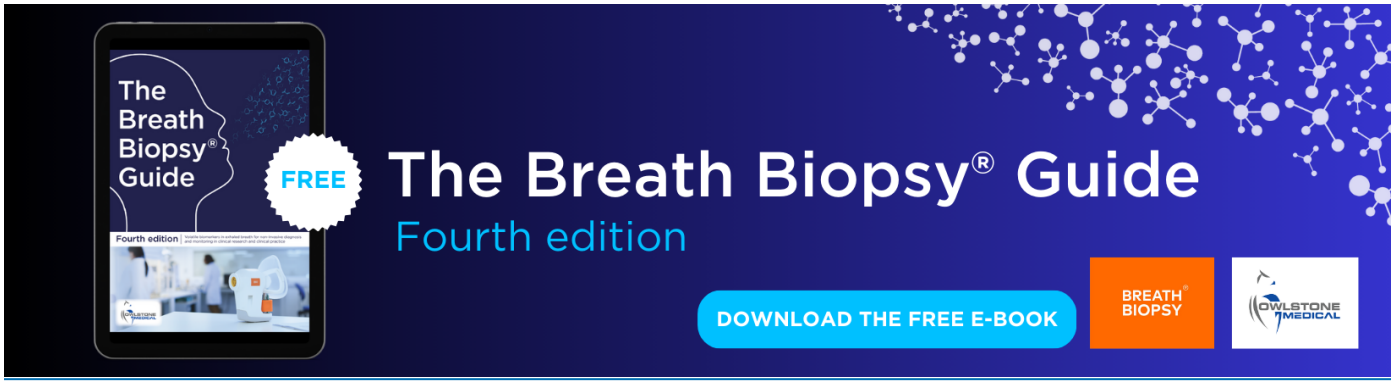
Rating by detection: an artifact detection protocol for rating EEG quality with average event duration



Journal of Neural Engineering

PAPER

Rating by detection: an artifact detection protocol for rating EEG quality with average event duration

Daniel Węsierski[1,*], **Mehrdad Rahimzadeh Rufuie**[2], **Olga Milczarek**[3], **Wojciech Ziembla**[2], **Paweł Ogniewski**[2], **Anna Kołodziejak**[2] and **Paweł Niedbalski**[2]

[1] Gdańsk University of Technology, Faculty of Electronics, Telecommunications, and Informatics, Gdańsk, Poland
[2] Elmiko Biosignals, Warsaw, Poland
[3] Department of Children's Neurosurgery, Jagiellonian University Medical College, Cracow, Poland
[*] Author to whom any correspondence should be addressed.

E-mail: daniel.wesierski@pg.edu.pl

Keywords: electroencephalography (EEG), artifact removal, average event duration, blind assessment, missing ground truth, rating by detection, real data

OPEN ACCESS

RECEIVED
10 May 2022

REVISED
1 February 2023

ACCEPTED FOR PUBLICATION
9 February 2023

PUBLISHED
22 March 2023

Abstract

Objective. Quantitative evaluation protocols are critical for the development of algorithms that remove artifacts from real electroencephalography (EEG) optimally. However, visually inspecting the real EEG to select the top-performing artifact removal pipeline is infeasible while hand-crafted EEG data allow assessing artifact removal configurations only in a simulated environment. This study proposes a novel, principled approach for quantitatively evaluating algorithmically corrected EEG without access to ground truth in real-world conditions. *Approach*. Our offline evaluation protocol uses a detector to score the presence of artifacts. It computes their average duration, which measures the recovered EEG's deviation from the modeled background activity with a single score. As we expect the detector to make generalization errors, we employ a generic and configurable Wiener-based artifact removal method to validate the reliability of our detection protocol. *Main results*. Quantitative experiments extensively compare many Wiener filters and show their consistent rankings agree with their theoretical assumptions and expectations. *Significance*. The rating-by-detection protocol with the average event duration measure should be of value for EEG practitioners and developers. After removing artifacts from real EEG, the protocol experimentally shows that reliable comparisons between many artifact filtering configurations are possible despite the missing ground-truth neural signals.

1. Introduction

Electroencephalography (EEG) records the electrical activity of the brain. It plays a key role in the diagnosis of many brain diseases. Scalp EEG is an indispensable tool for assessing seizure severity, training through biofeedback, monitoring sleep disorders, and deep brain stimulation. Intracranial EEG is the gold standard invasive approach for localizing epileptogenic zones during surgery. EEG interfaces the brain with a computer. Notwithstanding its versatility and unquestionable clinical significance, artifacts curtail the impact of EEG-based applications, thereby making EEG correction necessary.

Quantitative evaluation is critical to developing optimal artifact removal algorithms. The development phase should address numerous degrees of freedom of an artifact removal algorithm, such as hyperparameter settings, spatio-temporal locality models, and interference models of artifact classes with neurogenic signals. Evaluating event detection algorithms leverages manual data annotations and follows an established protocol for pattern recognition using, for example, the area under the precision-recall curves [1]. Tuning the hyperparameters of models and training procedures is feasible as detectors leave unaffected the input signals and should output labels that match prespecified labels. Unlike detection, artifact removal should correct noisy signals, leaving the recovered, hypothetically clean outputs with reference to neither the ground truth signals nor their true labels. Hence, this study addresses

the problem of evaluating the accuracy of one data-driven algorithm by another data-driven algorithm without access to the ground truth.

Evaluating a model on real data from different modalities without reference to true knowledge has been approached by taking advantage of another model. No-reference or blind image quality assessment [2] uses prior knowledge about image attributes and artifacts or learns a model for them from mean opinion scores. Since medical data can be ambiguous and noisy, the varying inter-rater agreement is evidence for the uncertain ground truth. Iterative refinement of labels from inaccurate raters improves model training and evaluation. Classifiers and noisy labels were optimized jointly via expectation-maximization [3] and further optimized on a small set of noise-free labels [4]. Text-to-speech synthesizers were evaluated for automatic speech recognition [5] by converting voices back to the text for text-to-text comparison. Supervised bootstrap was used successfully in [6], where a pair of positive–negative experts continuously retrained a detector of a moving object by exploiting video constraints. Common agreement [7] between image segmentation algorithms assessed the robustness and variability of the methods with respect to data and labels. The algorithms were trained in a supervised manner using manually annotated data. They then were evaluated automatically on data without ground truth under the premise that the result of one algorithm agrees with the votes of the remaining group of algorithms. Although the authors argued that the common agreement could not guarantee certainty that one algorithm outperforms the others, the principle was shown to be an unbiased estimator of the performance of algorithms as human raters validated the results. Reverse classification accuracy [8] is a rating protocol that adopts reverse testing [9] to quantitatively evaluate image segmentation algorithms on test sets without access to ground truth segmentation masks. A segmentation classifier was trained on the annotated images to label new images with no ground truth. The new labels, which came from the first classifier, served to train a reverse classifier. The hypothesis was that a good classifier should generate sufficiently good labels to train the reverse classifier to perform well at test time on the annotated training images of the first classifier.

In EEG, a traditional, scalable approach to evaluating artifact removal algorithms relies on simulated data by synthetically contaminating a clean signal. The method linearly mixes the clean signal with an instantiation of an artifact-like signal and then compares the algorithmically recovered signal to the clean signal [10–12]. It can clarify certain aspects of artifact correction algorithms, but [13] warns that EEG simulation methods are based on data and mixing models that may also be imprecise. Namely, if the cleaned EEG is unsatisfactory, a worse filtering result may be caused by the data simulation algorithm apart from an artifact correction algorithm.

Human raters often evaluate the performance of artifact removal algorithms on real EEG. Visual inspection of EEG after artifact correction is usually considered sufficient evidence for the efficacy of an artifact removal method [14, 15]. An EEG expert can assess the quality of EEG after artifact removal on a few-point scale. Manual artifact detection is based on the electroencephalographer's knowledge and the human eye's ability to register and mark changes in the EEG. However, manual ratings are time-consuming, subjective, and challenging to reproduce [16]. To visually inspect the corrected EEG and reduce variance in ratings, experts usually follow some questionnaires [17] and additionally analyze the results in the frequency domain [18, 19]. Employing more experts to rate the denoised EEG increases the objectivity of the manual evaluation. At the same time, multi-rater verification is hardly feasible in practice as it radically raises the costs and time of developing and fine-tuning artifact removal algorithms on real EEG data.

Quantitative comparisons of real artifacts to artifact-removed signals and artifact-removed signals to real, artifact-free signals have been proposed to overcome the dependence on laborious visual inspection [20]. After regressing real ocular artifacts to corrected data, a higher residual indicated better performance of ocular artifact removal in [13, 21]. This regression-based protocol measured artifact attenuation but not norm distortion, so residuals of least-squares regression in [22–24] quantified the similarities between intracerebral activity and signals with removed myogenic artifacts. In these equivalence tests, the instructed participants alternately tensed and relaxed their cranial muscles to alternate between myogenic and referential intracerebral activity. As regression-based methods require reference signals [25], in this study, we leverage a reference model of real artifacts and norm signals to evaluate corrected EEG fragments automatically.

We propose a scalable rating by detection protocol for the quantitative evaluation of EEG quality. The protocol uses a detector to automatically score ill-removed artifacts and measure EEG quality by the weighted duration of detected artifacts (figure 1). The underlying intuition behind the proposed evaluation procedure is that the better the artifact removal, the fewer artifacts should a multiclass detector find after correcting the EEG, and the more it should react to non-artifact classes of events. We train a multiclass classifier in a supervised manner using gradient boosting on manually labeled EEG recordings with normal background activity and ocular, muscle, and electrode-related artifacts, thus assuming that the EEG norm is the only non-artifact class. The discriminatively trained multiclass model assigns a probability-like distribution to corrected

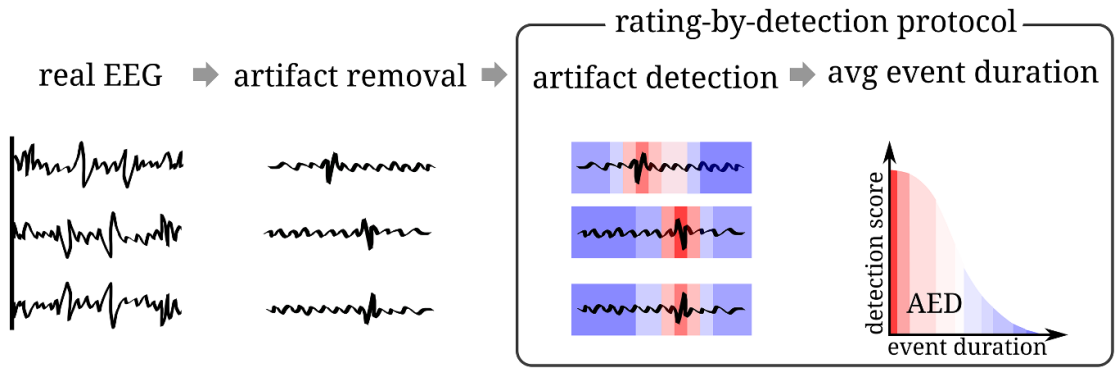


Figure 1. The proposed rating-by-detection protocol quantitatively rates the quality of EEG after removing real artifacts. As the ground truth neurogenic signals are unavailable after artifact removal, we propose automatically scoring ill-removed artifacts in the cleaned EEG to assess the performance of filtering algorithms in real-world conditions. Our threshold-free protocol computes the average event duration (AED) as the area under the curve that accumulates the duration of detected artifacts over the complete range of detection thresholds.

EEG instances over classes. The protocol then accumulates the duration of scored artifacts in the corrected EEG by progressively decreasing the detection thresholds over the entire probability range, making evaluation independent of any thresholds. The computed area under the curve (AUC) is the artifact average event duration (AED) that quantifies the presence of abnormal waveforms in the corrected EEG.

On simulated EEG data, two standard measures of accuracy of artifact removal algorithms are the signal-to-error ratio (SER) and artifact-to-residue ratio (ARR) [12]. The SER measures norm distortion, and the ARR measures artifact attenuation. However, evaluating the algorithm's accuracy with two complementary measures is less convenient than with a single measure. The ARR metric was extended to real data in [12] by substituting the true noise with the observed noisy signal. However, this substitution is valid for eye blink artifacts, which have a high amplitude, but numerous artifact classes generally can also have lower amplitudes. Moreover, SER and ARR require an oracle to partition an EEG epoch into clean and noisy multi-channel segments. Therefore, the SER- and ARR-based evaluation protocol is constrained only to these manually annotated EEG fragments and cannot scale to unlabeled data. The proposed rating by detection protocol for real EEG data jointly measures artifact attenuation and norm distortion with the single AED score to facilitate the quantitative evaluation of artifact removal. Provided that some classifier is trained on manually annotated smaller subset of these recordings, our protocol is then fully automatic and scales to an unlimited number of EEG recordings.

The closest work to ours is [15], which used a heuristic detector to evaluate EEG quality in real-world conditions. A differential evolution algorithm optimized the thresholds of an *if*-rule model of EEG features from manually labeled real EEG signals. The heuristic evaluation protocol required that all binary tests be satisfied jointly to classify an EEG epoch as a norm. The binary classifier showed good generalization on the test data but evaluated no artifact removal algorithms. In follow-up studies [16, 26, 27], a signal quality index (SQI) was defined using relaxed variants of the *if*-rules model of [15]. If an epoch passed more than a fixed percentage of tests, it was deemed as the norm in [26]. The SQI of EEG was defined as the mean of all epochs that were classified as artifacts to reject EEG trials from further analysis. Instead of thresholding the number of passed tests per epoch, the mean and standard deviation statistics of the percentage of passed tests per epoch determined an SQI of EEG recordings in [16, 27] to evaluate artifact removal algorithms. However, the reliability and usefulness of the SQI in designing and optimizing a particular artifact removal method has thus far not been addressed in the literature. The SQI compared different monolithic artifact removal algorithms on real EEG without delving into various design details and configurations of these algorithms. In contrast, we show that our verification protocol with AED measure is useful by gradually developing and configuring methods for filtering real artifacts. In this way, we follow a standard process of algorithm development that needs to evaluate various concepts and parameter settings to select the best one.

We argue that quantitative verification of cleaned EEG using a strong classifier of EEG events can provide useful feedback about the design and properties of artifact removal filters. Data-driven classifiers have been shown to approach expert-level performance in recognizing interictal [28] and ictal [29] events and sleep stages [30]. We propose using a classifier-based detector for artifact removal evaluation that uses binary decision trees trained in a gradient-boosting manner to yield soft scores according to the logistic regression function. Neither passing

a minimal number of tests [15, 26] nor heuristically normalizing the quotients of the passed tests [16, 27] is required. The trees of threshold-based rules and end scores are automatically learned from the labeled data. Our quantitative evaluation protocol can consistently rank various configurations of artifact removal algorithms in real EEGs. The detector densely scores the quotients of EEG waves to yield the AED of artifacts, which is our measure of EEG signal quality.

To validate the reliability and effectiveness of our rating protocol, we propose to fine-tune and configure a well-understood filtering method that works in the original space of EEG signals. We employ the state-of-the-art Wiener-based artifact removal method of [12]. It is a generic, multi-channel, and semi-automatic artifact removal method. The calibration of a Wiener filter manually partitions an EEG epoch into artifacts and non-artifacts. Options for optimizing filtering performance include selecting hyperparameters, such as filter delay and rank of artifact covariance. They configure a filter's structure and influence artifact removal's effectiveness, as shown in [12] on simulated and real EEG epochs.

We demonstrate that the proposed rating-by-detection protocol provides helpful feedback for developing an artifact removal algorithm for real EEG. We summarize our main contributions below:

- We develop a novel, offline *rating by detection* protocol that evaluates the accuracy of artifact removal algorithms in real-world conditions by measuring the quality of filtered real EEG signals. Our scalable rating procedure uses a detector trained in a gradient-boosting manner to score the presence of artifacts in every filtered EEG channel.
- We introduce the AED measure as a single-scored detection-based quantifier of signal quality. By jointly quantifying the norm distortion and artifact attenuation in the cleaned set of EEG recordings, AED is well suited for developing semi-automatic and automatic artifact removal algorithms, letting developers search for optimal filtering techniques effectively and efficiently.
- We validate the reliability and usefulness of our rating protocol with extensive experiments by fine-tuning and configuring a state-of-the-art Wiener filtering method. We show on hours-long EEG signals that (i) rating-by-detection protocol yields consistent rankings of differently configured Wiener filters that comply with the experimental results on synthetic and real data in [12], (ii) class-specific tuning of Wiener filter hyperparameters is better than class-agnostic tuning on real data, which agrees with intuition because artifact classes differ in waveform morphology, and that (iii) filter training has to be local for the best performance, which complies with the basic theoretical assumptions that underpin Wiener-based filtering.

2. Rating-by-detection protocol

We propose a detection protocol that quantitatively rates EEG quality using an EEG event detector and AED measure. The automatic protocol lets EEG practitioners and developers configure an optimal artifact removal algorithm for real EEG, as shown in figure 2. Our rating protocol is general. It can use any detector that scores ill-removed artifacts and background activity on some normalized scale, thereby automatically quantifying the accuracy of many artifact removal filters and selecting an optimal filtering configuration. It fixes no thresholds for detection. It evaluates an artifact removal method on real EEG with a single score as the AUC that determines the average duration of detected artifacts in corrected EEGs over the full range of detection thresholds.

2.1. Detection of artifacts

Our detection protocol is automatic, provided that EEG experts manually label EEG recordings beforehand with artifact classes that are subject to removal from EEG. In this study, a gradient-boosting classifier uses expert labels to train its parameters on selected EEG feature descriptors in a supervised manner. The detection protocol uses the trained discriminative model to assign soft scores from 0 to 1 to filtered EEG fragments. Better corrected EEG signals receive scores closer to 0, indicating that the morphology of the recovered signals resembles the morphology of manually labeled background activity. In the following sections, we describe a specific detector using decision trees but note that the rating-by-detection protocol can use any detector that scores EEG peaks in every channel.

2.1.1. Extracting features from EEG

Feature extraction in an EEG signal (figure 3(A)) commences by detecting positive and negative EEG peaks in every channel (figure 3(B)) with a peak detector. We define a peak as every local extremum that occurs within an EEG signal. At each peak location, the method instantiates a central window of duration 0.2 s and two side windows of duration 0.8 s to capture the temporal context of an EEG epoch (figure 3(C)). The total duration of the three windows is 1.8 s. In each window, the method computes a set of hand-crafted features that parameterize the morphological and spectral characteristics of the signal. The following features describe an EEG fragment:

- Time series anomaly score for the central window, defined by a linear model prediction error,
- Fast Fourier Transform (FFT) features (log power for frequency) in the central and neighborhood windows,
- Teager energy in the central and neighborhood windows,
- Quotient of waveform length in the central and neighborhood windows,

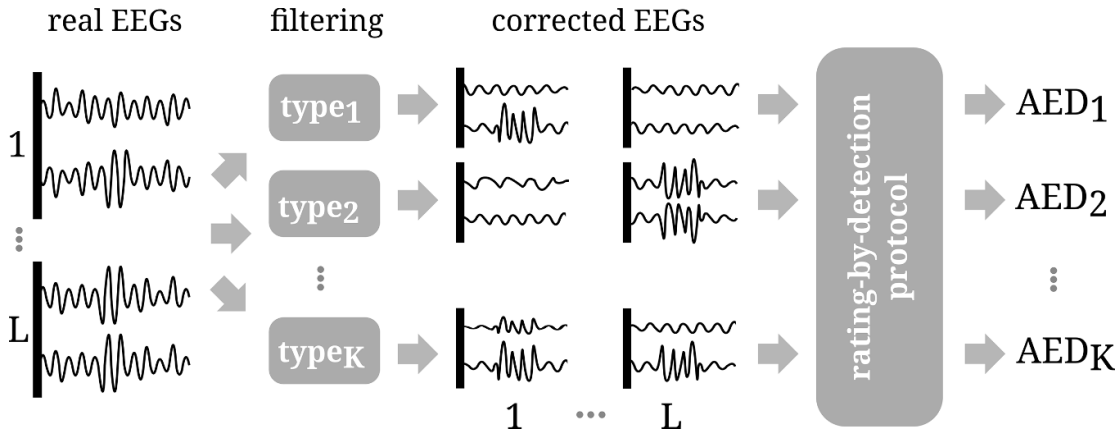


Figure 2. A general workflow for evaluating the accuracy of many types of artifact removal algorithms using our rating-by-detection protocol with the average event duration (AED) measure. A set of L recordings is subject to different types of artifact removal. The evaluated methods can be entirely different or refer to one method that admits various configurations and hyperparameter settings. Each of the K filtering types produces its own artifact-corrected set of L recordings. The protocol quantitatively evaluates each of such K sets. The filtering type with the minimal AED value is considered the most accurate in removing artifacts.

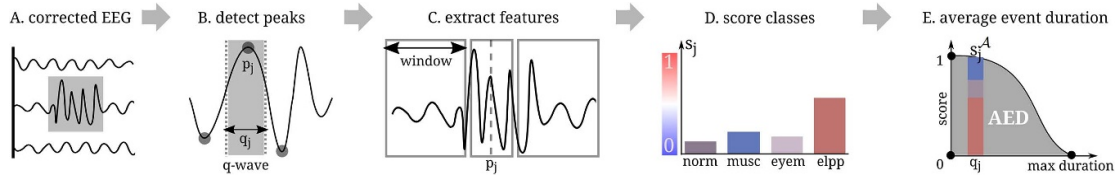


Figure 3. Rating-by-detection protocol. The proposed detection protocol rates the quality of artifact-corrected EEG channels (A) with the average event duration (AED) measure. Ill-removed artifact is denoted here by the gray box in the middle channel. After artifact removal, an EEG peak detector localizes peaks in every EEG channel (B). It determines the duration of quarter-waves (q-waves) between the middles of two consecutive pairs of peaks. The protocol requires a detector to score all q-waves for the presence of artifacts. Our specific detector uses a feature descriptor that locally parameterizes three neighboring EEG epochs (C) at every detected peak. A detector then classifies every three-window descriptor by distributing scores over predetermined event classes (D). Scores, distributed over artifact classes, are aggregated per q-wave. The q-waves are sorted in descending order according to their summed artifact scores to form a curve of accumulated artifact duration above the complete range of detection thresholds. The AUC amounts to the average artifact duration within the evaluated EEG signals (E).

- Standardized statistics (mean, standard deviation, skewness, min, max) of continuous wavelet (Ricker wavelet) transform coefficients for the central window; we use signal standardization with respect to the neighborhood windows,
- statistics (mean, standard deviation, skewness, min, max) of the EEG signal difference with lag 1 in the central window.

2.1.2. Training

The task of the classifier is to recognize background activity (`norm`) and three classes of artifact-related events, namely eye movement (`eyem`), muscle movement (`musc`), and electrode pop (`elpp`) in the feature space of EEG descriptors (section 2.1.1). Once J feature descriptors are extracted (section 2.1.1), we train a gradient-boosted classifier in the *one-vs-rest* supervised regime of XGBoost [31]. Logistic regression computes the mismatch error between the ground-truth labels y_j, which an EEG expert provides, and classifier predictions $\hat{y}_j$ as follows:

$$\sum_{j=1}^{J} y_j \ln(1+\exp(-\hat{y}_j)) + (1-y_j)\ln(1+\exp(\hat{y}_j)). \tag{1}$$

The maximum number of decision trees is 1000. The main configuration parameters for learning are maximal tree depth = 10; learning rate = 0.05; subsampling of examples = 0.5; and feature subsampling in each tree = 0.2. We trained XGBoost decision trees with early stopping.

2.1.3. Detection

The role of the detector at test time is to assign a score to a quarter-wave, hence the appellation q-wave. The softmax-based prediction s_j of the classifier of the jth q-wave distributes scores over four classes (figure 3(D)) that sum to unity. The duration $q_j > 0$ of the jth q-wave is determined by the centers between consecutive temporal locations of pairs of peaks (figure 3(B)):

$$q_j = \frac{1}{2}(p_{j+1} - p_{j-1}). \tag{2}$$

This detection procedure classifies all q-waves that are of interest to the user. In effect, the single-channel event detector semantically segments events in each EEG channel.

2.2. AED

We address the problem of quantifying artifact removal accuracy by summarizing the output scores of the detector. A hypothetical ideal artifact removal filter should clean the entire EEG recording of all artifacts, leaving only the true norm and other non-artifact-related events, such as pathological transients, seizures, or evoked potentials. This study assumes that only normal background activity and artifacts are present in the original EEG. Thus, the detector should score the norm more favorably than artifacts in the cleaned EEG. This notion is quantified by computing the total score $s_j^{\mathcal{A}} \in [0, 1]$ for all artifact classes $\mathcal{Z}_{\mathcal{A}}$ at the jth q-wave:

$$s_j^{\mathcal{A}} = \sum_{z \in \mathcal{Z}_{\mathcal{A}}} s_j^z. \tag{3}$$

Many EEG channels can be free of artifacts. An artifact removal algorithm should correct only the contaminated epochs per channel and omit background activity. Let N be the number of corrected EEG recordings with M channels. Let J_f denote the total number of q-waves in the contaminated epochs of the N recordings. The aggregated duration of all the contaminated epochs bounds the maximal duration $Q_{\max} > 0$ of J_f q-waves as:

$$Q_{\max} = \sum_{j=1}^{J_f} q_j. \tag{4}$$

Then, the total duration $Q(t)$ of artifact-related q-waves with a score of at least t is:

$$Q(t) = \sum_j q_j \quad \text{s.t.} \quad s_j^{\mathcal{A}} \geqslant t \tag{5}$$

such that $Q(t > 1) = 0$ and $Q(0) = Q_{\max}$.

Computing $Q(t)$ for thresholds t starting at $t = 1$ and descending to $t = 0$ with some fixed step $\delta \ll 1$ yields a monotonic, stepwise curve of threshold-weighted artifact duration, as shown in figure 3(E), that visually summarizes the quality of the corrected EEG fragments. The curve is bounded by $Q_{\max}$. A very accurate artifact removal method would produce high-quality EEG, having well-attenuated or no artifact-like morphology. In this case, its curve would have a maximal amplitude much lower than $t = 1$.

The rating protocol computes the AUC to quantify the overall accuracy of a given artifact removal technique. Therefore, the detection protocol depends on the performance of a particular artifact detector but not on a specific threshold of detection. The interpretation of the AUC is the AED in the filtered biosignal, defined by the threshold-weighted sum of the duration of J_f q-waves:

$$\text{AED} = \sum_{t>0}^{1} t(Q(t) - Q(t + \delta)). \tag{6}$$

The AED jointly measures the attenuation of artifacts and preservation of brain activity. At higher thresholds t, the function $Q(t)$ measures the total time of high-scoring artifacts. At lower thresholds t, the difference $Q_{\max} - Q(t)$ measures the total time of well-attenuated artifacts and norm distortion.

3. Validation of rating-by-detection protocol

We validate the reliability and effectiveness of our detection protocol on a dataset of hours-long EEG recordings that contain the most common artifact classes. We use the state-of-the-art Wiener-based artifact filtering method [12] as our testbed. The filtering method requires the manual selection of training data for computing filter parameters. To validate our protocol at scale, we use the detector to automatically segregate the EEG into multiple classes of artifact and norm segments, which is the same detector that rates EEG quality after filtering. Although the detection and filtering algorithms work differently and use different data-driven models, evaluating the filtering algorithm's effectiveness depends on the detector's reliability before and after filtering. Before filtering, some semantic masks may be classified incorrectly. After filtering, the true norm signal in effect could be distorted by artifact filtering in epochs that the detector incorrectly classified as artifacts. In our experiments, we train (section 2.1.2) two types of classification models for detection on different sets of recordings to evaluate four configurations of Wiener-based filtering. The filters were trained locally and globally. We use the workflow from figure 2 in our experiments on the testbed. In this way, our validation encapsulates a standard process of developing an artifact removal algorithm, in which one has to validate and fine-tune multiple concepts of an algorithm. Our results demonstrate that rating artifact removal on real EEG data complies with the theoretical assumptions that underpin the Wiener filters. Both detectors, though with models trained on two different datasets and misclassifying events differently, still lead to similar rankings of filters, trained with different hyperparameters, and to similar insights about optimal configurations for training the filters.

3.1. Dataset

Our dataset (table 1) consists of 39 real EEG recordings (Banana2 montage) that were selected from the publicly available TUHv1.0 EEG Artifact Corpus [32]. The dataset has a sampling rate of 250 Hz and

Table 1. Summary of the dataset. The first two rows refer to two different recording sets (A, B) that EEG experts manually labeled. The last two rows refer to the same set of recordings (B), automatically labeled by two detectors (A/B, B/B). The columns indicate types of sets, the number of recordings in each set, their total duration (multiplied by 18 channels), and the total number $\times 10^3$ (top row in the cell) and duration $\times 10^3$ [s] (bottom row in the cell) of four event classes.

set	#rec	dur (h)	norm	musc	eyem	elpp
A-exp	31	154.5	40	63	7.0	4.5
			11	17	1.6	1.3
B-exp	8	51.3	31	2.7	1.4	2.6
			8.3	0.8	0.3	0.7
A/B-det	8	51.3	143.8	15.5	17.8	7.5
B/B-det	8	51.3	157.6	9.6	3.9	13.6

16-bit resolution. The duration measurement treats each channel separately because our detector scores EEG q-waves independently in every channel. In our experiments, the 39 recordings were split into 31 recordings of 154.5 h in set A and 8 recordings of 51.3 h in set B.

3.1.1. Event classes

The dataset contains background activity (`norm`) and five categories of artifacts: eye movement (`eyem`), electrode pop, electrode static, and lead artifacts (`elpp`), muscle artifacts (`musc`), chewing, and shivering. We assigned chewing and shivering to `musc` artifacts as they are the least represented classes in the dataset and are related to muscle movements. In brain-computer interfaces, the most frequent `eyem` and `musc` artifacts testify to the physical condition of the subjects, where strictly controlling the acquisition settings can reduce the occurrence of the `elpp` artifacts. Involuntary movements of epileptic children and adults can severely move the attached electrodes and obscure the EEG background with frequent, high-amplitude `elpp` artifacts hindering the clinical EEG interpretation. Hence, we validated our detection protocol on `norm` and three classes of artifacts: `musc`, `eyem`, and `elpp`, which frequently occur in EEG [17].

3.1.2. Manual labeling

The selected epochs from sets A and B were labeled by EEG experts, thus forming the A-exp and B-exp sets. For each of the four event classes in the A-exp and B-exp sets, the top row of table 1 shows the number of annotated q-waves ($\times 10^3$), summarized over 18 channels, and the bottom row shows the total duration of q-waves ($\times 10^3$ [s]). The sets A-exp and B-exp train the detectors and evaluate their classification accuracy using ground truth labels from experts. Our EEG expert visually inspected all recordings and re-labeled the originally imprecise segments in time and semantics to ensure a higher ground truth quality. However, we note that determining the precise start and end of an EEG artifact is subjective and challenging. Generally, errors in EEG segmentation are common among experts and affect inter-rater agreement measures [33].

3.1.3. Automatic labeling

The sets A/B-det and B/B-det were automatically labeled in the entire EEG recordings from set B by A/B and B/B detectors, respectively, using our detection procedure from section 2.1. We use two different detectors to analyze their agreements in selecting the best filtering configurations under the AED measure. The total duration of automatically labeled q-waves with four classes equals the total duration of the recordings from set B. The duration of labels from B/B-det is considerably longer than from B-exp. The A/B detector is trained on the manually labeled data from set A, and the B/B detector is trained on the manually labeled data from set B. The last two rows of table 1 show the total duration of event classes ($\times 10^3$ [s]). The sets A/B-det and B/B-det evaluate the artifact removal accuracy of Wiener filters without ground truth by automatically detecting the ill-removed artifacts.

3.2. Testbed

A reliable protocol that convincingly evaluates EEG quality without access to ground truth is necessary to effectively inform developers what makes an effective artifact removal algorithm in real EEG data. Hyperparameter settings, spatio-temporal signal models, and modeling interference of artifact classes with neurogenic signals are examples of development directions that should be considered. We propose to fine-tune and configure a well-understood Wiener-based artifact removal method of [12] as a testbed for the reliability of the proposed rating-by-detection protocol. The filtering method (i) works in the original space of EEG signals, (ii) was evaluated extensively on simulated and real data in [12], (iii) has known theoretical properties, (iv) has interpretable hyperparameters, and (v) admits straightforward automation in this study.

We automate the method to validate the evaluation protocol at scale on 51.3 h of EEG channels, as shown in table 1, that count millions of q-waves. In effect, the quantitative results of our detection protocol should inform about the advantages and disadvantages of specific Wiener filter configurations that can be justified theoretically and experimentally by referring to the general properties of Wiener filtering.

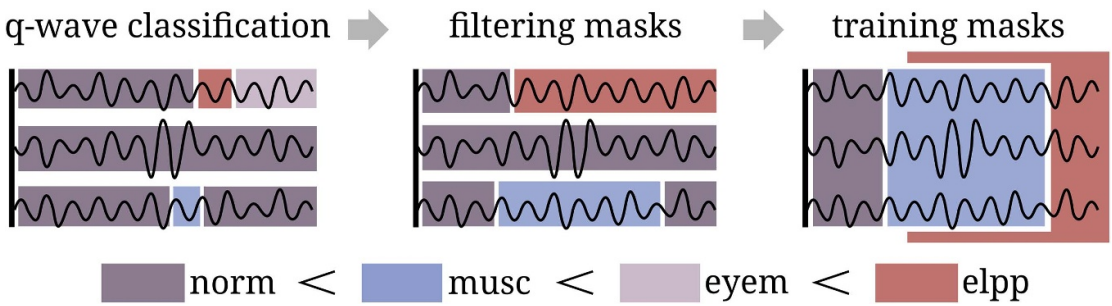


Figure 4. An example of sequential masking of real EEG for automatic training of Wiener filters [12] and for channel-wise filtering. Left: All q-waves of real EEG are classified with `argmax()` on detection scores. The adjacent q-waves of the same class are merged, likely forming fragmented, short semantic masks (colors encode classes) with insufficient data for training Wiener filters. Middle: An artifact mask is extended in both directions in its channel to increase the number of training data. As masks of different classes overlap after extension, we adopt a heuristic that ranks classes as shown at the bottom of the figure to create filtering masks that have zero overlaps. The `elpp` class is the most important class. It takes over some part of the `norm` and the whole `eyem` class in the top channel. Right: Training masks extend filtering masks over all channels to mark artifact and norm data for training temporally local, multi-channel Wiener filters. The norm mask is the residual of artifact training masks, and it never overlaps with them. The artifact training masks can overlap as one filter is trained independently from another filter across channels. In this example, the same norm data are used for training the `musc` and `elpp` Wiener filters. If the artifact training mask is longer than the past closest norm mask, the data from the other norm masks are aggregated from the past until their total duration exceeds its corresponding artifact mask.

For instance, one can expect that fitting a local filter to a local artifact should lead to better artifact removal accuracy throughout the whole EEG recording than learning a single, global filter as a canonical representation of all artifacts within the entire recording. Wiener filter estimation assumes the estimated signal is stochastically stationary. However, this assumption is not satisfied for EEG, which is a nonstationary signal [34]. Our protocol experimentally confirms that local Wiener filters are superior.

This section briefly reviews the Wiener-based artifact removal of [12] and retains its original notation. Using simple heuristics, we automate the Wiener filtering of a raw EEG recording with the help of the same detection method (section 2.1) that evaluates the quality of the filtered EEG. From the beginning to the end of an unprocessed EEG recording, the detector scores all q-waves and assigns the most probable class labels to them (figure 4). The q-wave labels integrate into longer, continuous semantic masks that segregate the real EEG into norm and artifact data blocks. Each artifact-related mask and the norm mask spawn a new Wiener filter. In the last step, the method applies Wiener filters only to channels masked as artifacts, leaving all channel-wise epochs under the norm masks unaltered throughout the recording.

3.2.1. Wiener filter in a nutshell

We briefly review the Wiener-based artifact removal of [12] to make the paper self-contained. Let an observed M-channel epoch of duration of $2\tau+1$ consecutive samples be denoted in a vector form as $y \in \mathbb{R}^k$, where $k = M(2\tau+1)$ and $\tau \in \mathbb{N}$ means time delay that symmetrically selects past and future samples at a given time instant. The observation is modeled in an additive manner as follows:

$$y = n + d \tag{7}$$

where $n \in \mathbb{R}^k$ represents the true neural signal and $d \in \mathbb{R}^k$ represents superimposed artifacts that can co-occur. The signals are preprocessed to zero-mean vectors. The signals n and d are assumed to be uncorrelated. Then, the covariances are governed by:

$$R_{yy} = R_{nn} + R_{dd} \tag{8}$$

where R_{yy}, R_{nn}, R_{dd} are defined as $E\{yy^T\}$, $E\{nn^T\}$, $E\{dd^T\}$, respectively, and $E\{\cdot\}$ is the expected value operator. The Wiener filter W of size $[k \times k]$ that estimates artifacts $\hat{d} = W^T y$ minimizes the mean squared error objective:

$$\min_W E\{\|d - W^T y\|^2\} \tag{9}$$

yielding $W = R_{yy}^{-1} R_{dd}$. In practice, the covariance matrices can be estimated from EEG segments Y_a and Y_c of size $[k \times T_a]$ and $[k \times T_c]$, respectively, as follows:

$$\hat{R}_{yy} = \frac{1}{T_a} Y_a Y_a^T \tag{10}$$

$$\hat{R}_{dd} = \hat{R}_{yy} - \hat{R}_{nn} = \hat{R}_{yy} - \frac{1}{T_c} Y_c Y_c^T \tag{11}$$

where T_a and T_c denote the number of EEG samples of the respective segment. The Wiener filter then amounts to:

$$\hat{W} = \hat{R}_{yy}^{-1} \hat{R}_{dd}. \tag{12}$$

Finally, using the additive model of equation (7), the neural EEG responses take the following approximated form:

$$\hat{n} = y - \hat{W}^T y. \tag{13}$$

The potential locations of artifacts in channels are class-specific. Ocular artifacts are bound to occur in

the frontal electrodes, while other types can occur at every site on the scalp. Moreover, artifacts typically corrupt only a subset of the EEG channels simultaneously. The number of artifacts affects the structure of covariance matrices $\hat{R}_{dd}$. The number of eigenvalues determines the subspace of a lower dimension of the matrices.

Designing the structure of the lower-dimensional subspace of $\hat{R}_{dd}$ can be approached using the generalized eigenvalue decomposition (GEVD) of $(\hat{R}_{yy}, \hat{R}_{nn})$. GEVD has been shown to improve Wiener-based filtering in [35]. The GEVD forces $\hat{R}_{dd}$ have a lower rank and be positive semi-definite. The low-rank approximation of $\hat{R}_{dd}$ decomposes the artifact covariance into generalized eigenvectors V and eigenvalues Σ_d as follows:

$$\hat{R}_{dd} = \hat{R}_{yy} - \hat{R}_{nn} = V^{-T}\Sigma_d V^{-1}. \tag{14}$$

The rank of $\hat{R}_{dd}$ can thus be controlled by editing the diagonal entries of Σ_d, with options including keeping some percentage of eigenvalues and keeping only positive eigenvalues or above other thresholds.

3.2.2. Training and filtering masks

Our procedure for automating the Wiener filtering uses simple heuristics. In [12], the EEG epochs of length T_a and T_c samples are determined by an expert that manually masks all EEG channels. In this study, we take advantage of the detector (section 2.1), which also rates the EEG quality after filtering, to compute the masks automatically. As shown in the example in figure 4, the detected event masks are often fragmented and short. Computing the covariance statistics using them would be unreliable. Hence, we define two types of masks: *training masks* and *filtering masks*. The main properties of the training and filtering masks are: (i) training masks expand over all channels to train the multi-channel Wiener filters, (ii) filtering masks cover only the channel fragments that are selected by the detector for artifact correction, (iii) filtering masks do not overlap because a specific Wiener filter can correct a given EEG fragment only once, (iv) artifact training masks can overlap, (v) norm training masks do not overlap with artifact training masks, (vi) a norm training mask is at least as long as the corresponding artifact training mask, (vii) the same norm training mask may repeatedly participate in training successive Wiener filters.

The method expands a short mask in a given channel with 2 s long margins to its left and right to merge the fragmented masks into continuous, longer masks. Depending on a particular filter training scenario, the method merges short masks either for each class of events separately (class-specific filtering) or jointly by treating all artifact classes as a single superclass (binary, class-agnostic filtering). The local Wiener filters are applied only once in a channel to a given EEG epoch. As the extended masks overlap in each channel, the adopted heuristic rule ranks event classes tentatively by their specific amplitudes following the order from the lowest to the highest amplitude, hence: `norm` → `musc` → `eyem` → `elpp`. The Wiener filters, tailored to the `elpp` class, have higher priority at the fragments where the `elpp` masks overlap with masks from three other classes in a given channel. If `musc` and `eyem` masks overlap, then the `eyem` mask has priority. After applying the ranking, we obtain filtering masks adjacent to each other. The artifact training masks extend the filtering masks over all channels to yield data Y_a to compute $\hat{R}_{yy}$ in equation (10). To select the norm data Y_c in a given epoch for computing $\hat{R}_{nn}$ in equation (11), the method requires all channels in the epoch to be detected as `norm`. However, the duration of detected `norm` in all channels may not be long enough to match at least the duration of the corresponding artifact training mask. Hence, the method aggregates past `norm` training masks until they exceed the duration of the corresponding artifact training mask. After training a filter for the artifact, the corresponding dimension of the filter output replaces the original, noisy channel under the corresponding filtering mask.

3.3. Quantitative evaluation of event detectors

The confidence in the AED results depends on the accuracy of event detection. We expect a detector to make classification errors on the test EEG data, which were unseen during the training of the detector. The amount of classification errors proportionally affects the reliability of our detection protocol. The protocol uses a detector to find artifacts in the filtered EEG recordings after correcting the real EEG recordings with a particular Wiener-based filtering configuration. The more artifacts are found, the lower the rated EEG quality and, thus, the lower the accuracy of a filtering method. Therefore, we propose to train two different detectors for the detection protocol and experimentally analyze whether they can indicate the same or similar improvements in the accuracy of different Wiener-based filtering configurations.

3.3.1. A/B detector

We assess the generalization ability of both detectors on the same test set B. We quantitatively evaluate the performance of both detectors on manually labeled q-waves that are unseen at the training time. Section 2.1.2 explains the training procedure of a detector. We trained the first detector, referred to as the *A/B detector*, on all manually labeled data from the set A-exp and tested it on all manually labeled test data from the set B-exp (table 1). The EEG channels were manually labeled with semantic masks of the four classes of events by EEG experts (section 3.1). At each q-wave from the sets A-exp and B-exp, detected by the peak detector, we extracted EEG features (figure 3) that served as our training and test sets, respectively.

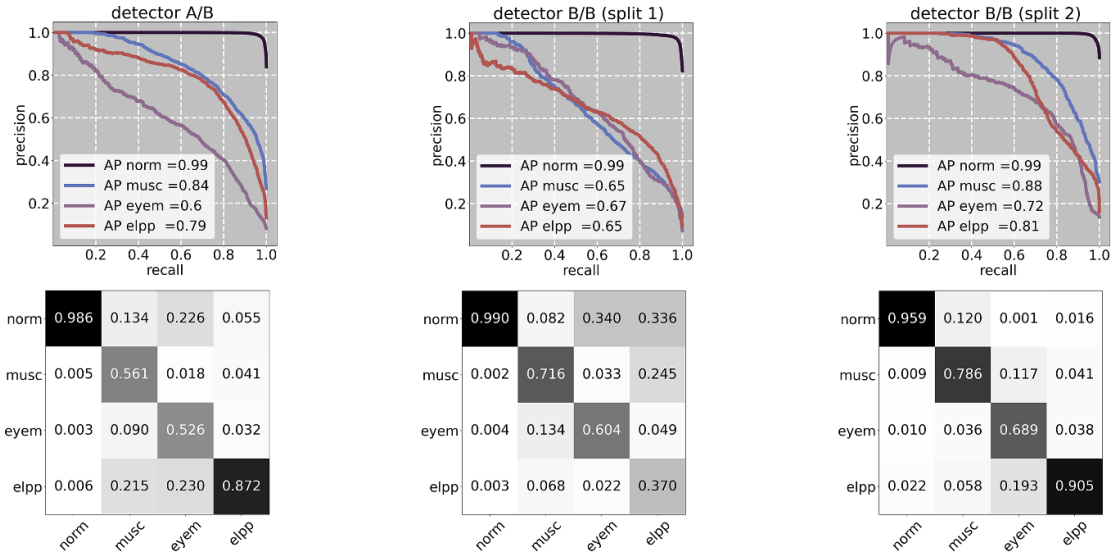


Figure 5. The precision-recall curves (top) and their confusion matrices (bottom) show the performance of detection of four classes. The average precision (AP) summarizes the accuracy of the A/B and B/B classifiers. The confusion matrices were normalized with respect to the predictions (columns). The first detector was trained on the A-exp set and evaluated on the B-exp set (left column). A 2-fold cross-validation evaluated the second detector on the B-exp set (middle and right columns). All three models almost perfectly recognized the norm in every test set, as indicated by AP scores of 0.99. The confusion matrices reveal that the true norm may be predicted as an artifact, but true artifacts are unlikely to be predicted as a norm.

3.3.2. B/B detector

The second detector, referred to as the *B/B detector*, was trained and evaluated in a 2-fold cross-validation manner with an equal number of examples per class over two splits of the set B-exp. The splits were created by randomly sampling the q-waves from the labeled fragments of the recordings. Hence, we enforced no temporal separation constraints for extracting EEG feature descriptors. A training example could be a close neighbor of a test example in an EEG fragment. To evaluate Wiener filters on the set B/B-det with the B/B detector, we trained its final form on all labeled q-waves from the set B-exp. In this way, we bias the trained model of the B/B detector to specialize it for automatic, class-specific masking of the recordings from set B.

3.3.3. Evaluation measures and results

As the classes are imbalanced in the set B-exp, the performance of the A/B and B/B detectors, respectively, is measured by the average precision (AP), which is the AUC of precision and recall, as shown in figure 5 (top). We gain additional insight into the performance of the detectors by computing the confusion matrices in figure 5 (bottom). As Wiener filters and the rating method depend on detection, the matrices are normalized with respect to the number of class-wise predictions (i.e. columns sum to unity).

3.3.4. Binary classification

The misclassification between norm and artifacts may negatively affect the accuracy of Wiener-based filtering. Training a Wiener filter is more sensitive to misclassifying artifacts as a norm than the norm as artifacts [12]. This Wiener filter property of asymmetric misclassification cost is advantageous for our detectors. As indicated by the confusion matrices, the predicted norm is rarely confused with the true artifact classes, as shown in the first column of the matrices. The detectors achieve high AP scores of 0.99 for the `norm` class. The detectors will thus train the filters well within the entire recording. However, they sometimes confuse the true norm with the predicted artifact classes, as shown in the first row of the confusion matrices. Some filters are learned unnecessarily when false artifacts are detected in the norm signal. In this case, the Wiener filters will distort the misclassified norm signals. The magnitude of this distortion partly depends on the Wiener filter training regime itself and partly on the correlation with the neighboring, correctly classified norm. The detection protocol jointly addresses artifact attenuation and norm distortion with the AED measure.

3.3.5. Multi-class classification

The misclassification within artifact classes may negatively affect the class-specific validation of our detection protocol but not class-agnostic (binary) validation. The mean AP of the artifact classes reached 0.73 over all three validations. The confusion matrices indicate that the artifact classes are confused mostly with each other. The `elpp` class has high and low EEG amplitudes leading to the highest intra-class variance. The low amplitudes of the `elpp` class often resemble `musc` artifacts, while the higher amplitudes are also characteristic for the `eyem` events. The predicted `musc` and `eyem` classes are confused mostly with the true `elpp` class, then with the true `norm` class. The detector

rarely confuses the eyem and musc with each other though in all three evaluations. Confusion of the predicted musc and eyem classes with the true norm class is more likely.

We conclude that both detectors discriminate well between the norm and artifacts but discriminating between artifact classes is more challenging. We attribute this limitation primarily to our single-channel EEG feature descriptor, which has insufficient capacity to accommodate discriminative patterns between artifact classes. In addition, the imprecise manual labeling of artifact segments also contributes to the measured class confusion. Given these results, we posit that our detectors can reliably rate the quality of cleaned EEG by summing the scores of the artifact classes (equation (3)). Hence, translating the multi-classification task for automatic initialization of Wiener filters into the binary classification task for rating EEG quality makes our detection protocol the most reliable.

3.4. Experiments and results for protocol

Some artifact classes may be harder to filter out than others, and generally, artifacts may require class-specific filter settings for more effective removal. The main factors that can optimize the performance of Wiener filters are filter delay τ and the rank of covariance R_{dd} of artifacts [12], which are defined in section 3.2.1. Our detection protocol confirms that in this section and further suggests that class-specific filter design with local filter training can further boost the performance of Wiener-based artifact removal on real EEG data.

We expect that local filter training estimates artifacts better than global filter training because EEG signals are only quasi-stationary within short epochs [34], and Wiener filters assume that noisy signals are stationary stochastic processes. Moreover, as artifacts vary significantly in morphology and duration, class-specific filter delays can discriminate between the spectral properties of artifacts and clean EEG spectra [12]. For example, the correlation of EEG samples of eye movement artifacts (eyem) extends over time more than the correlation of muscle movement (musc) and electrode-related artifacts (elpp), which are more chaotic.

To provide class-specific training masks for Wiener filters in the two evaluation variants, the detectors A/B and B/B score every q-wave in the entire EEG recordings from the same set B and produce sets of filtering masks A/B-det and B/B-det, respectively. We thus know that the A/B and B/B detectors disagree because they produce different labels, denoted by A/B-det and B/B-det, on set B, as shown in table 1. Both detectors agree indeed on the norm class. They recognize almost the same amount of norm in set B, with the quotient 0.91 of total norm duration between A/B-det and B/B-det sets. However, the quotients of the detected total duration of musc, eyem, elpp classes indicate higher disproportions and thus higher disagreement in q-wave classifications and amount to 1.61, 4.56, and 0.55, respectively.

3.4.1. Wiener-based filtering configurations

The experiments verified four configurations of Wiener filter training on the entire EEG test recordings (sets A/B-det and B/B-det). The filters were trained globally on the merged training masks or locally on each. Globally training the filters merges the data into two sets of norm and artifacts from all local training masks. For global and local training scenarios, artifact classes are kept separate or merged into a single artifact superclass, thereby realizing multiclass and binary filtering configurations, respectively. In this way, the validation of our detection protocol uses the following four Wiener-based filtering configurations: *global multiclass* G/M filters, *global binary* G/B filters, *local multiclass* L/M filters, and *local binary* L/B filters. The accuracy of the above four filtering configurations depends on the hyperparameter settings that our detection protocol measures with the average duration of artifacts in the artifact-corrected EEG recordings.

The detection protocol assesses the ability of G/M, G/B, L/M, and L/B filters to remove artifacts with the AED measure in figure 6. The delays that yielded minimal AEDs tuned the filters in figure 7. The average duration of artifacts for both detectors indicates that locally trained filters significantly outperform globally trained filters, leaving fewer high-scoring artifacts in the cleaned EEG. The results also indicate that class-specific filtering is better than binary filtering. However, the difference between class-specific and class-agnostic filter training is relatively less apparent.

3.4.2. Class-specific filter delays

Figure 7 illustrates a consistent ranking of filter delays on real EEG data. Both detectors show that the AED accuracy of the filters saturates with increasing delay, which [12] also observed. Moreover, the local training of Wiener filters significantly outperformed the global training in all scenarios and for all filter delays. Filters with no delay are the fastest to compute but consistently perform the worst. The detectors suggest that filters for the eyem class that contains more of lower frequencies than other classes require smaller delays than filters of other classes to remove artifacts more effectively. Interestingly, a delay of 17 for the local binary filter training led to minimal AED in both detection scenarios (A/B, B/B). For local training of class-specific filters, the best delays, which achieve minimal AED, vary between 9 and 17. Notably, the musc and elpp classes, which have lower intra-class temporal correlations than the eyem class, are best filtered with higher delays than the eyem class, as indicated by both detectors (figures 6 and 7). The delays at the knees of the curves vary less, with a mean delay of ≈5 for all three classes.

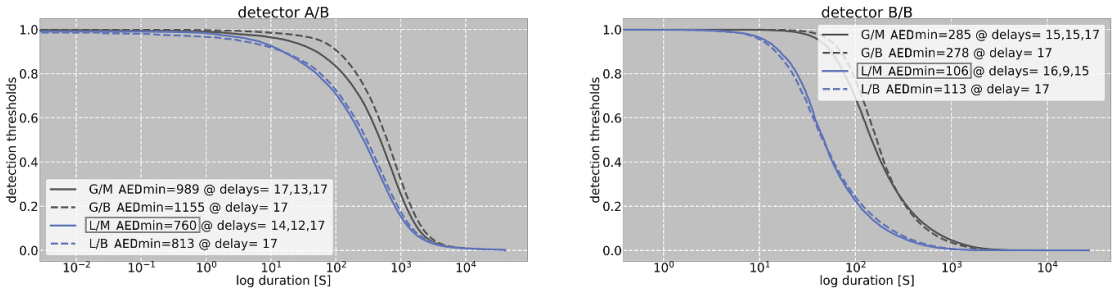


Figure 6. Accumulated duration of artifacts (log scale) in cleaned EEG with respect to the thresholded scores from detectors A/B (left) and B/B (right). The thresholds decrease from 1 to 0 with step $\delta = 10^{-3}$. The area under the curve is the average event duration (AED) of the artifacts in the cleaned EEG. The EEG recordings were corrected using four training variants of Wiener filters: global, multiclass (G/M), global, binary (G/B), local, multiclass (L/M), and local, binary (L/B) filters. The depicted curves refer to filters with the most optimal delay values (according to our AED measure). For multiclass filters, G/M and L/M, the ordering of the best three delays corresponds to `musc`, `eyem`, and `elpp`. Filter training keeps all the positive eigenvalues of the artifact covariances. Local filters correct real EEGs with evidently higher quality than global filters that cannot correctly estimate the covariance of norm and artifacts within an entire EEG recording. Moreover, the A/B and B/B detectors agree that the L/M filters work slightly better than the L/B filters.

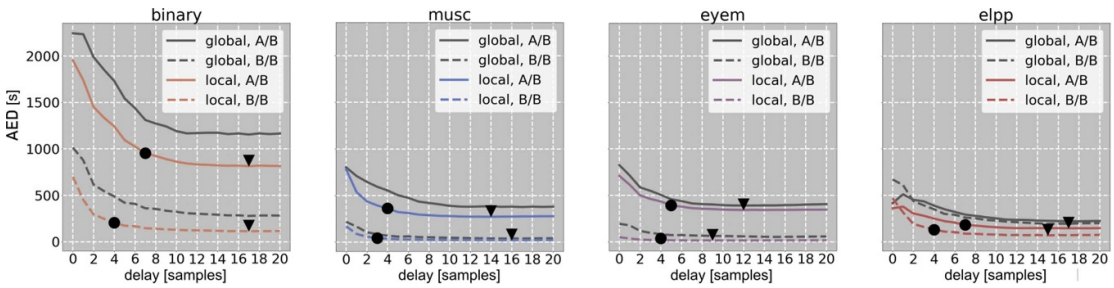


Figure 7. Average event duration (AED) measures the average duration of the artifacts that remain after applying Wiener filters with different delays. The artifacts are either merged into a single superclass (binary) or categorized into three classes: muscle movement (`musc`), eye movement (`eyem`), and related to electrodes (`elpp`). The filters were trained globally and locally. Artifacts were before and after artifact removal using the A/B and B/B detectors. The minimal AEDs of the local filter training are denoted by (▼) and the knees of the curves by (●). Both detectors, trained on different datasets, consistently agreed that AED minima and knees are achieved at similar delays.

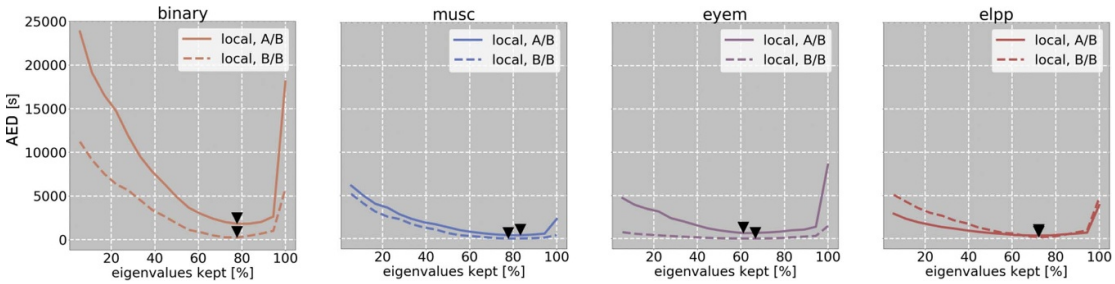


Figure 8. Average event duration (AED) measures the average duration of the artifacts that remain after applying Wiener filters with different ranks of the artifact covariance matrix. The Wiener filters are binary, class-specific, and trained with covariances of different ranks at a fixed filter delay of 15. Rank is equivalent to keeping the number of covariance eigenvalues. Both detectors A/B and B/B, which were trained on different datasets, consistently agreed that L/B and L/M filters achieve the best performance when the covariances of `musc`, `eyem`, `elpp` kept approximately 80%, 60%, and 70% of eigenvalues, respectively, as illustrated by the AED minima (▼).

3.4.3. Class-specific covariance ranks

Our experiments use real EEG data. Making assumptions about the expected rank of class-specific covariance matrices is hardly feasible in this setting. The detectors A/B and B/B search for the best structure of class-specific covariance matrices of artifacts to exploit the inherent pattern of artifact distribution over EEG channels. The detectors evaluate 18 rank-based configurations of covariance matrices for local class-specific filters that are trained with a fixed delay of 15, as shown in figure 8. Additionally, binary filters are evaluated to obtain a general, class-agnostic overview of the performance of Wiener-based filtering in this experiment. Each matrix configuration

retains only a certain percentage of eigenvalues, amounting to 18 types of filters for each of the four filtering configurations.

Despite different artifact misclassification errors of the A/B and B/B detectors, as indicated by apparent disproportions in the artifact classes in the A/B-det and B/B-det sets in table 1, both detectors point to almost the same percent of eigenvalues kept that attain the AED minima for classes musc and eyem class. The same percent of eigenvalues kept attains the minimum AED for the elpp class. The norm misclassification errors slightly differ between both detectors, but they still point to the same percent of eigenvalues kept to attain AED minima for the binary filtering configuration. The optimal number of retained eigenvalues differs across artifact classes. The eyem class requires the fewest eigenvalues, about 15%–20% fewer than the musc class, thereby confirming the findings of [12]. Keeping all eigenvalues (100%) implies that the GEVD is not used in optimizing filter parameters, but such training generally achieves poor performance. For the binary case, not using GEVD is better than keeping only several percent of all eigenvalues. For the multiclass case, though, the sensitivity of filtering performance to extreme rank settings is class-specific. Interestingly, for the eyem class, both detectors show that not using GEVD leads to worse results than retaining only a few percent of eigenvalues. We posit that this effect is specific to the eyem class because its occurrence is distributed locally over a few electrodes. In particular, the ocular artifacts, as opposed to muscle artifacts, can be well approximated with a small number of generalized eigenvectors and thus favor lower-rank filter approximations.

The number of maintained eigenvalues significantly affects the performance of all filters. In comparison to filters with varying delay parameters that keep all the positive eigenvalues of the artifact covariances using GEVD (figure 7), the AED-based filter performance varies significantly more with different ranks of artifact covariances (figure 8). Most importantly, training Wiener filters with locally adaptive rank selection outperforms Wiener filters with optimal but fixed ranks of their artifact covariances after comparing AED scores of filters with the delay of 15 in both figures. We attribute this result to dynamic artifact occurrence changes over time and channels. Electrodes can pop, and muscles may contract and relax locally and globally over channels, irrespective of the location on the scalp.

3.5. Qualitative comparison of filtering configurations

The quality of artifact removal was inspected visually by a neurological expert on the EEG recordings from the set A/B-det (table 1). The visual inspection of cleaned EEGs confirmed that the local filters L/M and L/B significantly outperform the global filters G/M, as shown in 5 section long epochs from the 3 EEG recordings in figure 9 (three columns). The top row of the figure shows the unprocessed, original EEG with the overlaid filtering masks (section 3.2.2), which indicated detected event classes (encoded by colors with the legend in the bottom left corner of the figure). The remaining rows of the figure show the filtered EEG after applying the filters L/M (second row), L/B (third row), and G/M (fourth row). The heatmap from 0 (norm) to 1 (artifact, after summing scores from all three artifact classes) that indicates the scores of the A/B detector is overlaid over the nine filtered EEG epochs. The global binary filter G/B produced the worst results, as indicated by both detectors A/B and B/B in figure 6, and it is omitted in the qualitative analysis. The filters are trained with the respective best delays (section 3.4.2) and by keeping only the positive eigenvalues of the artifact covariance matrices.

The L/M filters produce norm-like signals of higher quality than the other two filtering configurations, as shown in the heatmap with a low score in figure 9 (second row). The L/B filters performed slightly worse, as shown in figure 9 (third row). L/B and G/M filtering configurations tend to hallucinate signals that resemble pathological interictal events. This resemblance is particularly evident for the elpp artifacts in figure 9 (third column), where fake sharp wave groups and fake sharp wave–slow wave complexes emerge in the cleaned EEG. The A/B detector assigned high scores to these hallucinated fragments (as indicated by the heatmap). Using these filter training variants poses the risk of incorrect diagnosis and interpretation of activity caused by focal seizure lesions. The shape of the cleaned EEG after L/M filtering resembles the shape of the signals filtered by the L/B and G/M filters to some extent. However, the detector assigns lower scores to these fragments because their observable amplitudes are low. In this case, the lower amplitudes of the signal did not negatively affect EEG inspection either. Finally, the L/M filters corrected the EEG (first row) by attenuating high amplitude elpp activity, which generally obscures the background, making it difficult to interpret by a clinician.

Apart from optimizing Wiener filters with the above data configuration schemes, the performance of Wiener filters also depends on hyperparameter settings. As shown in figure 10, we were interested in how well the filtering performance from sections 3.4.2 and 3.4.3, indicated by the AED measure, aligns with the temporal and spectral characteristics of the corrected EEG. We analyzed the impact of several delays (top panel) and artifact covariance ranks (bottom panel) on filtering musc artifacts based on temporal and spectral properties of a four-channel, two section long EEG epoch. In the first experiment, the delay was increased while keeping all positive eigenvalues of the artifact covariance matrix. In the second experiment, the filter delay was fixed to 15 samples, and the

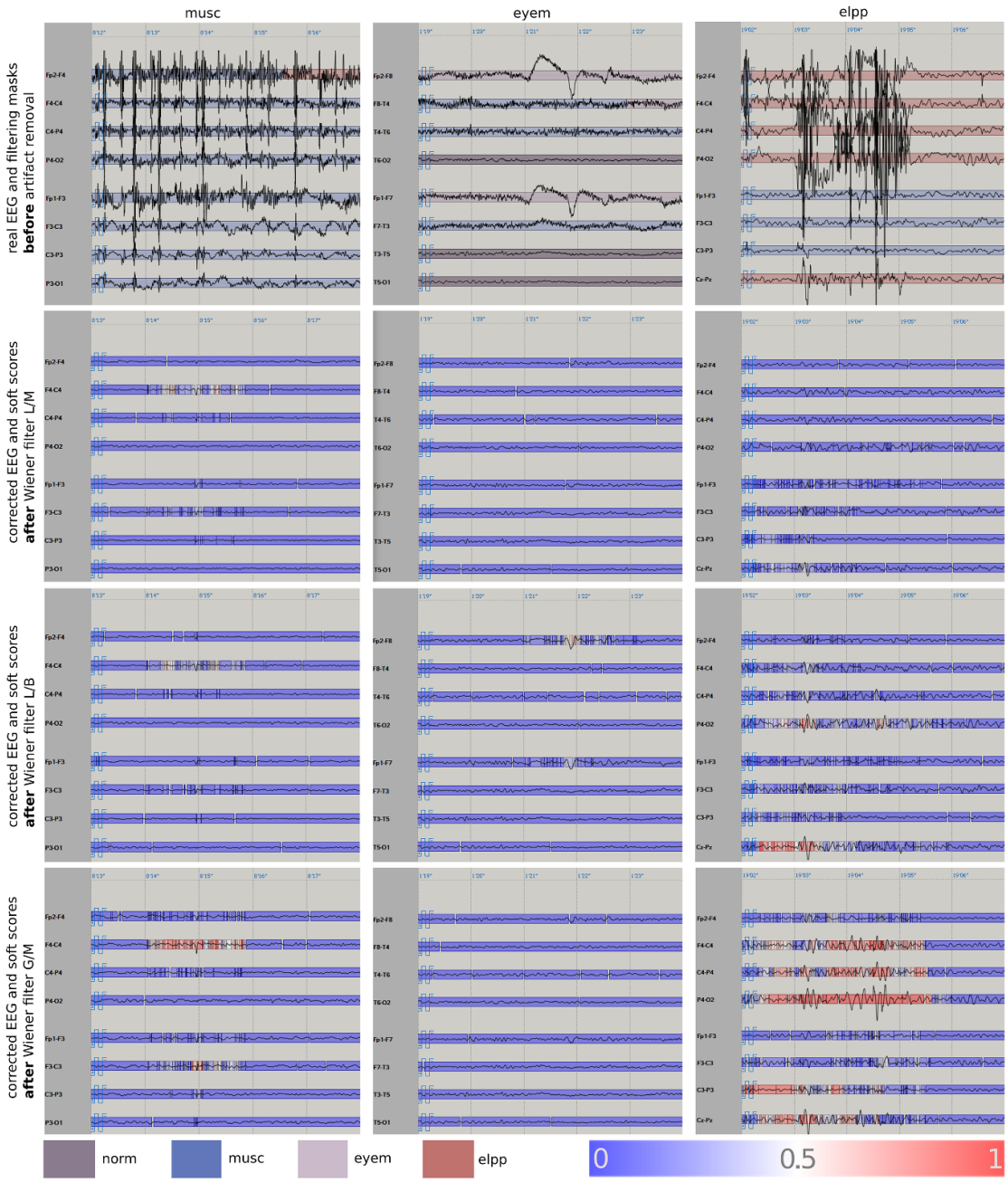


Figure 9. Qualitative comparison of Wiener filters on real EEG, contaminated with three artifact classes (columns). Row 1: Real, unprocessed EEG with colored *filtering masks*. Four colors encode the class labels (bottom left legend). The A/B detector correctly classifies the real EEG epochs. It masks real EEG data with artifact classes `musc`, `eyem`, `elpp`, and `norm` class to initialize the training of the filters automatically. Rows 2, 3, 4: Processed EEG after applying three different Wiener filtering configurations, L/M, L/B, G/M. The signals are overlaid with soft scores (heatmap) of the A/B detector. The detector assigns summed scores of artifact classes to each q-wave in corrected EEG. Detection scores near 0 and 1 indicate norm-like and artifact-like signals, respectively (bottom right legend). The detector indicates that the L/M filters perform somewhat better than the L/B filters while both outperform the G/M filters.

percent of kept eigenvalues rose from 20% to 100%. The L/M Wiener filters were trained on all EEG channels within 10 section long artifact mask and norm mask. We computed power spectral density (PSD) per channel by the Welsch method with default settings using the MNE-Python library [36]. We then averaged the PSDs (black) and computed their standard deviations (gray) over the selected four channels. The PSD plots cover five common waveform frequency bands: delta (δ), theta (θ), alpha (α), beta (β), and gamma (γ). Before filtering, the variance in amplitudes of signals above 20 Hz frequency was high across the channels in comparison to the variance in the band 0–20 Hz. After filtering, the PSD profiles differ mostly above the 20 Hz threshold as well. Thus, in addition to PSD plots between 0–40 Hz, we compute the mean total power, which was summed over all frequencies above 20 Hz until 125 Hz and averaged over all four channels ($\mathrm{MTP}_{f>20}$ [dB]), as a spectrum summary for comparative analysis.

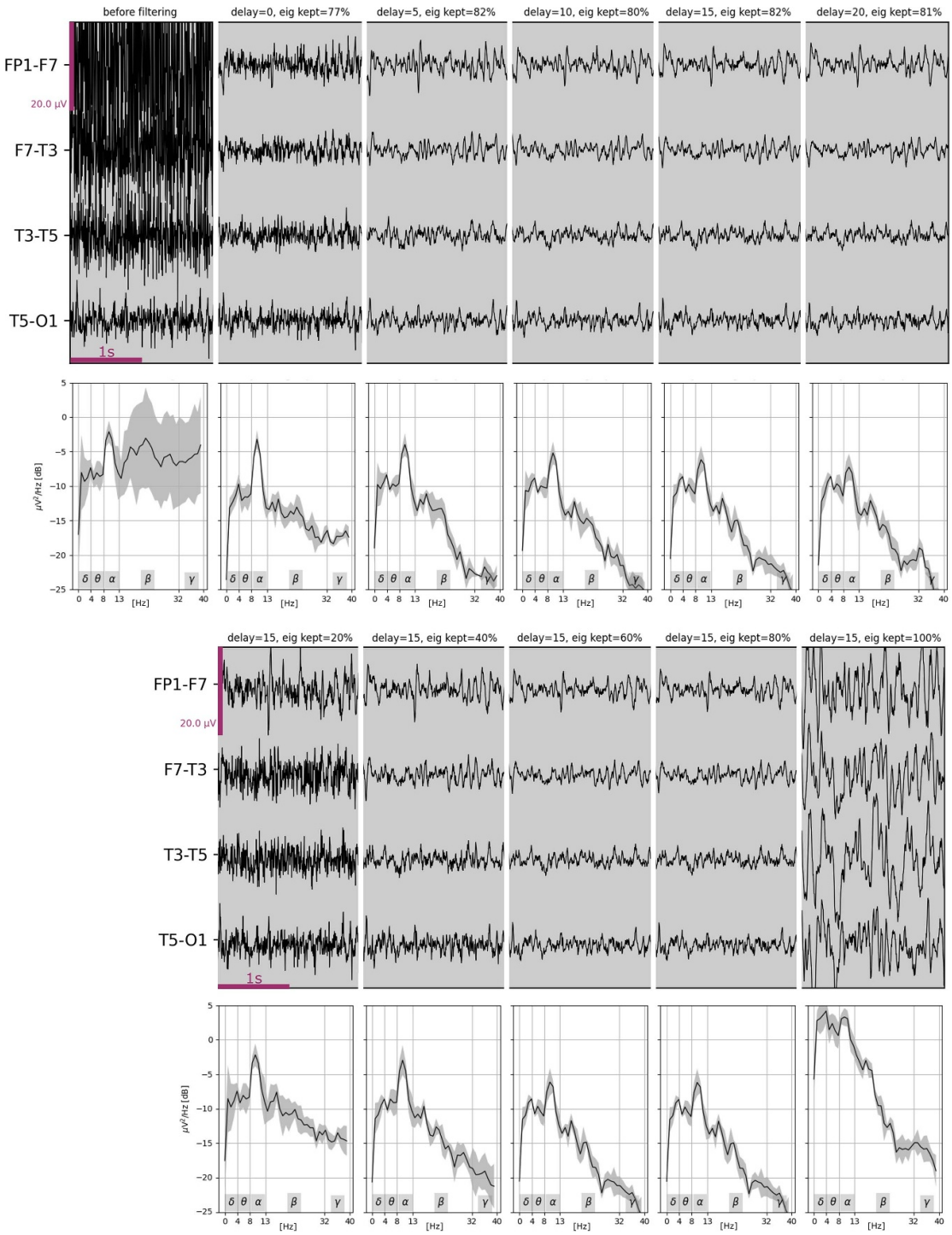


Figure 10. The correction of noisy EEG, here `musc` artifacts in four EEG channels, is sensitive to the specific hyperparameter settings of Wiener filters. The L/M Wiener filters, trained on the same norm and artifact data but with different delays (top panel) and artifact covariance ranks (bottom panel), correct the EEG differently. The EEGs are accompanied by respective Power Spectral Density plots (averaged over four channels, with mean in black and standard deviation in gray) over five common waveform frequency bands: delta (δ), theta (θ), alpha (α), beta (β), and gamma (γ). Alpha frequencies can be detected across most filter settings.

The filtered signals mostly have lower PSD profiles than before the correction. In all cases except one, the α band dominates the other bands. For the top panel, the Wiener filter with delay $= 0$ produces PSD plots where the α band power is the highest with respect to the powers in the lower bands. On the other hand, the Wiener filters with three delays $= 5, 10, 15$ may have a noticeable advantage over the filter with delay $= 0$ in the delta (δ) and theta (θ) band where they recover higher signal power, with PSD profiles following the original profile closer. These delays also lead to lower amplitudes after 20 Hz than delay $= 0$, with respective $\mathrm{MTP}_{f>20} = -3.4$, $\mathrm{MTP}_{f>20} = -4.3$, $\mathrm{MTP}_{f>20} = -4.6$ as compared to $\mathrm{MTP}_{f>20} = 2.3$ at

delay = 0. This suggests non-zero delays yield better performance in artifact attenuation, which is also clearly manifested in the corresponding temporal domains. In addition, the MTPs for delays = 10, 15 are lower than for delay = 5. Then, filter delay = 20 yields the lowest $\mathrm{MTP}_{f>20} = -4.7$ and has a PSD profile in the δ and θ band comparable to the profiles for the three shorter delays. However, the peak power in the α band tends to vanish over neighbor peak powers in the two lower bands at delay = 20 than in the case of the other four shorter delays. For all delay settings, the percent of automatically kept eigenvalues clustered around 80%, which is in concert with the findings in section 3.4.3. For the bottom panel, setting a filter with 20% eigenvalues kept produced PSDs of all four EEG channels closely following their original PSD profiles within the three lowest bands, as confirmed by the PSD mean and standard deviation. However, much power remained after 20 Hz, with $\mathrm{MTP}_{f>20} = 6.0$. Comparison of the $\mathrm{MTP}_{f>20} = -1.0$, $\mathrm{MTP}_{f>20} = -4.5$, and $\mathrm{MTP}_{f>20} = -4.6$ for the 40%, 60%, and 80% of kept eigenvalues, respectively, indicates that the 40% setting leads to weaker EEG correction, which also can be seen in the temporal domain. Keeping 100% of eigenvalues leads to filters with the PSD profile even higher than the original one. The amplitudes of temporal waveforms reveal improper behavior of the filtering under this setting.

3.6. Synthetic experiments

The optimal filtering configurations, which achieve the best performance under the AED measure, should also belong to the top-performing configurations under measures that rely on ground truth. Realistically though, the configurations should at least not be ranked below the average performers. Consequently, we also validate our AED-based detection protocol on synthetic data. Moreover, we note that the AED can be computed for any filtering method (see figure 2) to compare multiple different algorithms. The true, unknown, clean signal does not depend on a particular filtering procedure. Each filtering method produces its own version of this unknown clean signal. As long as all these filtered signals are scored with the same detector, AED will indicate which version better matches the classifier's model of norm and artifacts. Hence, our protocol can compare multiple configurations and parameterizations of the same algorithm, as shown in previous sections, but it can also compare different algorithms. To this end, this section compares two variants of L/B Wiener filtering by using different delays (`Wiener delays`) and ranks (`Wiener ranks`) with two ICA algorithms, Fast-ICA (`ICAfastica`) [37] and Picard-ICA (`ICApicard`) [38] on a synthetic EEG dataset. We used the MNE-Python library for testing both ICA algorithms.

3.6.1. Dataset and metrics

We use the EEGdenoiseNet dataset [39] for our synthetic data experiments. The dataset consists of single-channel segments that are 2-sec long and sampled at 256 Hz. The dataset is split into 4514 segments of ICA-cleaned EEG, 3400 segments of EOG artifacts, and 5598 EMG artifacts. We first normalize these fragments to zero-mean and unit variance signals. The training, validation, and test sets consist of 180, 20, and 500 epochs, respectively. The training and validation sets are used to train and validate a binary classifier (section 2.1). Each 6-sec long epoch has 18 channels and consists of 3 shorter epochs, each 2-sec long. The left and right shorter epoch is `norm` after randomly selecting 32 single-channel signals from the cleaned and normalized EEG set. The shorter central epoch contains 1 to 9 noisy single-channel fragments. We contaminate these fragments with artifacts of SNR noise from −7 to 2 dB following [39]. Namely, we randomly select pure artifacts from EOG and EMG single-channel sets and linearly mix them with the norm fragments. The resulting noisy fragments are randomly located across channels in the central, shorter epoch.

The three evaluation metrics in [39] are relative root mean squared error in the temporal and spectral domain (RRMSE temporal and RRMSE spectral) and the correlation coefficient (CC). Low values of MSE-related metrics and high values of the CC metric mean better performance. Together with AED, the four metrics were computed on all noisy channels in the central epoch to assess the algorithms' performance on the artifact attenuation task.

3.6.2. Quantitative comparison of artifact removal algorithms

The computation of all three metrics from section 3.6.1 requires access to the ground truth signals at test time, which are readily available in synthetic data. In contrast, our AED metric has no access to ground truth at test time and relies solely on a classification model that is trained on training data. When searching for the best configuration of a given filtering algorithm, the optima of these metrics individually inform about optimal selections of hyperparameters. Knowing the ground-truth data, ideally, the metrics should jointly point to the same single configuration of a filtering algorithm to be deployed in a given application. Moreover, the metrics should create the same rankings of different algorithms to specify a consistently best-performing algorithm. Hence, we investigate the degree of agreement of the metric optima with the optima of our AED measure for tuning the best configuration of a given algorithm and for ranking different algorithms.

Both ICA-based filtering algorithms were configured by searching for the optimal percentage of explained cumulative variance during ICA fitting,

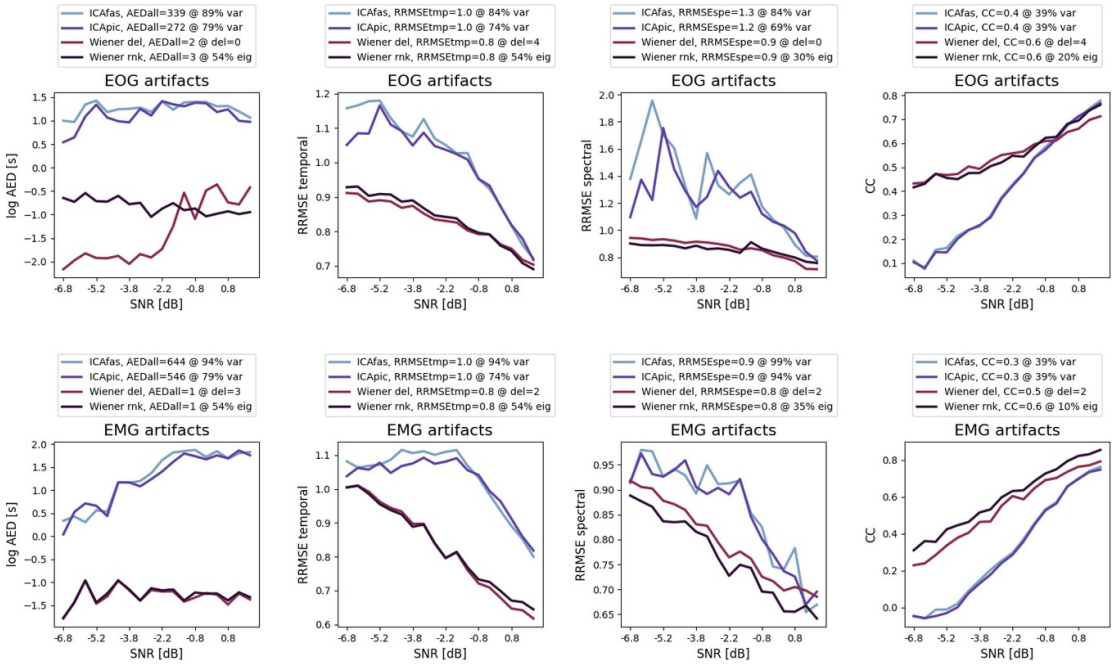


Figure 11. A quantitative comparison of ICA-based and Wiener-based filtering methods on suppression of EOG and EMG artifacts from synthetic data. As indicated in the legends, the plots illustrate the performance of each method at (@) its best configuration per metric (columns) and per artifact type (rows).

which is their main tuning parameter. The independent components that best correlated with the EOG and EMG artifacts were rejected before signal reconstruction. The Wiener-based filtering was configured by searching for optimal delay under adaptive covariance rank selection (GEVD), followed by searching for optimal covariance rank given the optimal delay, as in sections 3.4.2 and 3.4.3. The masks of norm and artifacts were indicated explicitly by the construction of epochs, where the left and the right side were the norm, and the central part was contaminated by numerous artifacts.

The quantitative comparison of ICA-based and Wiener-based filtering algorithms is shown in figure 11. The panels depict their behavior on four metrics across multiple EEG contamination levels by EOG and EMG artifacts. The plots refer to each method's best configuration per metric and artifact type, as indicated in the legend at the top of each panel. Both Wiener filtering algorithms outperform ICA-based filtering algorithms on both artifact types, according to the MSE-based and CC measures. Our AED (first column) reflects these rankings clearly. Furthermore, the MSE-based and CC measures correlate positively with the decreasing SNR for all filtering methods, suggesting that signal recovery is easier with lower SNR. On the other hand, the AED measure remains relatively insensitive to varying noise levels. We posit this effect may be attributed to discriminative training with hard labels (section 2.1.2) that force the model to predict all artifacts equally, regardless of their SNR levels.

The degree of agreement between metrics on optimal filtering configuration is shown in figure 12. For all four filtering methods, both MSE-based measures point to the best filtering configurations (x-axes) that are close to the best filtering configurations that are indicated by the AED measure, suggesting good agreement between these three measures. The CC measure hints at the configurations of three out of four filtering methods that are much different from the ones pointed to by the other three measures. This may raise questions about whether the CC measure is a reliable indicator of filtering performance. On the other hand, all measures agree that shorter filter delays (ranging from 2 to 4 samples) are optimal for Wiener filtering on our synthetic data.

3.6.3. Deteriorating detection by label noising

As our AED-based protocol depends on event detection, the detector should be sensitive to EEG artifacts. We leverage the synthetic data and the performance of the delayed Wiener filter with the adaptive rank selection from section 3.6.2 to simulate (i) the influence of noisy labels on training a detector and (ii) the ability of such detectors to keep selecting optimal filtering configuration despite noisy training. We trained 20 different detectors by gradually flipping more labels in the training set. The first model was trained on original training labels from the synthetic dataset. We then started by flipping 5% of randomly selected labels until 95% of labels were flipped. The results on the validation set in the form of precision-recall curves and APs are shown in the left and middle

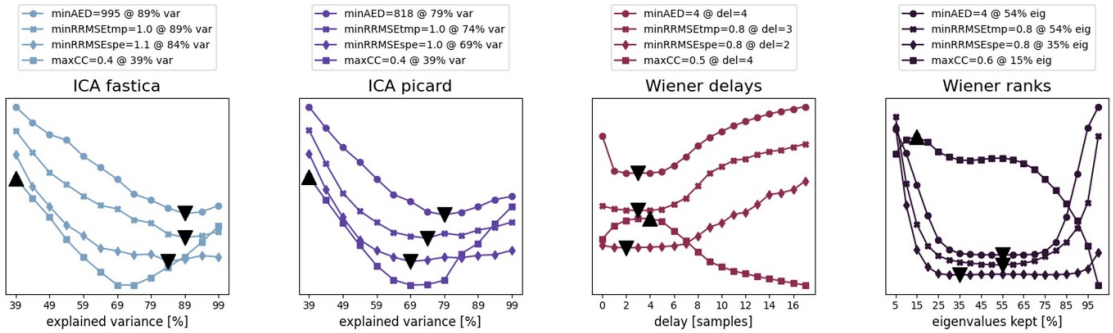


Figure 12. Illustration of agreements between four metrics in pointing to optimal (indicated by ▼ and ▲) filtering configurations jointly for EOG and EMG artifacts. The four metrics agree the most when searching for the best delay configuration for the Wiener filtering method.

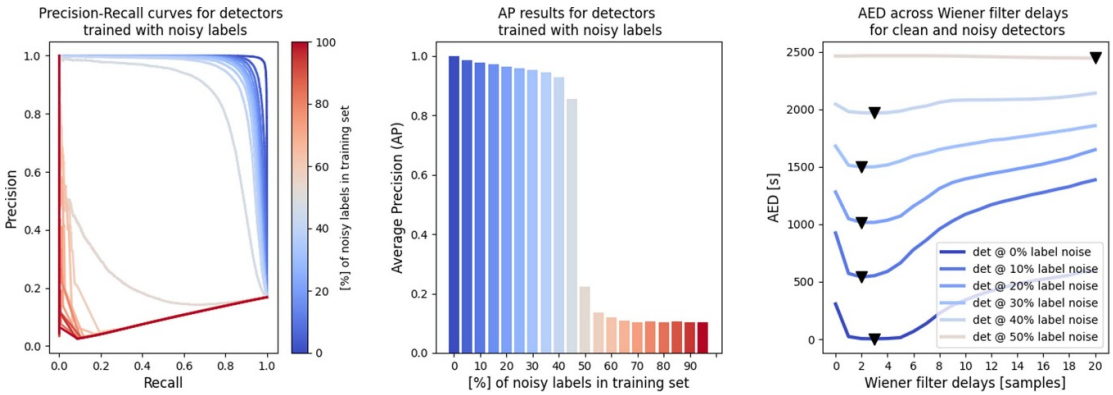


Figure 13. Assessing the usability of the detection-based evaluation protocol. Evaluation of detectors (left and middle panel) trained even with 40% of noisy labels indicates that such a protocol can still be useful to hint at optimal filtering configurations (right panel) unless the performance of detection drops significantly.

panels of figure 13, respectively. The rate of degrading APs is quite slow, eventually collapsing after 50% of corrupted labels. The right panel, in turn, shows that the AED-based protocol can still reliably hint at optimal Wiener filtering configuration of delay = 3 despite using detectors that were trained even with 40% of flipped labels. As observed in figure 12 from the previous section, the ground-truth-based CC and MSE measures hint at the same configuration as the AED measure.

4. Discussion

This study sought to answer an essential question for no-reference evaluation: *can a detector be trusted in rating EEG quality for developing an accurate removal algorithm of real artifacts?* The experimental results provide evidence for an affirmative answer. Our rating protocol goes beyond simulated EEG data and automatically evaluates the performance of many configurations of Wiener-based filtering [12] on real data with the help of a discriminatively trained detector. We intentionally keep true to the Wiener-based filtering from [12] as the testbed for experimentally showing the reliability of our rating-by-detection protocol. We expect the detector to make generalization errors during the automatic evaluation of Wiener filters under real-world conditions. However, in concert with the findings of [12] on simulated and real EEG data, our evaluation protocol similarly demonstrates on real EEG data that filter delay and artifact covariance rank impact the filter performance. The proposed rating procedure further finds that local class-specific Wiener filters remove artifacts most effectively, meeting theoretical expectations. Namely, canonical filters are not well suited for eliminating artifacts with patient-specific spectral signatures, such as artifacts of electromyogenic origin [23]. Moreover, Wiener-based filtering assumes the neurogenic and artifact signals are stationary, uncorrelated stochastic processes with known covariances [40]. As real EEG is a nonstationary signal [34], our rating-by-detection protocol correctly indicates that local modeling of artifacts with Wiener filters improves the accuracy the most.

We argue that standard event detection measures for real data, such as AP under recall-precision curves, constitute a good alternative to the standard SER measure, which quantifies norm distortion after artifact removal on simulated data. Automating an artifact removal algorithm with a detector is standard practice. A hypothetically perfect detector

will indicate only these EEG fragments for artifact removal that contain artifacts. It would thus leave the norm signal unaffected to lower the computational costs of signal filtering and eliminate the risk of norm distortion. Therefore, it may be more imperative to develop highly effective artifact detectors where the AP measure plays a key role instead of optimizing the SER measure during artifact removal development. For example, our detector leaves room for improvement. It is a simple classifier that parameterizes only temporal context in an EEG channel and neglects waveform correlations of artifacts across multiple EEG channels. Improving the multi-classification performance between artifacts would further increase the reliability of a rating-by-detection protocol. On the other hand, our A/B and B/B detectors, which are trained on different datasets, have an apparent agreement in discerning artifact classes from the background activity and point to similarly optimal filters.

The foundation of this study was a quantitative evaluation protocol based on detecting abnormal EEG waveforms after filtering real artifacts. The development of artifact removal methods is inhibited by unavailable ground truth. It thus should be noted that some care should be taken when analyzing the results. The rating by detection protocol with average duration measure provides no formal guarantees for being correct. The AED measure computes the duration intervals of artifact-like q-waves, weighting them by the complete range of detection thresholds. In effect, the EEG quality is measured independently of any specific threshold. Computing the duration of q-waves instead of counting EEG peaks in cleaned EEG is necessary. Different filtering types will produce signals containing a different number of peaks. Computing the duration of q-waves instead of counting the peaks ensures that rankings of filtering types are more meaningful and fair. The detection part of the protocol provides the AED measure with soft scores for the q-waves. A natural extension of this work could address the uncertainty of the detection scores, but the protocol has several critical dependencies. Firstly, it requires data and manual labels to train a classifier to recognize events at test time. Labeling data is time-consuming, and the labeled data may be of insufficient quality to train a classifier with good generalization. Secondly, manual labels can be noisy because of insufficient knowledge and precision, fatigue, and affect during labeling the data. Thirdly, hand-crafted features insufficiently model the morphology of the events limiting classifiers in separating the feature space with better margins during training. Then, the standard machinery of gradient-boosted decision trees yields detection scores without uncertainty estimates [41]. Additional evaluation measures that quantify uncertainty for instance by ensembling classifiers [42, 43], might affect and improve the filter selection process when a lower AED measure would be considerably more uncertain than a higher AED measure of two different filtering types.

5. Conclusions

This study developed and extensively validated a novel rating-by-detection protocol for measuring the EEG quality after removing real artifacts. The validation of the protocol exploited a state-of-the-art Wiener-based artifact filtering method as a testbed to gain confidence in the automatic evaluation of EEG quality. The results were summarized with the proposed AED measure that jointly measures artifact attenuation and norm distortion in corrected EEG. The AED rating-by-detection protocol should be valuable for EEG practitioners and developers who develop artifact removal algorithms. Our validation shows that reliable comparisons between many artifact removal configurations are possible despite the missing ground-truth neural signals in corrected EEG.

The key aspect for the success of our study is the consistent ranking of filter configurations by two detectors that are trained on different subsets of data. A consistent ranking indicates that some filters consistently work better than others. The evaluation method can show evident underperformers. The simplest but fastest Wiener filter performed the worst. It also yields theoretically expected results that agree with intuition, hinting at the advantages and disadvantages of specific designs of Wiener-based artifact filtering such as (i) local *vs.* global filter training, (ii) class-specific *vs.* class-agnostic filter training, and (iii) fine-tuning filter delays and artifact covariance ranks.

An acceptable standard for artifact removal evaluation on real EEG data is missing, but it is necessary to go beyond simulating EEG artifacts and advance the deployment of state-of-the-art EEG processing algorithms in real-world conditions. We will evaluate other artifact removal algorithms in a broader context of EEG analysis in the future. Our future work will address the problem of quantitative evaluation of artifact removal algorithms in the presence of more classes of events in scalp EEG, including interictal events for epilepsy diagnosis and event-related potentials for neurofeedback in children and adults.

Data availability statement

The data cannot be made publicly available upon publication because the cost of preparing, depositing and hosting the data would be prohibitive within the terms of this research project. The data that support the findings of this study are available upon reasonable request from the authors. https://isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml.

Acknowledgment

This work was financed in part by the National Centre for Research and Development under the Agreement STRATEGMED3/306306/4/NCBR/2017.

ORCID iD

Daniel Węsierski https://orcid.org/0000-0001-7093-8764

References

[1] Boyd K, Eng K H and David Page C 2013 Area under the precision-recall curve: point estimates and confidence intervals *Joint European Conf. on Machine Learning and Knowledge Discovery in Databases* (Springer) pp 451–66
[2] Shahid M, Rossholm A, Lövström B and Zepernick H-Jurgen 2014 No-reference image and video quality assessment: a classification and review of recent approaches *EURASIP J. Image Video Process.* **2014** 40
[3] Raykar V C, Yu S, Zhao L H, Valadez G H, Florin C, Bogoni L and Moy L 2010 Learning from crowds *J. Mach. Learn. Res.* **11** 4
[4] Hendrycks D, Mazeika M, Wilson D and Gimpel K 2018 Using trusted data to train deep networks on labels corrupted by severe noise *Advances in Neural Inf. Proc. Systems* vol 31, ed S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi and R Garnett (Curran Associates, Inc.)
[5] Vích R, Nouza J and Vondra M 2008 Automatic speech recognition used for intelligibility assessment of text-to-speech systems *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction* (Springer) pp 136–48
[6] Kalal Z, Mikolajczyk K and Matas J 2011 Tracking-learning-detection *IEEE Trans. Pattern Anal. Mach. Intell.* **34** 1409–22
[7] Bouix S, Martin-Fernandez M, Ungar L, Nakamura M, Koo M-S, McCarley R W and Shenton M E 2007 On evaluating brain tissue classifiers without a ground truth *Neuroimage* **36** 1207–24
[8] Valindria V V, Lavdas I, Bai W, Kamnitsas K, Aboagye E O, Rockall A G, Rueckert D and Glocker B 2017 Reverse classification accuracy: predicting segmentation performance in the absence of ground truth *IEEE Trans. Med. Imaging* **36** 1597–606
[9] Fan W and Davidson I 2006 Reverse testing: an efficient framework to select amongst classifiers under sample selection bias *Proc. 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 147–56
[10] Sweeney K T, Ayaz H, Ward T E, Izzetoglu M, McLoone S F and Onaral B 2012 A methodology for validating artifact removal techniques for physiological signals *IEEE Trans. Inf. Technol. Biomed.* **16** 918–26
[11] Urigüen J A and Garcia-Zapirain B 2015 EEG artifact removal—state-of-the-art and guidelines *J. Neural Eng.* **12** 031001
[12] Somers B, Francart T and Bertrand A 2018 A generic EEG artifact removal algorithm based on the multi-channel Wiener filter *J. Neural Eng.* **15** 036007
[13] Croft R J, Chandler J S, Barry R J, Cooper N R and Clarke A R 2005 EOG correction: a comparison of four methods *Psychophysiology* **42** 16–24
[14] Kumar P S, Arumuganathan R, Sivakumar K and Vimal C 2008 Removal of ocular artifacts in the EEG through wavelet transform without using an EOG reference channel *Int. J. Open Probl. Computer Sci. Math.* **1** 188–200
[15] Daly I, Pichiorri F, Faller J, Kaiser V, Kreilinger A, Scherer R and Müller-Putz G 2012 What does clean EEG look like? *2012 Annual Int. Conf. IEEE Engineering in Medicine and Biology Society* (IEEE) pp 3963–6
[16] Daly I, Nicolaou N, Nasuto S J and Warwick K 2013 Automated artifact removal from the electroencephalogram: a comparative study *Clin. EEG Neurosci.* **44** 291–306
[17] Hartmann M M, Schindler K, Gebbink T A, Gritsch G and Kluge T 2014 Pure EEG: automatic EEG artifact removal for epilepsy monitoring *Clin. Neurophysiol.* **44** 479–90
[18] Crespo-Garcia M, Atienza M and Cantero J L 2008 Muscle artifact removal from human sleep EEG by using independent component analysis *Ann. Biomed. Eng.* **36** 467–75
[19] Vos D M, Riès S, Vanderperren K, Vanrumste B, Alario F-X, Huffel V S and Burle B 2010 Removal of muscle artifacts from EEG recordings of spoken language production *Neuroinformatics* **8** 135–50
[20] Kobler R J, Sburlea A I, Lopes-Dias C, Schwarz A, Hirata M and Müller-Putz G R 2020 Corneo-retinal-dipole and eyelid-related eye artifacts can be corrected offline and online in electroencephalographic and magnetoencephalographic signals *NeuroImage* **218** 117000
[21] Pham T T H, Croft R J, Cadusch P J and Barry R J 2011 A test of four EOG correction methods using an improved validation technique *Int. J. Psychophysiol.* **79** 203–10
[22] McMenamin B W, Shackman A J, Maxwell J S, Greischar L L and Davidson R J 2009 Validation of regression-based myogenic correction techniques for scalp and source-localized EEG *Psychophysiology* **46** 578–92
[23] Shackman A J, McMenamin B W, Slagter H A, Maxwell J S, Greischar L L and Davidson R J 2009 Electromyogenic artifacts and electroencephalographic inferences *Brain Topogr.* **22** 7–12
[24] McMenamin B W, Shackman A J, Maxwell J S, Bachhuber D R W, Koppenhaver A M, Greischar L L and Davidson R J 2010 Validation of ICA-based myogenic artifact correction for scalp and source-localized EEG *Neuroimage* **49** 2416–32
[25] Mognon A, Jovicich J, Bruzzone L and Buiatti M 2011 ADJUST: an automatic EEG artifact detector based on the joint use of spatial and temporal features *Psychophysiology* **48** 229–40
[26] Scherer R, Moitzi G, Daly I and Müller-Putz G R 2013 On the use of games for noninvasive EEG-based functional brain mapping *IEEE Trans. Comput. Intell. AI Games* **5** 155–63
[27] Daly I, Scherer R, Billinger M and Müller-Putz G 2014 FORCe: fully online and automated artifact removal for brain-computer interfacing *IEEE Trans. Neural Syst. Rehabil. Eng.* **23** 725–36
[28] Jing J *et al* 2020 Development of expert-level automated detection of epileptiform discharges during electroencephalogram interpretation *JAMA Neurol.* **77** 103–8
[29] Constantino A C, Sisterson N D, Zaher N, Urban A, Richardson R M and Kokkinos V 2021 Expert-level intracranial electroencephalogram ictal pattern detection by a deep learning neural network *Front. Neurol.* **12** 673
[30] Abou Jaoude M, Sun H, Pellerin K R, Pavlova M, Sarkis R A, Cash S S, Westover M B and Lam A D 2020 Expert-level automated sleep staging of long-term scalp electroencephalography recordings using deep learning *Sleep* 11 **43**
[31] Chen T and Guestrin C 2016 Xgboost: a scalable tree boosting system *Proc. 22nd ACM Sigkdd Int. Conf. on Knowledge Discovery and Data Mining* pp 785–94
[32] Harati A, Lopez S, Obeid I, Picone J, Jacobson M P and Tobochnik S 2014 The TUH EEG CORPUS: a big data resource for automated EEG interpretation *2014 IEEE Signal Processing in Medicine and Biology Symp. (SPMB)* (IEEE) pp 1–5
[33] Halford J J *et al* 2015 Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings *Clin. Neurophysiol.* **126** 1661–9
[34] Kaplan A Y, Fingelkurts A A, Fingelkurts A A, Borisov S V and Darkhovsky B S 2005 Nonstationary nature of the brain

activity as revealed by EEG/MEG: methodological practical and conceptual challenges *Signal Process.* **85** 2190–212
[35] Serizel R, Moonen M, Van Dijk B and Wouters J 2014 Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants *IEEE/ACM Trans. Audio Speech Lang. Process.* **22** 785–99
[36] Gramfort A *et al* 2013 MEG and EEG data analysis with MNE-Python *Front. Neurosci.* **267** 267
[37] Hyvarinen A 1999 Fast and robust fixed-point algorithms for independent component analysis *IEEE Trans. Neural Netw.* **10** 626–34
[38] Ablin P, Cardoso J-F and Gramfort A 2018 Faster independent component analysis by preconditioning with Hessian approximations *IEEE Trans. Signal Process.* **66** 4040–9
[39] Zhang H, Zhao M, Wei C, Mantini D, Zherui Li and Liu Q 2021 EEGdenoiseNet: a benchmark dataset for deep learning solutions of EEG denoising *J. Neural Eng.* **18** 5
[40] Sweeney K T, Ward T E and McLoone S F 2012 Artifact removal in physiological signals–practices and possibilities *IEEE Trans. Inf. Technol. Biomed.* **16** 488–500
[41] Malinin A, Prokhorenkova L and Ustimenko A 2021 Uncertainty in gradient boosting via ensembles *Int. Conf. on Learning Representations (ICLR)*
[42] Ashukha A, Lyzhov A, Molchanov D, and Vetrov D 2020 Pitfalls of in-domain uncertainty estimation and ensembling in deep learning (arXiv:2002.06470)
[43] Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon J, Lakshminarayanan B and Snoek J 2019 Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift *Advances in Neural Inf. Proc. Systems* vol 32