ranges closer to the one achieved for Subpopulation1. Subpopulation1 Recall- CI ranges are similar to the ones estimated at state level with mid fusions (M2 and M5) being significantly better, M3 and M4 significantly worse and M1 and M6 with similar mid performance. This suggests that Subpopulation1 has highly homogeneous non-ictal patterns. For Recall+, all methods also perform worse in Subpopulation2, but differences in CI ranges are larger than for non-ictal detection. This not only identifies Subpopulation2 as having larger variability in ictal patterns, but also indicates the higher heterogeneity of epileptic seizures in comparison to non-ictal intervals. This is also reflected in Subpopulation1 CI ranges, which are closer to the ones obtained for Subpopulation2 in Recall- than to the top Recall- ranges of Subpopulation1. Regarding models performance, concatenation (M1 and M2) and weighted averages (M3 and M6) perform significantly better in Subpopulation1, while M1 is the best performer in both subpopulations.

5. Discussion and conclusions

5.1. Models' performance

Across-level comparisons confirm that window-level results are over-optimistic (Accuracy ≈ 0.99, Sensitivity ≈ 0.99 and Specificity ≈ 1.00) due to the lack of independence between train and test sets, confirming the need for higher-level validation schemes. On the contrary, state and subject-level splits yield more realistic performance estimates, with drops in metrics to Accuracy ≈ 0.91, Sensitivity ≈ 0.75 and Specificity ≈ 0.95 at state level and Accuracy ≈ 0.91, Sensitivity ≈ 0.60 and Specificity ≈ 0.91 at subject level. Notably, all models exhibited substantial variability in detecting seizure patterns, both across subjects in generalist models (Table 6) and intra-subject in personalized models (Table 4). This variability exhibits an increasing tendency across generalization levels, rising from 0.00 at window level to 0.24 at state interval level and up to 0.32 at subject level. At this last level, non-seizure detection also exhibits a substantial increase in variability, even for the most stable methods (M1,M2) from 0.00 and 0.04 at, respectively window and state interval levels to 0.17 at subject level. These findings highlight the challenges of generalizing to unseen conditions and the need for independence between train and test samples.

Regarding fusion architectures, the best performers are those approaches learning representation spaces that take into account channel-wise information. Fusions based on non-trainable averages summarize channel information into a single projection representation vector. Being projections some valuable discriminative information might be lost in the process. CAT considers the information of all channels independently to compute the representation vector and, thus, the network can learn and preserve a combination of more discriminative features. MW-AVG computes trainable averages independently for each dimension and channel and, thus, it can also learn more discriminative features. This potential loss of information in representations based on projections has a negative impact on the performance of architectures

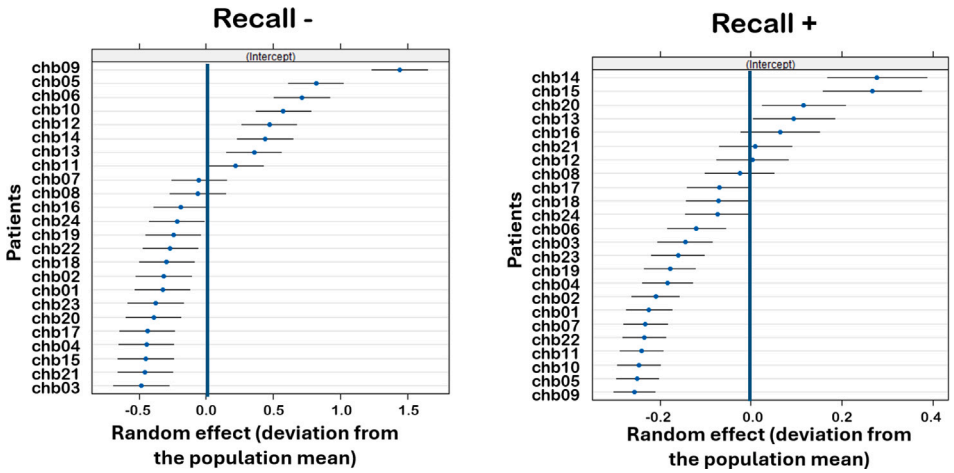

Fig. 3. Comparison of the methods at Subject Level. Visual analysis of random effects normality estimated for Recall- (left) and Recall+ (right) GMMs. For each plot, *y*-axis represents each subject and *x*-axis are confidence intervals for the estimated deviation (conditional modes) of each subject from the estimated metric average.

Table 7
Comparison of the methods. Regression Models (significant effects shaded) at Subject Level.

	Recall–					
	Subpopulation1			Subpopulation2		
Models	coefficient	p_val	95% CI	coefficient	p_val	95% CI
M1	−2.05e+00	–	(0.96,0.98)	−1.06e+00	–	(0.61,0.85)
M2	**−1.45e-01**	**<0.001**	**(0.98,1.00)**	−2.63e-01	0.25	(0.73,0.91)
M3	**1.36e-01**	**<0.001**	**(0.94,0.97)**	−6.28e-02	0.78	(0.64,0.87)
M4	**1.83e-01**	**<0.001**	**(0.93,0.96)**	2.31e-01	0.31	(0.49,0.79)
M5	**−1.40e-01**	**<0.001**	**(0.98,1.00)**	−1.83e-01	0.42	(0.70,0.90)
M6	4.79e-02	0.27	(0.95,0.98)	−4.26e-01	0.06	(0.78,0.94)
	Recall+					
	Subpopulation1			Subpopulation2		
Models	coefficient	p_val	95% CI	coefficient	p_val	95% CI
M1	3.71e-01	–	(0.75, 0.92)	6.09e-01	–	(0.30, 0.67)
M2	1.07e-02	0.39	(0.74, 0.90)	**1.77e-01**	**<0.01**	**(0.01, 0.48)**
M3	2.24e-02	0.08	(0.72, 0.88)	**1.37e-01**	**0.01**	**(0.11, 0.52)**
M4	**2.85e-02**	**0.03**	**(0.71, 0.86)**	8.44e-02	0.08	(0.19, 0.57)
M5	5.08e-03	0.68	(0.74, 0.91)	**1.08e-01**	**0.03**	**(0.16, 0.55)**
M6	**4.98e-02**	**<0.01**	**(0.68, 0.83)**	**2.72e-01**	**<0.01**	**(0.00, 0.38)**

(M3, M4) using averages with common weights for all channels. On the other hand, CAT (either at input level, like M1, or feature one, like M2) and MW-AVG (M5) with channel-wise weights are able to learn discriminant representation spaces which are significantly better for, both, personalized and generalist population models. The channel fusion using temporal information (M6) exhibits more stable Recall- metrics, with higher values for the subpopulation having more variability in interictal intervals. This suggests adding temporal information to channel-wise combinations.

We would like to highlight that performance of these low complexity models compares to SoA methods based on more complex architectures, including transformer-based [25], a CNN squeeze-and-excitation modules, EEGNET [26], Wei-CNN [27], and CNN-LSTM [28]. Specifically, the comparison using leave-one-patient-out scheme on CHB-MIT cases reported in show that results of specificity and sensitivity have moderate average values of, respectively, 0.86 and 0.79 for top performers that agree to the ranges achieved by our best fusion models (0.9+/−0.17, 0.77+/−0.22 for M1). This comparison suggests that appropriate fusion strategies with shallow architecture can achieve good results with less computational cost and data burden for training.

We would also like to note that ictal 80% window overlap used to compensate data unbalancing does not affect our conclusions on performance. Although, overlap might inflate ictal metrics at window level, this data partition already violates independence between train and test sets, introducing data leakage, and, thus, it always gives over-optimistic metrics, independently of window overlap. Another important remark is that partitions at state and subject levels prevent data leakage between overlapping windows, so that metrics are not inflated by their overlap. Our analysis does not depend on absolute quantities; rather, it focuses on their comparison across hierarchical levels. Given that such overlap is applied at all levels, it does not alter nor invalidate our findings and conclusions about generalization and over-optimistic data splits.

Finally, while we have used as benchmark examples end-to-end deep learning models processing EEG raw signals, classical feature-based approaches (e.g., entropy/complexity indices, spectral/parametric modeling) often provide physiologically interpretable markers of neural organization and variability. In this context, the proposed validation framework is representation-agnostic and can equally evaluate feature-engineered pipelines [49] that aim to explain neural dynamics.

5.2. Subpopulation analysis

The analysis of GMM random effects for seizure detection, identified a subpopulation of 8 subjects ($\{chb_{08}$, chb_{12}-chb_{16} and chb_{20}-$chb_{21}\}$), representing 29% of the cases, having ictal patterns so different from

the remaining population as to drastically drop performance of generalist models. Inclusion into models' training of ictal intervals from these cases increased performance of all fusion approaches (with more than 0.9 of average recall for models M1, M2 and M5), with the exception of the case chb_{14} which remain with a low detection rate. The analysis of GMM random effects for seizure detection, identified this case as an outlier of extreme low performance in personalized models. This indicates that chb_{14} ictal patterns might be significantly different from the remaining population, but also might have a significant inter-state variability difficult to model even using personalized approaches. The remaining subjects can be considered a compact population with uniform variability in seizure patterns, given that random effects modeling inter and intra subject variability followed a normal distribution.

Surprisingly, our GMM analysis identified two subpopulations in the detection of interictal patterns for, both, personalized (state level analysis) and population (subject level analysis) approaches. At both levels one of the subpopulations (the one labeled Subpopulation1) presented high homogeneity of non-ictal patterns with average metrics above 0.99 for top performers. Average performances for Subpopulation2 drastically drop with an increase in standard deviations indicating high inter and intra subject variability in interictal periods. As expected, the Subpopulation2 for personalized models (chb_{06}, chb_{13}, chb_{05}, chb_{08} and chb_{01}) was smaller than for population ones (chb_{05}, chb_{06} and chb_{10}-chb_{14}). Only 3 cases (chb_{05}, chb_{06} and chb_{13}) belong to Subpopulation2 at, both, personalized and population levels, indicating an interictal variability difficult to model even for personalized approaches. Also, the analysis identified chb_{09} and chb_{08} as outliers for, respectively, personalized and population models.

Finally, an analysis across the subpopulations with higher variability in patterns identifies the following cases with heterogeneous patterns especially difficult to model. Cases chb_{12}, chb_{13} and chb_{14} have ictal and interictal patterns so different from the remaining population that cannot be modeled by generalist approaches. Cases chb_{05}, chb_{06} and chb_{13} have high inter and intra subject variability in non-ictal patterns and so has chb_{14} for ictal periods. These cases are difficult to model even for personalized approaches.

Our findings of population subtypes are supported by recent literature conducting patient-specific analysis using different approaches (including non-linear dynamics [50], Wavelets [51], ensemble ML models [52], convolutional attention modules for channel fusion [53], GANs [54] and LSTM [55]). Like our GMM, the work in [51] also identified chb_{13} as being a especially bad case. In [52] cases chb_{05}, chb_{06}, chb_{08}, chb_{09}, chb_{10}, chb_{13}, chb_{14} and chb_{16} had a substantial increase in false positive rate in population models. All these cases were identified by our GMM as having complex interictal (5/6, 80%) or ictal (4/8, 50%) periods difficult to model. The analysis in [53] also detects an increase in false positives for chb_{08} and chb_{10}. In [50], cases chb_{06}, chb_{08} are also worst performers. The methods for seizure forecasting reported in [54] show systematic low performance for chb_{02}, chb_{09}, chb_{10} and chb_{14}, which GMM detects 3/4 (75%) of them. Finally, in [55], cases chb_{14} and chb_{18} achieve the lowest seizure prediction metrics. Considering cases having low metrics in at least two SoA methods as the most difficult ones, our GMM was able to detect all of them. In particular, chb_{14} which is detected by our analysis as a difficult case for, both, ictal and interictal periods coincides with 5/6 of the SoA methods.

The diversity in populations detected in this study has also a neurophysiological basis. Electrical recordings from people with epilepsy often show isolated epileptiform discharges between seizures known as Interictal Epileptiform Discharges (IEDs). IEDs are large intermittent electrophysiological events observed between seizures that exhibit multidien variation in their frequency that correlates with seizure likelihood. These abnormal neuronal oscillations reflect specific dynamic changes in network excitability, play various roles in epilepsy and their diversity suggests a sub-type classification of epileptic patients based on interictal patterns [56]. Aura Analysis

In our particular case, subpopulation groups can be associated to an heterogeneity in patients' age, being those belonging to the subpopulations with a drop in performance significantly younger for interictal groups at, both, subject and state levels. For interictal variability, the average ages of the patients subpopulations at subject-level are 5.94 (95%CI of (2.53, 9.34)) and 12.13 (95%CI of (9.20, 15.06)) for respectively, SubPopulation2 and SubPopulation1 and 5.94 (95%CI of (2.53, 9.34)) and 12.13 (95%CI of (9.20, 15.06)) for respectively, SubPopulation2 and SubPopulation1 detected at state level. A one-sided T-test for independent samples was significant in both cases with p-values equal to 0.0029 and 0.0093 for, respectively, subject and state subpopulations. For interictal variability at state level the average ages of the patients are 7.44 (95%CI of (3.27, 11.60)) and 11.33 (95%CI of (8.17, 14.49)) for respectively, SubPopulation2 and SubPopulation1. Although, according to one-side T-test for independent samples, these differences were not significant (p-val equal to 0.0544), the patients in SubPopulation2 are also younger.

These findings provide a clinical interpretation of our statistical population analysis and encourage its use in the implementation of improved personalized models

5.3. Use case limitations

First, we would like to clarify that the goal of this work is to present a hierarchical statistical validation framework of signal-based ML applications using GMMs. Epilepsy detection was selected as a clinically relevant use case, since it is one of the most extensively studied EEG applications with public datasets available for benchmarking. In this context, several DL methods for the integration of multi-channel EEG data were tested to highlight how misleading validation strategies can inflate results, and how our proposed framework provides a more rigorous assessment. Our aim was to demonstrate the validity and utility of our hierarchical validation framework in a well-established benchmark task, not to focus on methods for the detection of any particular pathology (epilepsy) or single/multi channel analysis

Our final decision to use CHB-MIT was motivated by its status of being one of the most popular datasets used to validate ML/DL models for automatic seizure detection in multi-channel problems [57]. Although relatively small in size and dated, it still has enough variability in patients and data to conduct analysis using DL methods compared to other datasets. Also, its extensive use enables direct comparison with previous studies and, as extensively discussed in the paper, allows us to highlight how validation choices critically affect reported performance and compare our findings to other subject-specific analysis. A recent review shows that Bonn and CHB-MIT EEG datasets are used the most to validate deep learning models for automatic seizure detection [57]. In particular, recent DL works [25,58,59] also perform their analysis only on this last data set, which validates our choice and should reduce the concerns about generalization of our findings

Regarding the choice of the DL models, we would like to emphasize that, while they do not explicitly estimate inter-regional coupling measures such as coherence or state synchrony, they process multi-channel EEG as a joint input and learn integrated feature representations across electrodes. For this reason, they should be considered multi-channel approaches, albeit of the "fusion" type rather than the "interaction-based" type. Being implicit fusion of channels might limit interpretability of inter-regional relationships (e.g., coherence, synchrony). Therefore, future studies should pair the framework with explicit connectivity or complexity features to examine whether interpretable channel interactions improve robustness and clinical insight. Another important remark is that the proposed hierarchical validation is broadly applicable to any EEG feature extraction paradigm because the framework operates at the level of validation design and statistical inference, independent of the specific features used.

6. Conclusions

Summarizing, this study proposes a hierarchical validation protocol that spans window, state, and subject levels, offering insights into the reproducibility of models under realistic conditions. As well, evidences that can exist subpopulations within EEG datasets that significantly impact model performance, emphasizing the need for tailored solutions in clinical applications. The main conclusion is that while AI methods achieve high performance under constrained settings (e.g., window-level analysis), their generalization to new subjects and conditions remains limited. By systematically addressing these challenges, the proposed framework provides a foundation for developing clinically viable AI systems for EEG analysis. We would like to emphasize that our methodological contribution is the validation framework itself, not new EEG representation/model. Model families are illustrative testbeds of the validation framework and not the target of our innovation.

There are some points of this work that could be improved. This study relies only on the CHB-MIT database, which potentially could limit generalizability. We plan to incorporate other public databases, like TUH EEG Seizure Corpus or EPILEPSIAE, to validate the framework across datasets with different demographics. The applicability of our framework would also benefit if we applied it to EEG data recorded with other neurological conditions (like cognitive workload or sleep apnea). Still, we consider that the experiments and discussion carried out are enough to validate the potential of our framework for the assessment of the personalization level required for systems analyzing cognitive states in EEG signals.

The proposed hierarchical validation is broadly applicable to any EEG feature extraction paradigm (including dual-channel coherence, inter-hemispheric synchrony, and multi-channel connectivity analyses) because the framework operates at the level of validation design and statistical inference, regardless of the specific features used. Future work will extend this framework to dual- and multi-channel analyses, particularly for conditions where functional connectivity or hemispheric asymmetries are key (e.g., ADHD, emotional dysfunctions). Future extensions of our framework may also combine fusion-based representations with explicit connectivity features, entropy/complexity descriptors and autoregressive/spectral parametrics as inputs to shallow classifiers [49] to further enrich the modeling of spatial dependencies across cortical regions

The proposed subpopulation analysis is a valuable tool, not only for the implementation of improved ML/DL models but also for an enhanced personalized clinical diagnosis and treatment of patients. If distinct EEG-defined subgroups are shown to correlate with differential responses to pharmacological or neuromodulatory interventions, clinicians could tailor therapeutic strategies accordingly, thereby reducing reliance on trial-and-error approaches and improving patient outcomes. EEG-based subpopulation detection could also contribute to prognostic assessments, as certain electrophysiological subtypes may be associated with differential disease trajectories, risk of relapse, or likelihood of treatment resistance. Finally, these subpopulation-specific signatures may function as biomarkers of treatment response, disease progression, or relapse risk, thereby enabling phenotypic stratification and supporting precision medicine approaches. In this context, since previous studies have shown that entropy and complexity-based EEG measures reflect neural system heterogeneity and organizational changes across populations, deep analysis of the connection between the observed inter-subject variability and underlying differences in EEG signal complexity will be performed as future valuable work.

CRediT authorship contribution statement

Debora Gil: Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Investigation, Formal analysis, Conceptualization. **Aura Hernández-Sabaté:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Funded by Ministerio de Economía, Industria y Competitividad, Gobierno de España, Spain grant number PID2025-173648OB-I00, and CERCA Programme/Generalitat de Catalunya, Spain.

Data availability

The data comes from a public database.

References

[1] K. Kingphai, Y. Moshfeghi, Mental workload assessment using deep learning models from EEG signals: a systematic review, IEEE Trans. Cogn. Dev. Syst. (2024).

[2] Y. Zhou, S. Huang, Z. Xu, P. Wang, X. Wu, D. Zhang, Cognitive workload recognition using EEG signals and machine learning: A review, IEEE Trans. Cogn. Dev. Syst. 14 (3) (2022) 799–818, http://dx.doi.org/10.1109/TCDS.2021.3090217.

[3] J.G. Booth, Generalized linear models with random effects: Unified analysis via H-likelihood by Y. Lee, J. A. Nelder, and Y. Pawitan, Biom. 63 (4) (2007) 1296–1297.

[4] U. Pale, T. Teijeiro, D. Atienza, Importance of methodological choices in data manipulation for validating epileptic seizure detection models, in: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, 2023, pp. 1–7.

[5] Sajeev Kunjan, Tyler S. Grummett, Kenneth J. Pope, David M.W. Powers, Sean P. Fitzgibbon, T. Bastiampillai, M. Battersby, Trent W. Lewis, The necessity of leave one subject out (LOSO) cross validation for EEG disease diagnosis, International Conference on Brain Informatics, Springer, 2021, pp. 558–567.

[6] Thomas Barbera, Jacopo Burger, Alessandro D.'Amelio, Simone Zini, Simone Bianco, Raffaella Lanzarotti, Paolo Napoletano, Giuseppe Boccignone, Jose Luis Contreras-Vidal, On using AI for EEG-based BCI applications: problems, current challenges and future trends, Int. J. Hum.–Comput. Interact. (2025) 1–20.

[7] Richard James Sugden, Phedias Diamandis, Generalizable electroencephalographic classification of Parkinson's disease using deep learning, Inform. Med. Unlocked 42 (2023) 101352.

[8] Federico Del Pup, Andrea Zanola, Louis Fabrice Tshimanga, Alessandra Bertoldo, Livio Finos, Manfredo Atzori, The role of data partitioning on the performance of EEG-based deep learning models in supervised cross-subject analysis: a preliminary study, Comput. Biol. Med. 196 (2025) 110608.

[9] A. Abdelhameed, M. Bayoumi, A deep learning approach for automatic seizure detection in children with epilepsy, Front. Comput. Neurosci. 15 (2021) 29, http://dx.doi.org/10.3389/FNCOM.2021.650050.

[10] S. Chakrabarti, A. Swetapadma, P.K. Pattnaik, A channel independent generalized seizure detection method for pediatric epileptic seizures, Comput. Methods Programs Biomed. 209 (2021) 106335, http://dx.doi.org/10.1016/J.CMPB.2021.106335.

[11] Z. Gao, X. Wang, Y. Yang, C. Mu, Q. Cai, W. Dang, S. Zuo, EEG-based spatio-temporal convolutional neural network for driver fatigue evaluation, IEEE Trans. Neural Netw. Learn. Syst. 30 (9) (2019) 2755–2763, http://dx.doi.org/10.1109/TNNLS.2018.2886414.

[12] M. Plechawska-Wójcik, M. Tokovarov, M. Kaczorowska, D. Zapała, A three-class classification of cognitive workload based on EEG spectral data, Appl. Sci. 9 (24) (2019) 5340, http://dx.doi.org/10.3390/app9245340.

[13] E.Q. Wu, X.Y. Peng, C.Z. Zhang, J.X. Lin, R.S. Sheng, Pilots' fatigue status recognition using deep contractive autoencoder network, IEEE Trans. Instrum. Meas. 68 (10) (2019) 3907–3919, http://dx.doi.org/10.1109/TIM.2018.2885608.

[14] A. Ahmadi, H. Bazregarzadeh, K. Kazemi, Automated detection of driver fatigue from electroencephalography through wavelet-based connectivity, Biocybern. Biomed. Eng. (2020) http://dx.doi.org/10.1016/j.bbe.2020.08.009.
[15] P. Zhang, X. Wang, J. Chen, W. You, W. Zhang, Spectral and temporal feature learning with two-stream neural networks for mental workload assessment, IEEE Trans. Neural Syst. Rehabil. Eng. 27 (6) (2019) 1149–1159.
[16] J. West, Z.D. Bozorgi, J. Herron, H.J. Chizeck, J.D. Chambers, L. Li, Machine learning seizure prediction: one problematic but accepted practice, J. Neural Eng. 20 (1) (2023) 016008.
[17] M. Zhou, C. Tian, R. Cao, B. Wang, Y. Niu, T. Hu, H. Guo, J. Xiang, Epileptic seizure detection based on EEG signals and CNN, Front. Neuroinform. 12 (2018) 95, http://dx.doi.org/10.3389/fninf.2018.00095.
[18] B. Gao, J. Zhou, Y. Yang, J. Chi, Q. Yuan, Generative adversarial network and convolutional neural network-based EEG imbalanced classification model for seizure detection, Biocybern. Biomed. Eng. 42 (1) (2022) 1–15, http://dx.doi.org/10.1016/J.BBE.2021.11.002.
[19] Z. Sun, B. Li, F. Duan, H. Jia, S. Wang, Y. Liu, A. Cichocki, C.F. Caiafa, J. Sole-Casals, WLnet: Towards an approach for robust workload estimation based on shallow neural networks, IEEE Access 9 (2020) 3165–3173.
[20] M.S. Hossain, S.U. Amin, M. Alsulaiman, G. Muhammad, Applying deep learning for epilepsy seizure detection and brain mapping visualization, ACM Trans. Multimed. Comput. Commun. Appl. 15 (1s) (2019) http://dx.doi.org/10.1145/3241056.
[21] J. Zhou, L. Liu, Y. Leng, Y. Yang, B. Gao, Z. Jiang, W. Nie, Q. Yuan, Both cross-patient and patient-specific seizure detection based on self-organizing fuzzy logic, Int. J. Neural Syst. 32 (6) (2022) http://dx.doi.org/10.1142/S0129065722500174.
[22] J. Turner, A. Page, T. Mohsenin, T. Oates, Deep belief networks used on high resolution multichannel electroencephalography data for seizure detection, in: Proc. AAAI Spring Symp. Ser, 2014.
[23] M.A. Almogbel, A.H. Dang, W. Kameyama, Cognitive workload detection from raw EEG-signals of vehicle driver using deep learning, in: Proc. 21st Int. Conf. Adv. Commun. Technol. (ICACT), IEEE, 2019, pp. 1–6.
[24] H. Zeng, X. Li, G. Borghini, Y. Zhao, P. Aricò, G. Di Flumeri, N. Sciaraffa, W. Zakaria, W. Kong, F. Babiloni, An EEG-based transfer learning method for cross-subject fatigue mental state prediction, Sensors 21 (7) (2021) 2369.
[25] Yu Zhou, Yuxin Gao, Qiang Li, Ruiheng Wu, Aiping Yang, Ming-Lang Tseng, The interpretable deep learning framework and validation for seizure detection in pediatric electroencephalography: An improved accuracy and performance analysis, Artif. Intell. Med. (2025) 103276.
[26] V.J. Lawhern, A.J. Solon, N.R. Waytowich, S.M. Gordon, C.P. Hung, B.J. Lance, EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces, J. Neural Eng. 15 (5) (2018) 056013, http://dx.doi.org/10.1088/1741-2552/AACE8C.
[27] Zuochen Wei, Junzhong Zou, Jian Zhang, Jianqiang Xu, Automatic epileptic EEG detection using convolutional neural network with improvements in time-domain, Biomed. Signal Process. Control. 53 (2019) 101551.
[28] Jeff Craley, Emily Johnson, Christophe Jouny, Archana Venkataraman, Automated inter-patient seizure detection using multichannel convolutional and recurrent neural networks, Biomed. Signal Process. Control. 64 (2021) 102360.
[29] V. Shah, E. Von Weltin, S. Lopez, J.R. McHugh, L. Veloso, M. Golmohammadi, I. Obeid, J. Picone, The temple university hospital seizure detection corpus, Front. Neuroinform. 12 (2018) 83.
[30] G. Xu, T. Ren, Y. Chen, W. Che, A one-dimensional CNN-LSTM model for epileptic seizure recognition using EEG signal analysis, Front. Neurosci. 14 (2020) 1253, http://dx.doi.org/10.3389/FNINS.2020.578126/BIBTEX.
[31] Y. Li, Z. Yu, Y. Chen, C. Yang, Y. Li, X. Allen Li, B. Li, Automatic seizure detection using fully convolutional nested LSTM, Int. J. Neural Syst. 30 (4) (2020) http://dx.doi.org/10.1142/S0129065720500197.
[32] G. Jiang, H. Chen, C. Wang, P. Xue, Mental workload artificial intelligence assessment of pilots' EEG based on multi-dimensional data fusion and LSTM with attention mechanism model, Int. J. Pattern Recognit. Artif. Intell. 36 (11) (2022) 2259035.
[33] D.D. Chakladar, S. Dey, P.P. Roy, D.P. Dogra, Eeg-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm, Biomed. Signal Process. Control. 60 (2020) 101989.
[34] R.T. Schirrmeister, J.T. Springenberg, L.D.J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for EEG decoding and visualization, Hum. Brain Mapp. 38 (11) (2017) 5391–5420, http://dx.doi.org/10.1002/HBM.23730.
[35] X. Wang, X. Wang, W. Liu, Z. Chang, T. Kärkkäinen, F. Cong, One dimensional convolutional neural networks for seizure onset detection using long-term scalp and intracranial EEG, Neurocomputing 459 (2021) 212–222, http://dx.doi.org/10.1016/J.NEUCOM.2021.06.048.
[36] C. Hua, H. Wang, J. Chen, T. Zhang, Q. Wang, W. Chang, Novel functional brain network methods based on CNN with an application in proficiency evaluation, Neurocomputing 359 (2019) 153–162.
[37] C. Chatzichristos, J. Dan, A.M. Narayanan, N. Seeuws, K. Vandecasteele, M. De Vos, A. Bertrand, S. Van Huffel, Epileptic seizure detection in EEG via fusion of multi-view attention-gated U-net deep neural networks, in: 2020 IEEE Signal Processing in Medicine and Biology Symposium, SPMB, IEEE, 2020, pp. 1–7.
[38] Y. Yuan, G. Xun, K. Jia, A. Zhang, A multi-view deep learning framework for EEG seizure detection, IEEE J. Biomed. Health Inform. 23 (1) (2018) 83–94.
[39] A. Hernández-Sabaté, J. Yauri, P. Folch, M. Piera, D. Gil, Recognition of the mental workloads of pilots in the cockpit using EEG signals, Appl. Sci. 12 (5) (2022) 2298.
[40] P. PhysioBank, Physionet: comp onents of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.
[41] A.H. Shoeb, Application of machine learning to epileptic seizure onset detection and treatment, (Ph. D. thesis), Massachusetts Institute of Technology, 2009.
[42] Balu Krishnan, Simon Tousseyn, Chetan Sateesh Nayak, Thandar Aung, Ammar Kheder, Z. Irene Wang, Guiyun Wu, Jorge Gonzalez-Martinez, Dileep Nair, Richard Burgess, et al., Neurovascular networks in epilepsy: correlating ictal blood perfusion with intracranial electrophysiology, Neuroimage 231 (2021) 117838.
[43] Elisa Bruno, Andrea Biondi, Mark P. Richardson, RADAR-CNS Consortium, et al., Pre-ictal heart rate changes: a systematic review and meta-analysis, Seizure 55 (2018) 48–56.
[44] Y. Gao, B. Gao, Q. Chen, J. Liu, Y. Zhang, Deep convolutional neural network-based epileptic electro encephalogram (EEG) signal classification, Front. Neurol. 11 (2020) 375, http://dx.doi.org/10.3389/FNEUR.2020.00375/BIBTEX.
[45] S. Toraman, Automatic recognition of preictal and interictal EEG signals using 1D-capsule networks, Comput. Electr. Eng. 91 (2021) 107033, http://dx.doi.org/10.1016/J.COMPELECENG.2021.107033.
[46] F.E. Grubbs, Sample Criteria for Testing Outlying Observations, University of Michigan, 1949.
[47] R.S. Fisher, A.M. Bonner, The revised definition and classification of epilepsy for neuro diagnostic technologists, Neurodiagnostic J. 58 (1) (2018) 1–10, http://dx.doi.org/10.1080/21646821.2018.1428455.
[48] J.J. Falco-Walter, I.E. Scheffer, R.S. Fisher, The new definition and classification of seizures and epilepsy, Epilepsy Res. 139 (2018) 73–79, http://dx.doi.org/10.1016/J.EPLEPSYRES.2017.11.015.
[49] S. Aydin, Determination of autoregressive model orders for seizure detection, Turk. J. Electr. Eng. Comput. Sci. 18 (1) (2010) 23–30.
[50] M. Zabihi, S. Kiranyaz, V. Jantti, T. Lipping, M. Gabbouj, Patient-specific seizure detection using nonlinear dynamics and nullclines, IEEE J. Biomed. Health Inform. 24 (2) (2019) 543–555.
[51] L.A. Moctezuma, M. Molinas, EEG channel-selection method for epileptic-seizure classification based on multi-objective optimization, Front. Neurosci. 14 (2020) 593, http://dx.doi.org/10.3389/FNINS.2020.00593/BIBTEX.
[52] E. Ali, M. Angelova, C. Karmakar, Epileptic seizure detection using CHB-MIT dataset: The overlooked perspectives, R. Soc. Open Sci. 11 (5) (2024) 230601.
[53] H. Ji, T. Xu, T. Xue, T. Xu, Z. Yan, Y. Liu, B. Chen, W. Jiang, An effective fusion model for seizure prediction: Gamrnn, Front. Neurosci. 17 (2023).
[54] N.D. Truong, L. Kuhlmann, M.R. Bonyadi, D. Querlioz, L. Zhou, O. Kavehei, Epileptic seizure forecasting with generative adversarial net- works, IEEE Access 7 (2019) 143999–144009.
[55] J. Zhang, S. Zheng, W. Chen, G. Du, Q. Fu, H. Jiang, A scheme combining feature fusion and hybrid deep learning models for epileptic seizure detection and prediction, Sci. Rep. 14 (1) (2024) 16916.
[56] N. Lai, Z. Li, C. Xu, Y. Wang, Z. Chen, Diverse nature of interictal oscillations: EEG-based biomarkers in epilepsy, Neurobiol. Dis. 177 (2023) 105999.
[57] Ijaz Ahmad, Chen Yao, Lin Li, Yan Chen, Zhenzhen Liu, Inam Ullah, Mohammad Shabaz, Xin Wang, Kaiyang Huang, Guanglin Li, et al., An efficient feature selection and explainable classification method for EEG-based epileptic seizure detection, J. Inf. Secur. Appl. 80 (2024) 103654.
[58] S. Poorani, Pjijosae Balasubramanie, Deep learning based epileptic seizure detection with EEG data, Int. J. Syst. Assur. Eng. Manag. (2023) 1–10.
[59] Yanna Zhao, Jiatong He, Fenglin Zhu, Tiantian Xiao, Yongfeng Zhang, Ziwei Wang, Fangzhou Xu, Yi Niu, Hybrid attention network for epileptic EEG classification, Int. J. Neural Syst. 33 (06) (2023) 2350031.