

PAPER • OPEN ACCESS

A gated artifact management pipeline for low-density eyewear EEG

To cite this article: Andrea Costanzo Palmisciano *et al* 2026 *Physiol. Meas.* **47** 095010

View the [article online](#) for updates and enhancements.

You may also like

- [Report from the IRPA task group on awareness of tissue reactions in the eye lens, cardiovascular system and skin](#)
Marie Claire Cantone, Colin J Martin, Mercè Ginjaume et al.
- [Turning glasses into sensors: capacitive facial expression recognition under frame geometry constraints](#)
Francesco Latino, Matteo Rossi, Andrea Costanzo Palmisciano et al.
- [Attenuation assessment of medical protective eyewear: the AVEN experience](#)
M Bertolini, G Benecchi, M Amici et al.



PAPER

OPEN ACCESS

RECEIVED
17 May 2026REVISED
10 August 2026ACCEPTED FOR PUBLICATION
13 August 2026PUBLISHED
29 September 2026

Original content from
this work may be used
under the terms of the
[Creative Commons
Attribution 4.0 licence](#).

Any further distribution
of this work must
maintain attribution to
the author(s) and the title
of the work, journal
citation and DOI.



A gated artifact management pipeline for low-density eyewear EEG

Andrea Costanzo Palmisciano^{1,*} , Andrea Farabbi¹ , Francesco Latino¹ , Matteo Rossi² ,
Niccolò Antonello² , Pietro Cerveri^{3,1} and Luca Mainardi¹

¹ Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy

² Smart Eyewear Lab, EssilorLuxottica, Milano, Italy

³ Department of Industrial and Information Engineering, Università di Pavia, Pavia, Italy

* Author to whom any correspondence should be addressed.

E-mail: andreacostanzo.palmisciano@polimi.it

Keywords: EEG artifacts, eyewear EEG, artifact correction

Abstract

Objective. To develop an artifact-management method suitable for low-density, dry-electrode electroencephalogram (EEG) recorded from a wearable eyewear platform, enabling artifact detection, classification, and correction while avoiding unnecessary processing of clean neural activity. **Approach.** A gated two-stage pipeline is proposed in which a lightweight gate classifier first identifies artifact-contaminated EEG windows and selectively triggers further processing only when needed. The complete pipeline combines: (i) a time-domain feature-based detector, (ii) a convolutional–recurrent network for three-class artifact discrimination (eye blinks, horizontal eye movements, facial movements), and (iii) a U-Net-based denoising autoencoder. **Main results.** Evaluated on the eyewear EEG dataset (29 subjects) using a leave-one-subject-out scheme, the gate classifier achieved a balanced accuracy of 0.89. The artifact classifier reached a balanced accuracy of 0.84 with a median macro-F1 of 0.83. The denoiser achieved a median normalized RMSE lower than 0.20 and a spectral cosine similarity above 0.84. In the end-to-end evaluation, the pipeline achieved a median normalized RMSE of 0.19 and a spectral cosine similarity of 0.89. Evaluation on an independent external benchmark showed that the proposed pipeline maintained a slightly lower gate detection performance (balanced accuracy 0.82) and achieved higher two-class classification performance. The denoising performance was comparable to that obtained on the dataset collected for this study, with absolute median differences not exceeding 0.04 for both normalized RMSE and spectral cosine similarity. **Significance.** These results support the effectiveness of the proposed approach for artifact management in wearable eyewear EEG, while maintaining a design compatible with resource-constrained wearable implementations.

1. Introduction

Electroencephalogram (EEG) recordings are frequently contaminated by artifacts, particularly those arising from eye blinks and muscle movements, which can be orders of magnitude larger than the underlying neural signals (Teplan 2002). EEG artifacts can be classified into two categories: physiological artifacts and non-physiological artifacts. Physiological artifacts originate from body parts near the head, such as eyes, muscles, and heart, while non-physiological artifacts stem from environmental and equipment factors (Croft and Barry 2000).

Traditional artifact removal approaches encompass several key methods. independent component analysis (ICA) has been widely adopted, though it often requires expert intervention for component selection (Delorme and Makeig 2004). Wavelet Transform and Empirical Mode Decomposition have shown promise but may struggle with the dynamic nature of physiological artifacts (Islam *et al* 2016). Signal space projection projects the EEG onto a subspace orthogonal to the noise direction to remove artifacts from the original EEG (Uusitalo and Ilmoniemi 1997). It is suitable for stereotypical artifacts

like blinks and heartbeats. A recent method in the family of regression-based denoising is artifact-reference multivariate backward regression (ARMBR), which estimates the EOG reference and spatial filters from the EEG itself without the need for dedicated channels (Alkhoury *et al* 2025).

Deep learning provides a modern framework for EEG artifact removal. Various architectures have been explored, from convolutional neural networks to more specialized encoder–decoder and embedding-decomposition approaches (Sun *et al* 2020, Yu *et al* 2022). One-dimensional residual convolutional neural networks have shown effectiveness for denoising raw time-domain waveforms, although earlier convolutional architectures generally achieved more reliable results for ocular than for myogenic artifacts (Zhang *et al* 2021). Benchmark studies have also evaluated multiple architectures, including fully-connected networks, simple and complex CNNs, and recurrent neural networks, providing comprehensive comparisons for both ocular and myogenic artifact removal (Zhang *et al* 2021). The classification of EEG artifacts represents another crucial challenge in EEG processing. Traditional approaches have relied on manual expert classification or semi-automated methods using predefined features (Pion-Tonachini *et al* 2019). Recent deep learning approaches have attempted to automate this process, with CNNs showing promise in distinguishing between different artifact types (Peh *et al* 2022). The growing field of wearable EEG devices equipped with a limited number of dry electrodes poses an additional challenge for existing artifact correction and classification algorithms. Compared with conventional EEG systems, wearable EEG devices are characterized by limited spatial resolution and the use of dry electrodes, which leads to suboptimal electrode-skin contact, higher impedance and signals more prone to artifacts (Arpaia *et al* 2025). Several studies have addressed this challenge by developing new methods or adapting existing ones to wearable characteristics. These approaches often rely on public datasets recorded with gold-standard equipment and wet electrodes in clinical or laboratory settings (Zhang *et al* 2021). Xing and Casson developed a deep autoencoder (DAE) for artifact correction, designed for real-time artifact removal in single-channel EEG recordings (Xing and Casso 2024). Training was performed using public datasets containing high-density, annotated EEG recordings with EOG, motion, and EMG artifacts. The single-channel design and training on signals from different channels enabled an architecture independent of specific montages or spatial information, making it adaptable to different wearable use cases. Its suitability for portable, real-time processing was evaluated through deployment on a smartphone and embedded edge-computing platforms (Saleh *et al* 2025). Artifact correction was also addressed using a combined deep learning and adaptive filtering approach, consisting of a CNN cascaded with LMS filtering (Nair *et al* 2025). This design aimed to improve robustness when artifacts differ from the training data. The method was developed on a two-channel polysomnography EEG dataset with annotated EOG and chin EMG artifacts, with particular attention to optimizing the algorithms for efficient FPGA implementation. The problem of artifact detection in few-channel EEG was addressed by Ingolfsson *et al* in the context of real-life epilepsy monitoring (2024). They used four temporal channels from an annotated clinical-grade EEG dataset. Artifact detection relied on FFT and discrete wavelet transform features, with a gradient-boosted decision tree classifying each window and channel as artifact-affected or artifact-free. Using the dataset annotations, the same approach was extended to 12-class artifact classification. The algorithm was also optimized for deployment on a parallel ultra-low-power platform. A semi-supervised artifact detection approach combining an autoencoder and a CNN classifier has been proposed, using a limited number of wet electrodes and an autoencoder to learn spatial features from multichannel EEG inputs (Hermans *et al* 2023).

As noted above, a key difference between wearable and clinical-grade EEG is the use of dry rather than wet electrodes. Arpaia *et al* proposed a hybrid artifact removal framework combining Multivariate Empirical Mode Decomposition and Artifact Subspace Reconstruction for a four-channel dry-electrode EEG setup (2023, 2024). Artifact detection was addressed in Grosselin *et al* (2019) using a two-parietal dry-electrode system, where a machine learning classifier used time-, frequency-domain, and entropy features from 1-second EEG windows to assign each window to one of three noise-level classes.

A specific domain of wearable EEG devices is represented by EEG-enabled eyewear designed for health monitoring and consumer-level applications, such as the AttentivU (Kosmyna *et al* 2019) and the e-Glass platform (Zanetti *et al* 2025). These platforms adopt dry electrodes on the temples for EEG recording. Zanetti *et al* introduced e-Glass for cognitive workload monitoring and seizure detection in non-clinical environments. In Zanetti *et al* (2022), they identified artifact susceptibility as the main obstacle to wearable EEG adoption and included an artifact mitigation pipeline in their eyewear-based cognitive workload monitoring workflow. The method targeted eye-blink correction through two steps: artifact detection via relative energy peaks on a frontal channel (F7), followed by multi-channel correction by subtracting an artifact waveform estimated with Savitzky–Golay filtering.

Despite these advances, state-of-the-art methods for EEG artifact detection, correction, and categorization in dry-electrode, low-density, montage-constrained devices, such as smart eyewear, remain limited. Many approaches operate continuously, causing unnecessary EEG processing. Moreover, models are often trained on standard high-density recordings with wet electrodes and conventional references, thereby overlooking the specific characteristics of eyewear-based EEG systems. While single-channel approaches proposed in previous works are not dependent on specific electrode montages, they disregard the available, albeit limited, spatial information. Finally, the inter-subject variability specifically introduced by eyewear setups (Palmisciano *et al* 2026) was not adequately addressed in prior studies. To address these limitations, we introduce a selective artifact management paradigm tailored to wearable EEG constraints, in which artifact detection acts as a gating mechanism to restrict computationally intensive processing to contaminated segments, thereby preserving clean neural activity while reducing unnecessary processing overhead.

1.1. Work contribution

We propose a novel two-stage pipeline composed of modular and independently trained blocks designed to: reduce computational demands, leverage montage-specific information, and incorporate validation strategies that explicitly account for inter-subject variability. The first stage consists of a lightweight gate classifier that uses simple time-domain features to classify each EEG window as artifact-affected or artifact-free. The second stage comprises two independent neural networks designed for distinct purposes: artifact correction (denoising) and artifact classification. The denoising network employs a UNet architecture adapted specifically for EEG data, capable of processing individual time windows independently to enable short window-based operation. The artifact classification network employs a classifier that operates directly on the multichannel signal windows to identify the type of artifact that has occurred. Our aim is to provide rapid artifact classification and signal-preserving artifact correction, both critical for short-window-based processing applications where traditional methods fall short. The proposed pipeline was trained and evaluated using data acquired with a custom EEG-enabled frame device and a dedicated artifact elicitation protocol. The training and testing procedures were designed to assess the influence of subject-specific characteristics on performance. In addition to eye blinks and horizontal eye movements, the design considers facial movements that strongly affect the nose-mounted reference electrode and therefore represent an artifact specifically related to the adopted eyewear montage. The pipeline was also tested under realistic routing decisions that consider the interaction of the two stages. Overall, the pipeline was designed to reflect the signal characteristics and processing constraints of the target application in low-density, dry-electrode EEG-enabled eyewear, from the selection of the gate features and artifact classes to the short-window operation and modular architecture.

2. Materials and methods

In this section, we introduce the methods employed to build and validate a novel system for eyewear EEG artifact denoising and classification. The experimental protocol for data acquisition is described along with the architecture implementation. The collected data are used to validate our models and evaluate performance through established metrics.

2.1. Experimental protocol

A total of 29 subjects (age: 25.6 ± 2.1 years, 16 males) participated in the study. The study and data acquisition were approved by Politecnico di Milano Ethical Committee (Opinion 46/2023, 18 December, 2023). The study was conducted in accordance with the Declaration of Helsinki, and all adult participants provided written informed consent to participate in this study. They were asked to sit comfortably in front of a screen used to display text instructions for artifact elicitation. The protocol consisted of four phases with different tasks. The first three phases were focused on the on-demand generation of specific types of artifacts, eye blinking (EB), horizontal eye movements (EM), and facial movements (FM), respectively. The considered FM can be described as nasal wrinkling, a movement of the nasal and perinasal muscles that produces a visible wrinkling of the nasal bridge and elevates the nose. This specific facial movement was selected as it directly affects the contact between the reference electrode and the skin. Subjects were asked to perform voluntary eye blinks, eye movements, or facial movements when prompted by a text appearing on the screen. A representation of the three tasks adopted to produce artifacts is reported in figure 1. Each artifact was repeated 60 times with an inter-stimulus interval of 5 seconds, yielding a total of 60 EEG epochs per artifact type. After completing the artifact trials, subjects proceeded to the fourth phase, which consisted of a 4-minute-long eyes-open (EO) eyes-closed

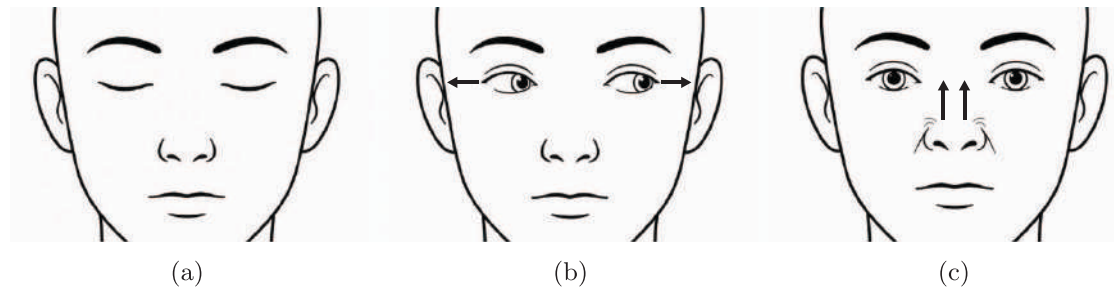


Figure 1. Examples of the task performed during the experiment to generate EEG artifacts. (a) Blink. (b) Horizontal eye movement. (c) Facial movement.

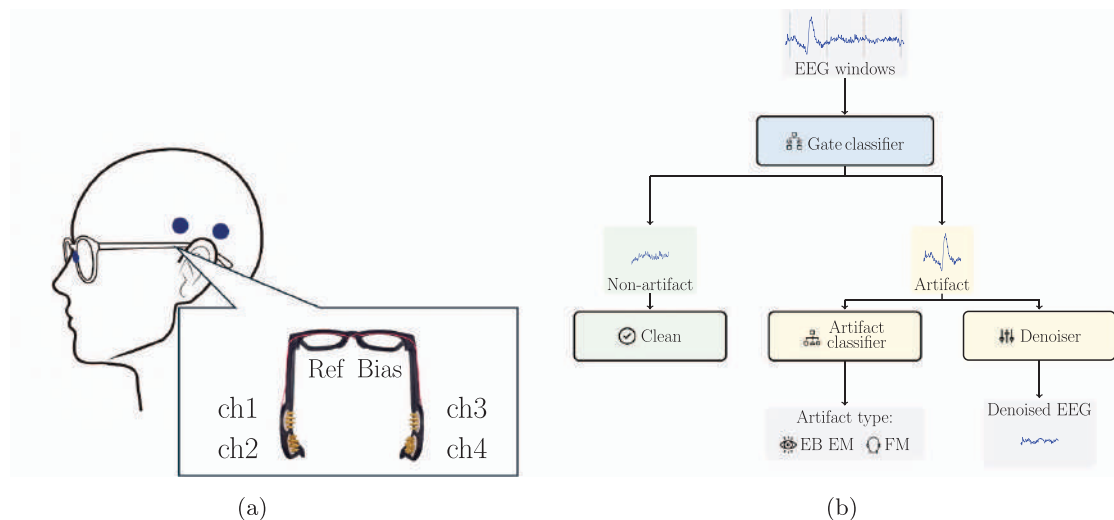


Figure 2. (a) Representation of the setup used in the experimental protocol. The blue markings indicate the position of the electrodes on the temporal area (sensing) and on the nose (reference and bias). The eyewear prototype is also reported. (b) Schematic representation of the artifact management pipeline. Each block is reported, as well as its output.

(EC) protocol, with EO and EC tasks lasting 2 minutes each. The total duration of the acquisition protocol was approximately 20 minutes, 5 minutes for each artifact elicitation phase and 4 minutes for the EOEC phase, including brief rest periods between phases to prevent fatigue.

EEG signals were recorded using a custom-developed eyewear frame (Palmisciano *et al* 2026) adapted to host 4 dry gold-plated spring-loaded electrodes (Brainbit, USA), in positions corresponding approximately to T3, T4, T5, and T6 positions of the 10–20 montage system figure 2(a). The reference and bias electrodes were made of SoftPulse conductive elastomer (Datwyler, Switzerland) and placed on the nose pads. The electrodes were connected to a Cyton OpenBCI biosensing board, recording at a sampling frequency of 250 Hz. The board was attached to the back of the participant's head with an adjustable band. EEG recordings were synchronized with the visual triggers of the artifacts using the lab streaming layer protocol (Kothe *et al* 2025).

2.2. Data preprocessing

The raw EEG signal epochs underwent specific preprocessing steps. These steps included mean subtraction and [0.5–30] Hz band-pass filtering, followed by downsampling from 250 Hz to 125 Hz. The target EEG for the artifact correction network was obtained by applying the artifact subspace reconstruction (ASR) algorithm (Cataldo *et al* 2022) to the original EEG to remove transient artifacts. The baseline statistics to calibrate the ASR algorithm were computed on the EEG recorded during the EC phase of the protocol. The signals were subsequently segmented into 2-second windows for input to the artifact management pipeline. Each window was defined starting from the marker indicating the onset of the corresponding artifact trial. The resulting windows were then inspected to identify potential signal alterations due to insufficient electrode–scalp contact.

2.3. EEG artifact management pipeline

The proposed artifact management system is designed as a three-block pipeline that operates on 2-second EEG windows and performs (i) artifact detection, (ii) artifact classification, and (iii) artifact correction (denoising). The main design principle is to apply computationally intensive correction and classification only to artifact-affected EEG windows, while minimizing the risk of distorting physiological activity in artifact-free data. Therefore, the pipeline adopts a gated architecture in which a lightweight decision module controls whether further processing is executed (figure 2(b)).

2.3.1. Gate classifier

The gate classifier (GC) operates on a set of time-domain features computed channel-wise on the incoming windows for all four available channels (Vidaurre *et al* 2009, Grosselin *et al* 2019). The following features were selected from those adopted in the literature for artifact detection (Kauppila *et al* 2018) and were chosen to have low complexity and to be suitable for fast computation:

- maximum variance-normalized sum of absolute differences, computed separately on non-overlapping 500 ms and 1000 ms sub-windows (yielding two distinct features at different time scales)
- maximum amplitude range on 500 ms non-overlapping sub-windows
- average skewness, on 500 ms non-overlapping sub-windows
- kurtosis, on entire window
- zero-crossing rate, on entire window
- Hjorth parameters (activity, mobility, and complexity), on the entire window.

Three model families were considered for the GC: logistic regression (LR), linear support vector machine classifier (SVM), and random forest (RF). All the candidate models are lightweight and allow fast inference, potentially even in the case of embedded deployment.

2.3.2. Artifact classifier

The artifact classifier was implemented as a three-class, signal-based network that operates on artifact-contaminated windows and discriminates between eye blinks, horizontal eye movements, and facial-movement (nose-wrinkle-induced reference transients) artifacts. Each input consisted of a preprocessed 2-second window, which was further normalized using z-scoring before being fed to the network. The front-end follows a VGG-inspired feature extractor (Simonyan and Zisserman 2015) composed of four sequential convolutional blocks using 1-D kernels with ReLU activations and interleaved max-pooling, progressively increasing the number of feature maps from 32 to 256 while reducing the temporal resolution to form compact, hierarchical representations. The resulting tensor is permuted to a time-major format and processed by a stack of three bidirectional LSTM layers to capture long-range temporal dependencies and bidirectional context typical of nonstationary artifact dynamics. The recurrent output is then permuted back and aggregated by adaptive global average pooling to obtain a fixed-length embedding independent of the input window length. Finally, a fully connected classifier with progressively expanding hidden dimensions, ReLU nonlinearities, and dropout ($p = 0.2$) maps the pooled representation to three output logits.

2.3.3. Denoiser

A modified version of the IC-U-Net (Farabbi *et al* 2025), a specific deep denoising autoencoder already described in the literature (Chuang *et al* 2022), was used in this study. A deep denoising autoencoder is a specific neural network designed to map artifact-affected inputs, y , to their reconstructed outputs, y_d , via encoding and decoding processes. This enables the network to approximate the clean target signals. The objective function, $\mathcal{L}$, is a reconstruction loss function minimizing the disparity between the clean targets and the outputs of the network. In this study, the proposed network consists of n parallel input channels, where n represents the number of EEG signals recorded from the n channels under examination. In the encoder, each input EEG channel is processed independently by a dedicated branch operating in its own latent space. This design enables the extraction of channel-specific feature representations. The encoders subsequently feed into a shared bottleneck, followed by a unified decoder. The model input is a three-dimensional tensor (batch-size, window-length, channel-number). The model provides n distinct outputs, corresponding to each channel of denoised EEG signals for each channel.

Ensemble loss function The denoising network was trained using a composite loss function combining point-wise reconstruction error with temporal and spectral similarity terms. Similar multi-component

objectives, integrating time- and frequency-domain criteria, have previously been employed for EEG denoising (Chuang *et al* 2022). In this study, the loss function was formulated to promote sample-wise reconstruction accuracy together with waveform similarity, preservation of local temporal dynamics, and spectral-shape consistency between the denoised and target signals. The composite loss consisted of four terms:

1. Mean squared error (MSE), defined as:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_{d,i} - y_{t,i})^2 \quad (1)$$

where $y_{t,i}$ is the ASR-derived target and $y_{d,i}$ is the denoiser-predicted sample of the signal at instant i , and N is the number of samples in the window. This term penalizes pointwise differences between the predicted and target signals.

2. Correlation penalty ($\mathcal{L}_{\text{corr}}$):

$$\mathcal{L}_{\text{corr}} = 1 - \frac{\sum_i (y_{t,i} - \bar{y}_t)(y_{d,i} - \bar{y}_d)}{\sqrt{\sum_i (y_{t,i} - \bar{y}_t)^2 \sum_i (y_{d,i} - \bar{y}_d)^2}} \quad (2)$$

This term penalizes differences in the overall waveform structure independently of their absolute scale.

3. Temporal loss ($\mathcal{L}_{\text{temp}}$): to capture temporal dynamics, the difference between the temporal derivatives of the ground-truth signal and predicted signals is considered:

$$\mathcal{L}_{\text{temp}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left| (y_{t,i+1} - y_{t,i}) - (y_{d,i+1} - y_{d,i}) \right| \quad (3)$$

This term compares the first-order temporal differences of the target and predicted signals and penalizes discrepancies in their local temporal dynamics.

4. Spectral similarity loss ($\mathcal{L}_{\text{ss}}$):

$$\mathcal{L}_{\text{spectral}} = \frac{1}{F} \sum_{k=1}^F \left(\frac{|Y_{t,k}|}{\sum |Y_{t,k}|} - \frac{|Y_{d,k}|}{\sum |Y_{d,k}|} \right)^2 \quad (4)$$

where $Y_{t,k}$ and $Y_{d,k}$ represent the spectral coefficients of the target signal and of the denoised signal, respectively. The normalized spectral magnitudes are compared to reduce differences in the relative distribution of signal energy across frequencies.

The total training objective was obtained as the weighted sum of the four components:

$$\mathcal{L}_{\text{tot}} = \alpha \cdot \text{MSE} + \beta \cdot \mathcal{L}_{\text{corr}} + \gamma \cdot \mathcal{L}_{\text{temp}} + \delta \cdot \mathcal{L}_{\text{spectral}} \quad (5)$$

where α , β , γ , and δ determine the relative contribution of the individual terms.

2.3.4. Training strategy

The three blocks of the pipeline were trained independently. The subject-level dataset split and the leave-one-subject-out (LOSO) validation scheme ensured that data leakage was avoided during training (Brookshire *et al* 2024).

Gate classifier The obtained dataset was split at the subject level into two separate subsets: a development set (5 subjects) for hyperparameter tuning and an evaluation set (24 subjects) for the following training and testing of the model. This dataset split was adopted to prevent any possible subject-level data leakage. Hyperparameter tuning was performed on the development set using the LOSO cross-validation scheme. This strategy is particularly relevant for wearable-EEG applications, as the subject-specific signal characteristic of the application may introduce selection bias and adversely affect the generalizability of the model. Preprocessing, such as feature normalization and class weight computation, was performed using only the training subjects of each fold.

Artifact classifier The artifact classifier was trained and tested using a LOSO scheme, applied to all the subjects in the dataset. All convolutional and linear layers were initialized with HeNormal (He *et al* 2015) initialization, which is particularly suited for deep networks with ReLU activation functions.

Denoiser The denoiser was trained and evaluated using a LOSO scheme across all subjects in the dataset. The target signals for training were obtained by cleaning the input signals with the ASR algorithm. To mitigate overfitting and enhance network robustness, multiple regularization strategies were adopted. In particular, dropout layers were introduced after each convolutional block, and L2 regularization was applied to constrain weight magnitudes and promote smoother optimization dynamics. The loss weights were set as follows: $\alpha = 0.4$, $\beta = 0.35$, $\gamma = 0.20$, $\delta = 0.05$.

2.3.5. Performance evaluation

Performance evaluation was performed independently for each of the three pipeline components, as well as for the complete end-to-end pipeline. The inference time for the models was estimated on a consumer-level laptop equipped with an Intel Core i7 CPU and 64 GB of RAM. Measurements were performed on a per-window basis using CPU execution only.

Gate classifier The performance of the gate classifier was evaluated using standard classification metrics, namely balanced accuracy, precision, and recall, computed on a per-subject and per-artifact-type basis. The operating behavior of the selected gate classifier was characterized with a post-hoc threshold analysis using the predictions generated within the LOSO evaluation scheme. The ROC curve was computed by pooling the out-of-subject predictions from all LOSO folds. The classifier decision threshold was then varied over the interval $[0, 1]$, and threshold-dependent metrics were computed separately for each subject. Residual artifact leakage was quantified as the false-negative rate of the artifact class

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

representing the proportion of artifact-contaminated windows incorrectly routed through the artifact-free path. The predicted positive rate

$$\text{PPR} = \frac{\text{TP} + \text{FP}}{N} \quad (7)$$

was used as a proxy for denoising load, as it corresponds to the fraction of windows routed to the denoising network. For each threshold, both metrics were summarized across subjects using the median and interquartile range, thereby preserving subject-level variability while quantifying the trade-off between reducing residual artifact leakage and increasing the computational burden of the denoising stage.

Artifact classifier The artifact type classifier was evaluated using classification metrics, including overall and class-wise precision, recall, and $F1$ -score, to assess both the overall performance and the ability to discriminate between eye blinks, horizontal eye movements, and facial movements.

Denoiser The denoising quality was evaluated using time- and frequency-domain metrics computed between the denoiser output y_d and the target signal y_t over each window. The metrics are defined as follows:

- range-normalized RMSE (nRMSE): the root-mean-square error between the target y_t and the denoised output y_d is normalized by the peak-to-peak amplitude of the target window to obtain a scale-invariant error measure;
- spectral cosine similarity: the spectral cosine similarity was computed between y_t and y_d L1-normalized magnitude spectra. This metric assesses the agreement between the spectral profiles of the target and the denoised outputs, with an emphasis on the shape of the spectra rather than absolute scale;
- log-distance of normalized PSD: the PSDs of y_t and y_d were computed and normalized by broadband power in the 0.5–30 Hz range. Then, the distance between the $\log_{10}$ -PSD was computed in the 0.5–30 Hz range

$$\text{LSD} = \sqrt{\frac{1}{K} \sum_{k=1}^K (10 \log_{10}(P_d[k]) - 10 \log_{10}(P_t[k]))^2} \quad (8)$$

where P_d and P_t are the normalized PSDs of the denoised output and target, respectively.

The contribution of the temporal and spectral components was examined through an ablation analysis, comparing MSE alone, MSE with correlation penalty and temporal loss terms, MSE with the spectral term, and the complete loss.

Complete pipeline The full pipeline was evaluated in an end-to-end setting in which the predicted output of the gate classifier determines whether an EEG window is routed to the correction and classification branch. This protocol reflects the intended deployment behavior and provides a realistic estimate of the correction quality achievable by the complete system. The end-to-end evaluation of the artifact management pipeline was conducted with the LOSO scheme in accordance with the assessment of the single blocks. The subject-level split was maintained for all the parts to prevent data leakage. The windows presented to the input of the pipeline are all extracted from the EEG recorded in the artifact elicitation phases. The artifact-free windows included in each LOSO fold validation set are extracted from clean segments of the EEG and preprocessed as described in 2.2. The gate classifier was applied to each EEG window to detect artifact-affected and artifact-free windows. The artifact type classification performance was evaluated with multiclass classification metrics. The denoiser evaluation was performed by applying the denoiser to all windows, independently of the gate decision. The resulting metrics (nRMSE and spectral cosine similarity) were subsequently stratified according to the gate outcome. This analysis was used to characterize denoiser performance on the different subsets identified by the gate, including erroneous routing conditions.

2.4. Architecture assessment on an external dataset

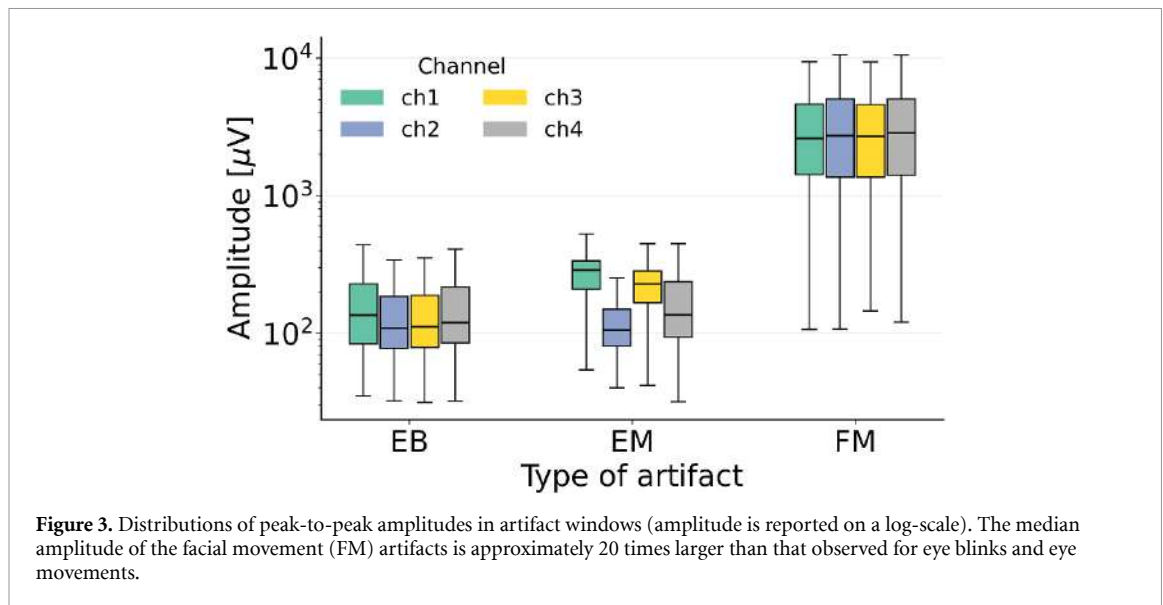
The proposed architecture was further evaluated on an independent public benchmark, the EEG eye artifact dataset (EEAD) (Kobler *et al* 2021), considering 24 subjects from two studies included in the dataset. The EEAD comprises EEG recorded with conventional high-density EEG caps in a laboratory environment. The EEAD studies focused on blink, horizontal eye movement, and vertical eye movement artifacts. To ensure consistency with the artifact types considered in this work, only blink and horizontal eye movement artifacts were included in the comparison. Furthermore, four channels located in positions comparable to those of the EEG-enabled eyewear prototype (namely FT7, T7, FT8, T8) among those available in the dataset were used for the artifact management pipeline assessment. The same preprocessing and target extraction procedures, as well as the LOSO validation scheme, were used for the wearable dataset and applied to the external data in order to ensure methodological consistency. Because direct transfer of EEG models across datasets is known to be difficult (dataset shift problem), the models were retrained on the external dataset. In machine learning EEG applications, cross-dataset variability can markedly weaken the generalization ability of models, due to signal non-stationarity and to differences in instrumentation, electrode placement, contact impedance, and recording environment across datasets (Jayaram and Barachant 2018, Apicella *et al* 2024). Consequently, this experiment should be interpreted as an assessment of the proposed architecture on a different EEG domain.

2.5. Denoiser assessment on semi-synthetic data

Semi-synthetic contaminated EEG windows were generated using a procedure based on EEGDenoiseNet (Zhang *et al* 2021), while deriving artifact signals directly from the wearable EEG recordings. For each subject and artifact type, artifact-affected windows were aligned at the center of the window and combined using a channel-wise median. The resulting multichannel templates were then smoothed using a Savitzky–Golay filter to reduce residual EEG fluctuations while preserving the artifact morphology. This subject-level extraction preserved differences in artifact amplitude and spatial distribution across participants. Each artifact template n was added to a clean EEG window x according to

$$y = x + \lambda n \quad (9)$$

where y is the semi-synthetic contaminated window and λ controls the artifact strength. For each subject and artifact type, λ was selected to reproduce the corresponding median empirical SNR (computed between artifact template and clean windows), as well as two additional contamination levels at -3 and $+3$ dB relative to the median. The semi-synthetic data were used only for testing. The network trained on the original real-data training subjects was kept fixed and applied to semi-synthetic windows generated from clean recordings using the same LOSO scheme as in the previous evaluations. Denoising performance was assessed by comparing the network output with the original clean window on denoising metrics. The semi-synthetic dataset was also used to compare the network output with a no-processing



reference by comparing the semi-synthetic contaminated inputs with the corresponding clean windows using the same metrics.

2.6. Comparison with deep-learning denoising methods

The proposed denoising method, a modified IC-U-Net architecture, was compared with DeepSeparator (Yu *et al* 2022) and Novel CNN (Zhang *et al* 2021). These architectures were selected because their native single-channel input allows channel-wise retraining and inference, thereby enabling their application to the four-channel eyewear EEG configuration. Both comparison networks were retrained on the eyewear EEG dataset using the same LOSO scheme adopted for the modified IC-U-Net. Since DeepSeparator and Novel CNN operate on single-channel inputs, each of the four EEG channels was treated as an independent training example and processed separately during inference, using shared model weights across channels. Their denoising performance was evaluated on the same test windows used for the modified IC-U-Net.

3. Results

This section reports the performance of the proposed artifact-management pipeline under the LOSO evaluation protocol to detect, classify, and correct the three artifact types considered: eye blinks, horizontal eye movements, and facial movements. First, it should be noted that the FM artifact exhibits an amplitude approximately 20 times larger than EB and EM, as it is clearly shown by the peak-to-peak amplitude distributions (figure 3) and by the example signals reported in figure 4. The gate classifier is assessed in terms of its ability to detect artifact-contaminated windows and to support selective processing. Then, the performance of the artifact type classifier in discriminating among the three artifact classes is reported. The denoiser signal reconstruction performance is quantified, reporting results per channel and artifact type, aggregated across LOSO folds. Finally, the results of the end-to-end pipeline evaluation, the external dataset evaluation, the semi-synthetic dataset evaluation, and the comparison with baseline models are presented.

3.1. Gate classifier

The balanced accuracy (BA) obtained with the three evaluated model families, LR, SVM, and RF, is reported in figure 5 for each held-out subject in the LOSO scheme. A BA > 0.8 can be observed for the majority of subjects. Table 1 summarizes the gate classifier performance across all subjects, reported as median and quartiles over LOSO folds. For each model family, balanced accuracy, precision, and recall for the artifact class are reported. The RF achieved higher scores and slightly lower variability than LR and SVM on balanced accuracy and precision, with a recall slightly lower than LR but higher than SVM. The better performance of the RF classifier was confirmed by statistical analysis on the balanced accuracy achieved across subjects. The Friedman test confirmed the presence of significant differences among the three classifiers at a 0.05 significance level ($p = 0.002$). The Wilcoxon test with Holm correction was

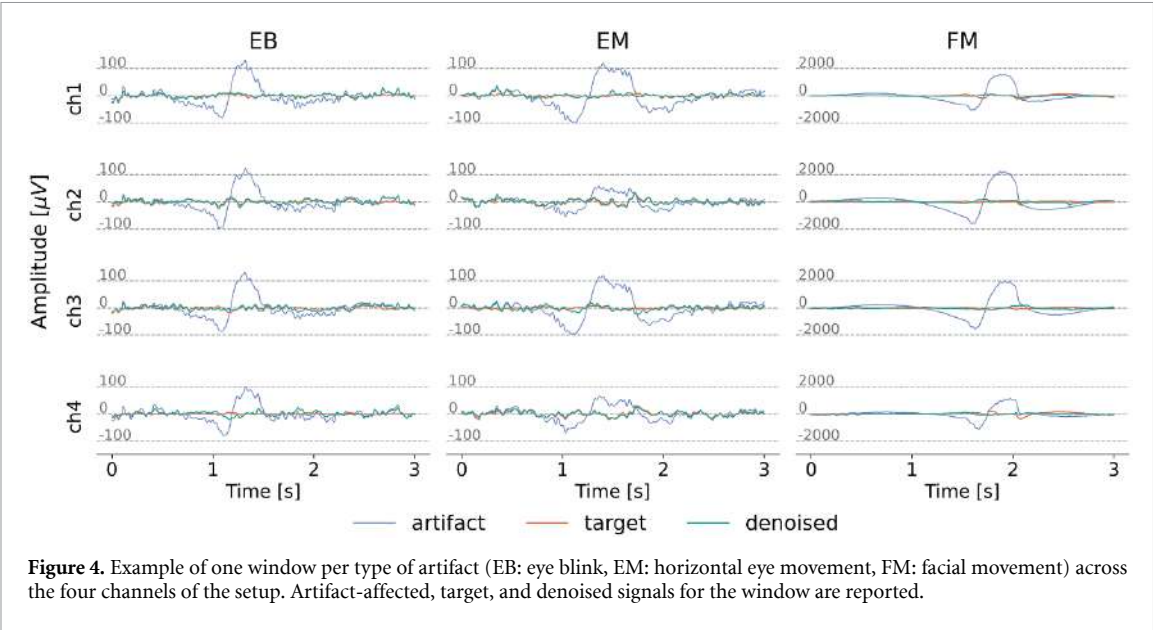


Figure 4. Example of one window per type of artifact (EB: eye blink, EM: horizontal eye movement, FM: facial movement) across the four channels of the setup. Artifact-affected, target, and denoised signals for the window are reported.

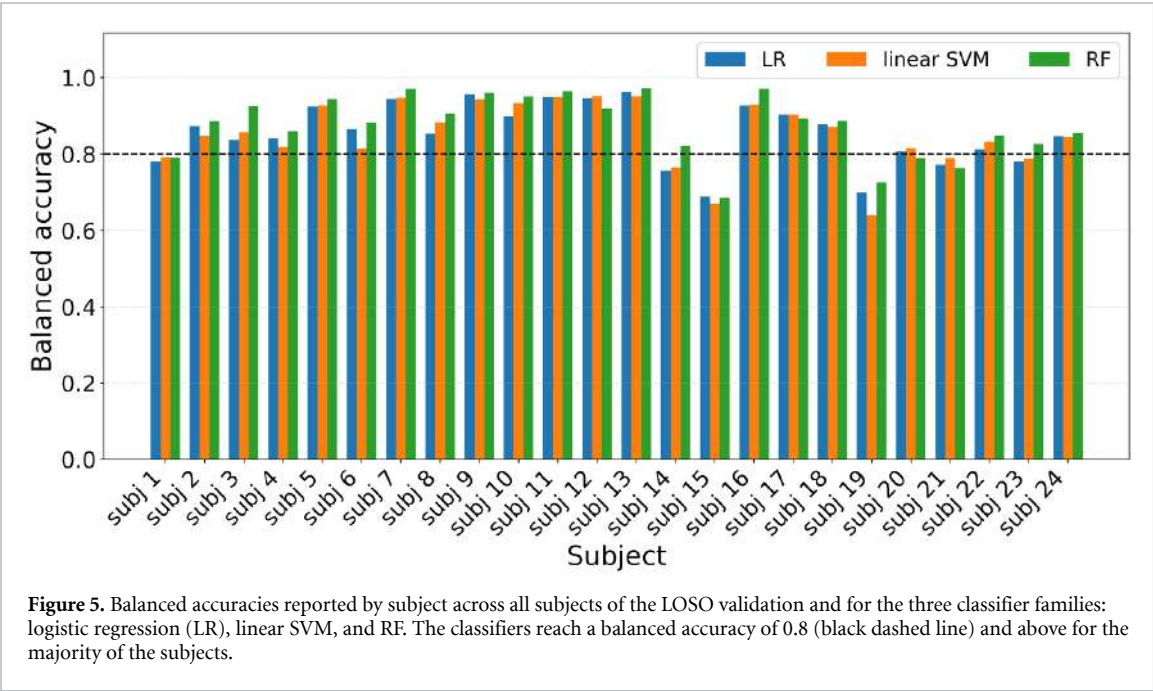


Figure 5. Balanced accuracies reported by subject across all subjects of the LOSO validation and for the three classifier families: logistic regression (LR), linear SVM, and RF. The classifiers reach a balanced accuracy of 0.8 (black dashed line) and above for the majority of the subjects.

Table 1. Gate classification metrics across subjects reported as median and quartiles.

Classifier	Balanced accuracy	Precision	Recall
LR	0.86 [0.80, 0.92]	0.90 [0.82, 0.96]	0.94 [0.77, 0.99]
SVM	0.85 [0.81, 0.93]	0.89 [0.83, 0.94]	0.92 [0.76, 0.99]
RF	0.89 [0.82, 0.94]	0.93 [0.87, 0.97]	0.93 [0.78, 0.98]

used for pairwise comparisons of the classifiers, showing significant differences at 0.05 significance level for RF compared to LR ($p = 0.004$) and SVM ($p = 0.006$). The performance breakdown by artifact type (figure 6) indicates that eye blink windows are the most difficult to identify as artifact-contaminated, with large variability across subjects, whereas facial movement windows are the most easily separable across all subjects. Among the evaluated models, the RF classifier provides the best performance across all artifact types. The execution time of this block was estimated per processed window as 7 ± 1 ms (mean $\pm$ standard deviation).

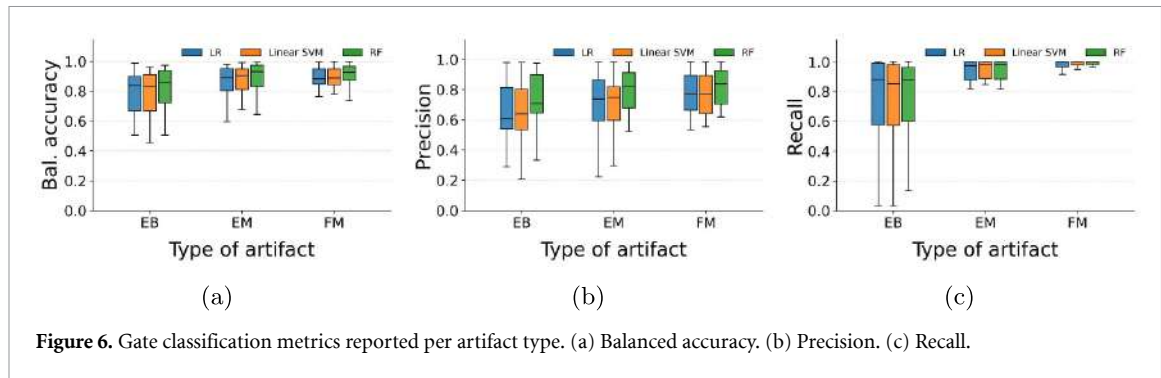


Figure 6. Gate classification metrics reported per artifact type. (a) Balanced accuracy. (b) Precision. (c) Recall.

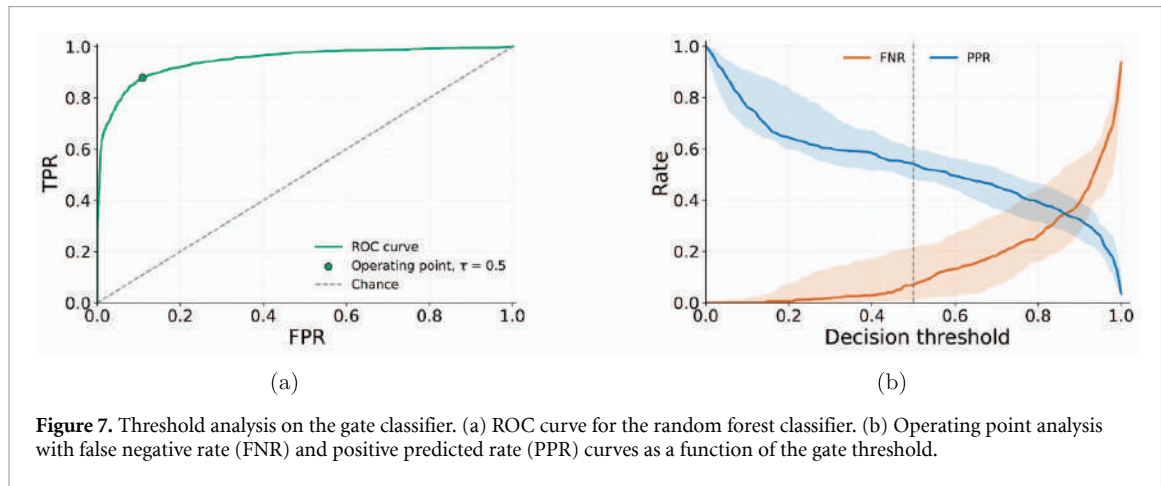


Figure 7. Threshold analysis on the gate classifier. (a) ROC curve for the random forest classifier. (b) Operating point analysis with false negative rate (FNR) and positive predicted rate (PPR) curves as a function of the gate threshold.

3.1.1. Gate threshold analysis

The pooled ROC curve obtained from the held-out LOSO predictions showed that the RF classifier achieved good separation between clean and artifact-contaminated windows (figure 7(a)). At the default decision threshold of $\tau = 0.5$, the operating point corresponded to a true-positive rate of 0.88 and a false-positive rate of 0.11. The threshold analysis further characterized the trade-off between residual artifact leakage (FNR) and denoising workload. At $\tau = 0.5$, the median artifact false-negative rate across subjects was below 0.1, while approximately half of the input windows were routed to the denoising network. The interquartile ranges indicated subject-dependent variability in both quantities, which becomes more pronounced at higher thresholds. Overall, the default threshold provides a favorable operating point, maintaining a low residual artifact rate without routing all windows to the denoising stage.

3.2. Artifact type classifier

Artifact type classifier performance is reported under the LOSO validation protocol for the three artifact types. Table 2 reports the confusion matrix aggregated across all LOSO folds (i.e. pooled predictions over all held-out subjects) for the three-class discrimination problem. Table 3 reports the overall balanced accuracy and $F1$ -score, as well as the class-wise precision, recall, and $F1$ -score. Metrics are computed for each subject separately and then aggregated. The balanced accuracy achieved a median of 0.84, with first and third quartiles of 0.68 and 0.92, respectively, highlighting a strong dependence on the considered subject. The artifact type classifier achieved higher and less dispersed classification scores for the facial movement artifacts (median $F1$ -score 0.91). The classification of ocular artifacts yielded lower (median $F1$ -score EB: 0.80, EM: 0.87) and more dispersed performance across subjects. The inference time of this block was estimated to be 10 ± 2 ms (mean $\pm$ standard deviation), providing a practical indication of its computational cost.

3.3. Denoiser

The denoiser output y_d is evaluated against the target y_t using the metrics defined in section 2.3.5. Results are summarized per artifact type (EB, EM, FM) and per channel. The distribution of range-normalized RMSE (nRMSE) is reported in figure 8(a). All the groups show similar medians (< 0.19 for EB, ≤ 0.20 for EM, and ≤ 0.18 for FM) and dispersion. Figure 8(b) reports the distribution of spectral cosine similarity with similar medians for EB (> 0.90) and EM (> 0.87), and slightly lower median

Table 2. Confusion matrix reporting the classification predictions of the artifact classifier. The classification shows a high ability to distinguish between facial movements (FM) and ocular artifacts, but a lower capacity to separate eye blinks (EB) from eye movements (EM). The numbers of correctly classified EEG windows are shown in bold.

		Predicted label		
		EB	EM	FM
True label	EB	1121	264	46
	EM	235	1312	40
	FM	128	236	1362

Table 3. Per-class and overall performance (LOSO) of the artifact type classifier, with median and quartiles across subjects.

Class	Precision	Recall	F1-score
EB	0.88 [0.63, 0.97]	0.84 [0.72, 0.91]	0.80 [0.68, 0.91]
EM	0.81 [0.67, 0.93]	0.93 [0.81, 0.98]	0.87 [0.65, 0.93]
FM	0.97 [0.91, 0.99]	0.93 [0.62, 0.98]	0.91 [0.75, 0.97]
		Balanced accuracy	Macro-F1
Overall		0.84 [0.68, 0.92]	0.83 [0.62, 0.91]

and more dispersed values for FM (> 0.84). The log-spectral distance shows the distance between the broadband-normalized PSD between y_d and y_t (figure 8(c)). While the median LSD remains close across channels for EB (< 2.3) and EM (< 2.4), the FM artifact type has much higher medians (> 3.54) and dispersion. The loss ablation study was conducted on a subset of 15 subjects. No statistically significant differences were found among the loss configurations for either nRMSE or spectral cosine similarity. The observed variations were small and did not indicate a consistent advantage of the composite loss over MSE alone. The inference time of this block was estimated to be 29 ± 10 ms (mean $\pm$ standard deviation).

3.4. Complete pipeline

The RF model was selected as the gate classifier in this evaluation for its greater ability to separate artifact-affected and artifact-free EEG windows. Table 4 reports the nRMSE and spectral cosine similarity metrics computed by applying the denoiser to all EEG windows, stratified according to the gate classification outcome (TP: true positives; FP: false positives; FN: false negatives; TN: true negatives). The results show similar reconstruction metrics in the time and frequency domains across all gate-outcome groups, including both windows that would be passed to the denoising branch during pipeline operation (TP and FP) and those that would bypass it (FN and TN).

Regarding classification metrics for the artifact classifier evaluated in the complete pipeline, the confusion matrix in table 5 shows the overall classification, including the GC classification (clean class). The metrics are reported in table 6. The lowest classification metrics are obtained for the EB class, with the presence of artifact-free windows among those routed to the classification network substantially affecting the overall classification performance.

3.5. Architecture assessment on an external dataset

The three blocks of the artifact management pipeline were evaluated on an external dataset as described in 2.4. For the GC, the best-performing artifact detector on the external dataset was the RF, with a median balanced accuracy of 0.82, with first and third quartiles of 0.77 and 0.87, respectively, achieving a lower artifact detection performance result than with the eyewear EEG dataset (0.89 [0.82, 0.94], cfr. Table 1). The artifact classification and denoising metrics are reported for the EB and EM classes, which are included in both datasets. The artifact classifier obtained the following metrics on the EEAD for EB class: precision 0.98 [0.95, 1], recall 0.98 [0.96, 1], and F1-score 0.99 [0.97, 1]; for EM class, the results were: precision 0.99 [0.96, 1], recall 0.99 [0.97, 1], and F1-score 0.98 [0.97, 1]. The results obtained on the external dataset were remarkably higher and less dispersed across subjects than those obtained on the eyewear EEG (cfr. Table 3). Regarding the denoiser, it obtained across channels nRMSE of 0.17 [0.14, 0.20] for EB and 0.15 [0.13, 0.17] for EM, spectral cosine similarity of 0.89 [0.82, 0.94] on EB and of 0.91 [0.87, 0.94] on EM, and log-spectral distance of 1.23 [0.98, 1.57] on EB and of 0.97 [0.76, 1.24].

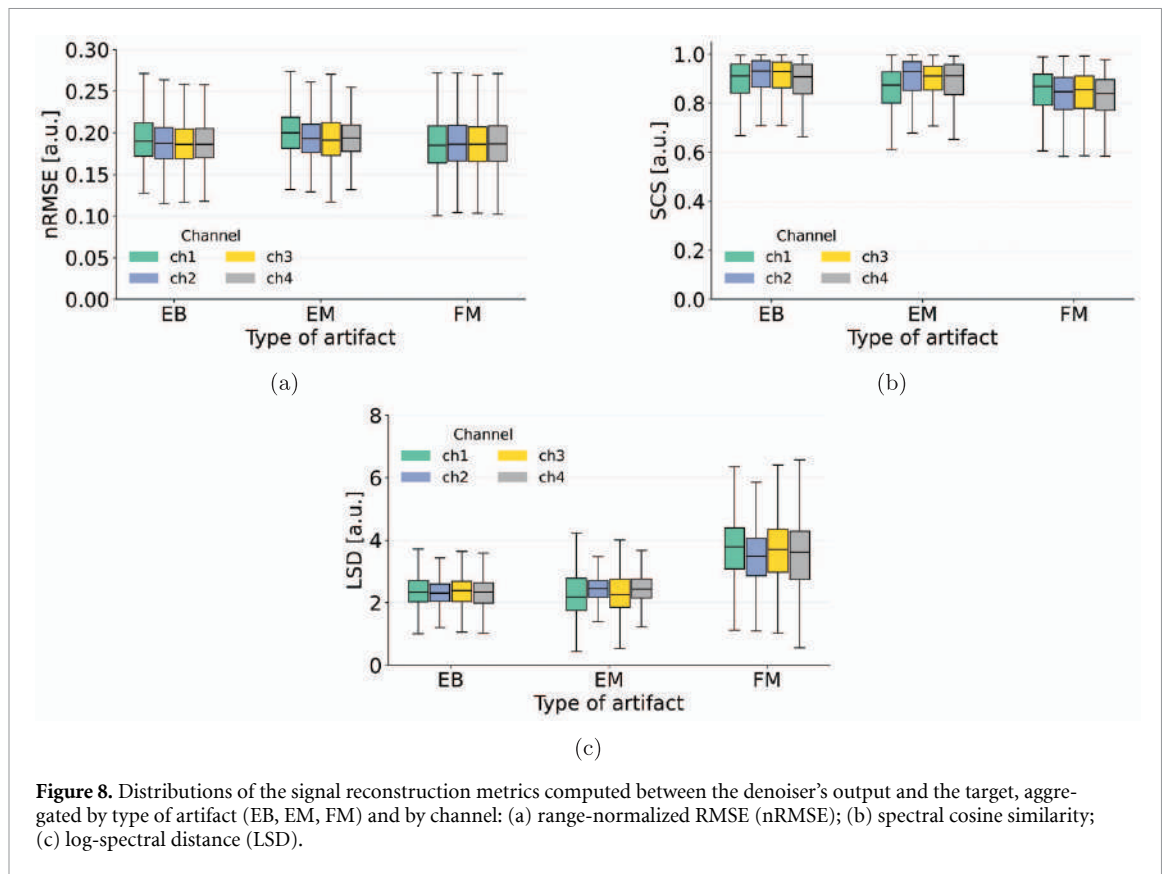


Figure 8. Distributions of the signal reconstruction metrics computed between the denoiser's output and the target, aggregated by type of artifact (EB, EM, FM) and by channel: (a) range-normalized RMSE (nRMSE); (b) spectral cosine similarity; (c) log-spectral distance (LSD).

Table 4. Signal reconstruction metrics obtained by the denoiser. Normalized RMSE and spectral cosine similarity are reported as *median [first quartile, third quartile]*, stratified by correct and incorrect gate classifications (TP: true positives; FP: false positives; FN: false negatives; TN: true negatives).

	nRMSE	Spectral cosine similarity
Overall	0.19 [0.17, 0.21]	0.89 [0.81, 0.94]
TP	0.19 [0.17, 0.21]	0.88 [0.81, 0.94]
FP	0.20 [0.18, 0.22]	0.87 [0.81, 0.93]
FN	0.19 [0.17, 0.21]	0.92 [0.85, 0.97]
TN	0.20 [0.18, 0.23]	0.89 [0.81, 0.93]

The corresponding values on the eyewear EEG dataset were: nRMSE of 0.19 [0.17, 0.21] for EB and 0.19 [0.18, 0.21] for EM, spectral cosine similarity of 0.92 [0.85, 0.97] on EB and of 0.90 [0.83, 0.95] on EM, and log-spectral distance of 2.34 [2.02, 2.65] on EB and of 2.36 [1.96, 2.75]. The comparison of denoising performance shows similar overall results in terms of nRMSE and spectral cosine similarity, with a marked improvement observed particularly for EM artifacts.

3.6. Denoiser assessment on semi-synthetic data

The models trained on real data acquired with eyewear EEG were then evaluated on semi-synthetic signals. The denoising performance across the three types of artifact in terms of nRMSE and spectral cosine similarity is reported in table 7. As a no-processing reference, the semi-synthetic contaminated inputs were also compared directly with the corresponding clean signals. At the empirical SNR levels, the unprocessed signals showed median nRMSE values of 0.27, 0.44, and 10.27 for EB, EM, and FM, respectively, compared with 0.19, 0.19, and 0.20 after denoising. The corresponding median spectral cosine similarities increased from 0.88, 0.83, and 0.75 for the unprocessed signals to 0.90, 0.88, and 0.78 after denoising. Although the denoised nRMSE values were comparable across artifact types, the no-processing reference indicates substantially different initial contamination levels, with the largest

Table 5. Confusion matrix of the complete pipeline, reporting the combined predictions of the gate classifier and the artifact classifier. The predictions are reported pooled for all subjects in the LOSO validation. The numbers of correctly classified EEG windows are shown in bold.

		Predicted label			
		clean	EB	EM	FM
True label	clean	2767	146	49	135
	EB	273	717	191	32
	EM	136	137	1005	27
	FM	81	96	148	1104

Table 6. Classification metrics for the overall pipeline, reported as median across subjects, considering the artifact-free class (clean) selected by the GC, and the artifact types detected by the classification branch.

		Overall		
Macro-F1		0.76		
Balanced accuracy		0.76		
	Clean	EB	EM	FM
F1	0.87	0.62	0.79	0.88
Recall	0.91	0.60	0.89	0.92

Table 7. Denoising metrics obtained on the semi-synthetic dataset at empirical-based SNR levels.

		nRMSE	Spectral cosine similarity
EB	0.19 [0.17, 0.20]	0.90 [0.85, 0.93]	
EM	0.19 [0.18, 0.21]	0.88 [0.82, 0.92]	
FM	0.20 [0.18, 0.21]	0.78 [0.72, 0.84]	

difference observed for FM. The semi-synthetic evaluation confirms the results obtained on real contaminated EEG, with comparable post-denoising nRMSE across artifact types despite their different unprocessed baseline values, but with a drop in spectral cosine similarity for FM artifacts. These results support the ability of the network to generalize beyond the ASR-derived training targets. The lower performance observed for FM in the spectral metric indicates an artifact-specific limitation, suggesting either reduced robustness to its spectral characteristics or a larger discrepancy between the semi-synthetic template and naturally occurring artifacts. Overall, the results support the validity of the original evaluation while highlighting that denoising effectiveness depends on artifact morphology.

3.7. Comparison with deep-learning denoising methods

The proposed modified IC-U-Net was compared with two deep-learning methods for EEG denoising, namely DeepSeparator and Novel CNN. The results of the comparison presented in figure 9 show that the modified IC-U-Net achieved nRMSE values comparable to those of Novel CNN, whereas DeepSeparator obtained the lowest median nRMSE for EB and EM artifacts, with values of 0.16 and 0.19, respectively, although its results were more dispersed for EB. Novel CNN achieved the lowest median nRMSE for FM artifacts, with a value of 0.19. In contrast, the modified IC-U-Net consistently provided the highest spectral cosine similarity across all artifact types, reaching median values of 0.94 for EB, 0.94 for EM, and 0.84 for FM. Overall, the comparison indicates that the modified IC-U-Net better preserved the spectral characteristics of the signals, while the single-channel comparison methods achieved slightly lower time-domain reconstruction errors for some artifact types.

4. Discussion

The proposed pipeline demonstrates that selective artifact management under wearable EEG constraints is a viable strategy to balance artifact suppression, signal preservation, and computational efficiency in

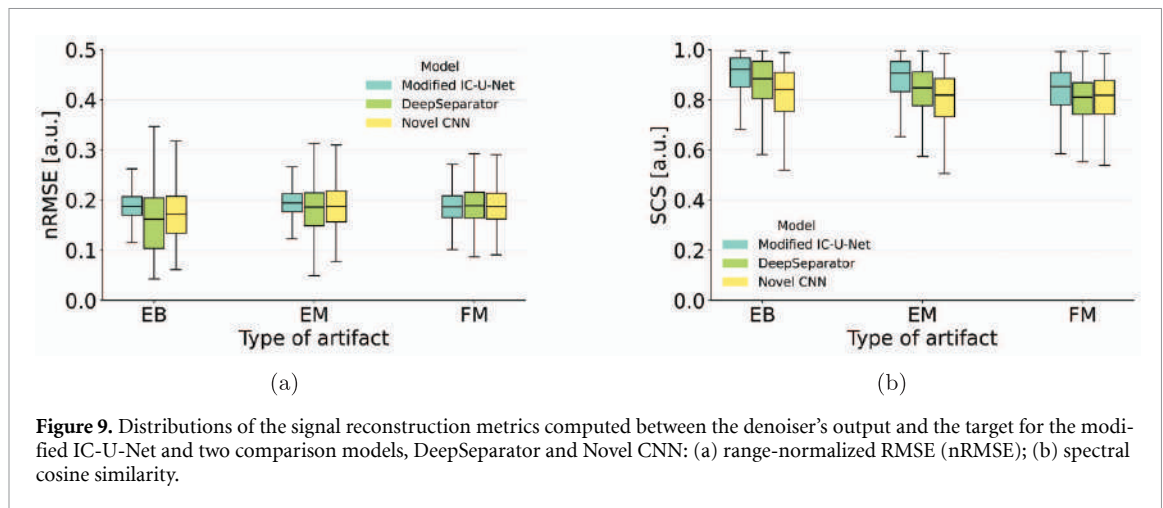


Figure 9. Distributions of the signal reconstruction metrics computed between the denoiser's output and the target for the modified IC-U-Net and two comparison models, DeepSeparator and Novel CNN: (a) range-normalized RMSE (nRMSE); (b) spectral cosine similarity.

low-density configurations. Across the LOSO folds, the gate classifier exhibited inter-subject variability in balanced accuracy (figure 5), with a dispersion that was consistent across the three model families. Despite this variability, the overall performance summarized in table 1 indicates that a lightweight, feature-based classifier is sufficient to support the gating stage, yielding aggregated median balanced accuracy greater than 0.85 and median artifact-class precision and recall greater than 0.90 across model families. The detection accuracy achieved by the gate classifier is slightly lower than the accuracies of 90% and 93% reached by Grosselin *et al* and Ingolfsson *et al* respectively (2019, 2024). Among the tested models, the RF achieved the highest balanced accuracy. This outcome is consistent with the ability of this classifier to capture nonlinear relationships and feature interactions that may not be represented by linear decision boundaries. The logistic regression and linear SVM provided lower but comparable performance, indicating that the classification problem is effectively captured by the selected feature set. The threshold analysis (figure 7) showed that the gate operating point can be adjusted according to the application requirements. The default threshold of 0.5 provided a suitable compromise between residual artifact leakage and computational workload. Nevertheless, the observed inter-subject variability suggests that subject-specific threshold optimization could further improve performance. The favorable results for this stage obtained with time-domain features and lightweight classifiers are promising for the integration of this approach into a short window-based embedded pipeline.

The artifact type classifier showed marked variability across held-out subjects. The aggregated confusion matrix (table 2) indicates that facial movements were better distinguished from ocular artifacts (EB and EM), whereas confusions were primarily concentrated between eye blinks and horizontal eye movements. This pattern is reflected by the class-wise metrics distributions in table 3, where FM achieved higher and more concentrated values, while EB and EM exhibited comparatively lower and more dispersed precision, recall, and *F1*-scores.

The relative ease of separating FM from ocular artifacts is consistent with the distinct temporal morphology and amplitude characteristics typically associated with facial muscle activity in the considered wearable montage, as the reference electrode is directly and highly impacted by this type of artifact. In contrast, EB and EM share overlapping spectral and temporal signatures in short windows, which can reduce separability when only a limited number of channels is available (Plöchl *et al* 2012).

Artifact correction results indicate that the artifact correction network reduces the artifact effect as quantified by both range-normalized RMSE and spectral cosine similarity computed between the denoiser output and ASR-obtained target figures 8(a) and (b). The nRMSE distributions across channels and artifact categories show moderate agreement with the target in the time domain. Spectral cosine similarity, however, indicates a good preservation of the spectral profile after correction with respect to the target. Differences across artifact types are slightly visible in the frequency domain, with lower median cosine similarity and larger dispersion for FM artifacts. The results are in line with those obtained by Xing and Casso with a denoising autoencoder (nRMSE 0.52 and 0.60 on EOG and EMG artifacts, respectively) (2024), and by Arpaia *et al* (median nRMSE 0.3 and 0.46 for ocular and blink artifacts) (2023). The log10 distance of PSD (LSD) was treated as a secondary indicator. In the present formulation, the artifact correction network's loss does not optimize explicitly for power reconstruction; rather, it includes a term for spectral shape similarity. Under these conditions, LSD is expected to be more sensitive to residual mismatches in broadband power and localized spectral deviations. The largest LSD values were observed for FM (medians across channels greater than 3.5) relative to EB and EM

(medians across channels lower than 2.4). This trend is consistent with the stronger disruption induced by FM artifacts in the adopted wearable configuration, where large-amplitude peaks can dominate short windows and alter the estimated spectra more substantially. Additional analyses are required to disambiguate whether elevated LSD in FM is primarily driven by incomplete suppression, ASR target limitations, or by montage-specific artifact effects (reference electrode on the nose pad). The results of the ablation analysis on the loss suggest that, under the evaluated conditions, the additional loss terms had a limited measurable influence on denoising performance. The evaluation of the complete pipeline considers the behavior of the three blocks working together in the intended usage scenario. The results suggest that the gated pipeline does not introduce large reconstruction errors in any gate-outcome group. The classification is, as expected, more affected due to the presence of artifact-free windows, which are mainly classified as eye blinks. This specific misclassification can be explained by the low amplitude of the eye blinks at the temporal level. Even though the amplitudes are similar to those observed for horizontal eye movements, for these latter artifacts, the artifact classifier can exploit the asymmetries in the signals recorded from the two sides of the head (cfr. Figure 3). The results of the assessment of the artifact management pipeline on the EEG eye artifact public dataset did not show a uniform increase or decrease in performance across pipeline stages, reflecting stage-specific challenges. The slightly lower artifact detection performance observed on the external dataset suggests a better alignment between the adopted statistics-based time domain features and the characteristics of the eyewear EEG domain. This is consistent with the domain shift problem in EEG models (Apicella *et al* 2024). The slightly better nRMSE results on the external dataset likely reflect a more favorable setup for denoising in the external dataset. In particular, the external recordings involved a narrower and more stereotyped set of artifact cases. Regarding the artifact-type classification, the markedly higher performance on the external dataset was expected, since the external benchmark involved only two artifact classes with more distinct waveform characteristics, whereas the wearable dataset included three classes and less-separable waveform shapes for EB and EM, due to fundamental differences in montage and reference. These findings suggest that the relative difficulty of the individual pipeline stages depends strongly on dataset-specific factors, including acquisition setup, artifact taxonomy, and class separability.

The semi-synthetic evaluation provides additional evidence that the denoiser can generalize beyond the ASR-derived training targets. However, this evaluation should take into account that assessment using a semi-synthetic dataset represents a partial domain shift and that the injected artifacts only approximate the complexity of naturally occurring contamination. The comparison with the unprocessed semi-synthetic signals further supports the effectiveness of the denoiser, showing improvements especially in nRMSE for the more strongly contaminated EM and FM signals. The lower spectral cosine similarity observed for FM supports the hypothesis that this artifact type presents greater challenges for preserving the spectral characteristics of the EEG signal during denoising. Future work should therefore investigate more realistic models of contamination and evaluate whether artifact-specific training or augmentation strategies can improve spectral reconstruction.

The comparison of the modified IC-U-Net with DeepSeparator and Novel CNN denoising models shows that, in the eyewear EEG domain, the modified IC-U-Net consistently achieved the highest spectral cosine similarity across all artifact types, indicating better preservation of the spectral profile, while the comparison models achieved slightly lower median nRMSE for some artifact classes. Therefore, the main advantage of the proposed denoiser lies in the spectral preservation while maintaining broadly comparable, although not consistently higher, time-domain reconstruction accuracy.

Although the proposed method for artifact management was evaluated in an offline setting, the specific use case of consumer-level EEG-enabled eyewear and the general evolution of edge-AI hardware clearly suggest wearable deployment as a relevant future development direction. There is growing interest in moving AI inference to wearable devices, particularly to reduce latency, limit data transmission, and improve privacy (Diab and Rodriguez-Villegas 2022). TinyML optimization strategies and dedicated biomedical benchmarks further support the deployment of machine- and deep-learning models on constrained devices (Samakovlis *et al* 2025). The GAP9 (GreenWaves Technologies), MAX78000 (Analog Devices), Apollo510 (Ambiq), STM32N6 series (STMicroelectronics), MCX N series (NXP), and RA8D1 (Renesas) are examples of recent embedded-AI platforms, which illustrate the growing availability of microcontroller- or system-on-a-chip-class devices with DSP extensions or neural accelerators.

In conclusion, a central challenge arises from adapting an artifact correction network originally designed for higher-density recordings to a low-density wearable setup with four temporal channels and nose referencing. The reduced spatial resolution limits the availability of redundancy and spatial cues, and the reference configuration can amplify certain artifact components, especially for facial movements. These constraints may affect both artifact separability (EB versus EM) and correction stability across subjects. The overall performance reliability of the pipeline could also benefit from the inclusion

of a probability-weighted gate classifier instead of the proposed hard-threshold one. A viable practical direction for improving robustness is the inclusion of a short subject-specific calibration phase, which is a common step in many wearable devices. A brief calibration session could be used to record subject-specific artifacts and estimate scaling to adjust the gate threshold and fine-tune a small subset of model parameters. Such calibration is expected to reduce inter-subject variability observed in both gating and artifact-type classification and to improve artifact correction consistency.

5. Conclusions

This work presented a gated artifact-management pipeline for low-density wearable EEG comprising (i) a lightweight gate classifier for artifact detection, (ii) a three-class artifact type classifier for eye blinks, horizontal eye movements, and facial movement artifacts, and (iii) a deep artifact correction network for denoising. Under LOSO evaluation, the gate classifier achieved good aggregated balanced accuracy, with the RF providing the strongest overall performance among the tested model families. Artifact type classification performance exhibited substantial inter-subject variability. Facial movements were reliably separated from ocular artifacts, whereas eye blinks and eye movements remained the most confusable artifacts. Artifact correction performance was quantified using range-normalized RMSE and spectral cosine similarity across channels and artifact types, showing moderate time-domain and strong spectral profile agreement with the target. The log10 distance of normalized PSD was reported as a secondary metric, with the largest deviations observed for FM, motivating further analysis of montage-specific effects and residual spectral mismatches. Overall, the proposed pipeline implemented a selective artifact management strategy under wearable EEG constraints, in which gating limited intensive processing to artifact-contaminated segments while preserving artifact-free ones. Architecture assessment on EEAD with dataset-specific retraining, together with semi-synthetic analyses at controlled signal-to-noise ratios, further supported the robustness of the proposed approach while highlighting its sensitivity to dataset-specific characteristics. Finally, the comparison with single-channel denoising models on the eyewear EEG dataset provided better preservation of the spectral profile across artifact types while maintaining broadly comparable time-domain reconstruction accuracy, although slightly lower nRMSE values were achieved by the comparison models for some artifact types. These findings supported selective artifact management as a viable design principle for future wearable EEG systems operating under strict spatial and computational constraints. Future work will investigate subject-specific calibration strategies and architectural adaptations to improve the preservation of frequency content and account for the specific montage characteristics of the EEG setup.

Acknowledgments

This work was carried out in the EssilorLuxottica Smart Eyewear Lab, a Joint Research Center between EssilorLuxottica and Politecnico di Milano. This study was supported by a PNRR scholarship funded by the European Union NextGenerationEU M4C2-I3.3.

Data availability statement

The data cannot be made publicly available upon publication due to legal restrictions preventing unrestricted public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

Conflict of interest

Matteo Rossi and Niccoló Antonello are employees of EssilorLuxottica. Andrea Costanzo Palmisciano, Andrea Farabbi, Francesco Latino, Pietro Cerveri, and Luca Mainardi declare no conflict of interest. The authors affirm that the study design and the analysis of the results, interpretation, and reporting were not influenced by the commercial interests or technological development priorities of EssilorLuxottica.

ORCID iDs

Andrea Costanzo Palmisciano  0009-0007-2379-1465

Andrea Farabbi  0000-0001-5582-4654

Francesco Latino  0009-0008-9699-6274

Matteo Rossi  0000-0003-2519-0720

Niccoló Antonello  0000-0002-0803-5385

Pietro Cerveri  0000-0003-3995-8673

Luca Mainardi  0000-0002-6276-6314

References

- Alkhoury L, Scanavini G, Louviot S, Radanovic A, Shah S A and Hill N J 2025 Artifact-reference multivariate backward regression (ARMBR): a novel method for EEG blink artifact removal with minimal data requirements *J. Neural Eng.* **22** 036048
- Apicella A, Arpaia P, D'Errico G, Marocco D, Mastrati G, Moccaldi N and Prevete R 2024 Toward cross-subject and cross-session generalization in EEG-based emotion recognition: systematic review, taxonomy and methods *Neurocomputing* **604** 128354
- Arpaia P et al 2025 A systematic review of techniques for artifact detection and artifact category identification in electroencephalography from wearable devices *Sensors* **25** 5770
- Arpaia P, De Benedetto E, Esposito A, Natalizio A, Parvis M, Pesola M and Sansone M 2024 Artifacts removal from low-density EEG measured with dry electrodes 2024 *IEEE Int. Conf. on Metrology for EXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRaine)* (IEEE) pp 195–200
- Arpaia P, Esposito A, Natalizio A, Parvis M and Pesola M 2023 Low-density EEG correction with multivariate decomposition and subspace reconstruction *IEEE Sens. J.* **23** 23621–8
- Brookshire G, Kasper J, Blaich N M, Wu Y C, Glatt R, Merrill D A, Gerrol S, Yoder K J, Quirk C and Lucero C 2024 Data leakage in deep learning studies of translational EEG *Front. Neurosci.* **18** 1373515
- Cataldo A, Crisculo S, De Benedetto E, Masciullo A, Pesola M, Schiavoni R and Invitto S 2022 A method for optimizing the artifact subspace reconstruction performance in low-density EEG *IEEE Sens. J.* **22** 21257–65
- Chuang C-H, Chang K-Y, Huang C-S and Jung T-P 2022 IC-U-Net: a U-Net-based denoising autoencoder using mixtures of independent components for automatic EEG artifact removal *NeuroImage* **263** 119586
- Croft R and Barry R 2000 Removal of ocular artifact from the EEG: a review *Neurophysiol. Clin./Clin. Neurophysiol.* **30** 5–19
- Delorme A and Makeig S 2004 EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis *J. Neurosci. Methods* **134** 9–21
- Diab M S and Rodriguez-Villegas E 2022 Embedded machine learning using microcontrollers in wearable and ambulatory systems for health and care applications: a review *IEEE Access* **10** 98450–74
- Farabbi A, Ballabio F, Rossi M, Palmisciano A C, Antonello N, Trojaniello D, Ongarello T, Cerveri P and Mainardi L 2025 A two-stage deep learning approach for EEG artifact removal and classification: towards reliable wearable applications 2025 *47th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE) pp 1–5
- Grosselin F, Navarro-Sune X, Vozzi A, Pandremmenou K, De Vico Fallani F, Attal Y and Chavez M 2019 Quality assessment of single-channel EEG for wearable devices *Sensors* **19** 601
- He K, Zhang X, Ren S and Sun J 2015 Delving deep into rectifiers: surpassing human-level performance on imageNet classification 2015 *IEEE Int. Conf. on Computer Vision (ICCV)* (IEEE) pp 1026–34
- Hermans T, Smets L, Lemmens K, Dereymaeker A, Jansen K, Naulaers G, Zappasodi F, Van Huffel S, Comani S and De Vos M 2023 A multi-task and multi-channel convolutional neural network for semi-supervised neonatal artefact detection *J. Neural Eng.* **20** 026013
- Ingolfsson T M, Benatti S, Wang X, Bernini A, Ducouret P, Rylvlin P, Beniczky S, Benini L and Cossetti A 2024 Minimizing artifact-induced false-alarms for seizure detection in wearable EEG devices with gradient-boosted tree classifiers *Sci. Rep.* **14** 2980
- Islam M K, Rastegarnia A and Yang Z 2016 Methods for artifact detection and removal from scalp EEG: a review *Neurophysiol. Clin./Clin. Neurophysiol.* **46** 287–305
- Jayaram V and Barachant A 2018 MOABB: trustworthy algorithm benchmarking for BCIs *J. Neural Eng.* **15** 066011
- Kauppila M, Vanhatalo S and Stevenson N J 2018 Artifact detection in neonatal EEG using Gaussian mixture models *Embec & NBC 2017* (Springer) pp 221–4
- Kobler R J, Shurlea A I, Dias C L, Schwarz A, Mondini V and Müller-Putz G R 2021 Eeg eye artifact dataset (available at: osf.io/2qgrd)
- Kosmyna N, Morris C, Sarawgi U, Nguyen T and Maes P 2019 AttentivU: a wearable pair of EEG and EOG glasses for real-time physiological processing 2019 *IEEE 16th Int. Conf. on Wearable and Implantable Body Sensor Networks (BSN)* (IEEE) pp 1–4
- Kothe C, Shirazi S Y, Stenner T, Medine D, Boulay C, Grivich M I, Artoni F, Mullen T, Delorme A and Makeig S 2025 The lab streaming layer for synchronized multimodal recording *Imaging Neurosci.* **3** IMAG.a.136
- Nair S, James B P and Leung M-F 2025 An optimized hybrid approach to denoising of EEG signals using CNN and LMS filtering *Electronics* **14** 1193
- Palmisciano A C, Farabbi A, Rossi M, Antonello N, Trojaniello D, Cerveri P and Mainardi L 2026 Form factor meets function: anatomy-dependent electrode-skin coupling and signal content in consumer eyewear EEG systems *J. Neural Eng.* **23** 016035
- Peh W Y, Yao Y and Dauwels J 2022 Transformer convolutional neural networks for automated artifact detection in scalp EEG
- Pion-Tonachini L, Kreutz-Delgado K and Makeig S 2019 ICLABEL: an automated electroencephalographic independent component classifier, dataset and website *NeuroImage* **198** 181–97
- Plöchl M, Ossandón J P and König P 2012 Combining EEG and eye tracking: identification, characterization and correction of eye movement artifacts in electroencephalographic data *Front. Hum. Neurosci.* **6** 278
- Saleh M, Xing L and Casson A J 2025 EEG artifact removal at the edge using ai hardware *IEEE Sens. Lett.* **9** 1–4
- Samakovlis D, Albini S, Álvarez R R, Constantinescu D-A, Schiavone P D, Peón-Quirós M and Atienza D 2025 BiomedBench: a benchmark suite of TinyML biomedical applications for low-power wearables *IEEE Des. Test* **42** 45–54
- Simonyan K and Zisserman A 2015 Very deep convolutional networks for large-scale image recognition 3rd *Int. Conf. on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9, May 2015, Conf. Track Proc.*, ed Y Bengio and Y LeCun

- Sun W, Su Y, Wu X and Wu X 2020 A novel end-to-end 1D-ResCNN model to remove artifact from EEG signals *Neurocomputing* **404** 108–21
- Teplan M 2002 Fundamentals of EEG measurement *Meas. Sci. Rev.* **2** 1–11
- Uusitalo M A and Ilmoniemi R J 1997 Signal-space projection method for separating MEG or EEG into components *Med. Biol. Eng. Comput.* **35** 135–40
- Vidaurre C, Krämer N, Blankertz B and Schlögl A 2009 Time domain parameters as a feature for eeg-based brain–computer interfaces *Neural Netw.* **22** 1313–9
- Xing L and Casson A J 2024 Deep autoencoder for real-time single-channel EEG cleaning and its smartphone implementation using tensorflow lite with hardware/software acceleration *IEEE Trans. Biomed. Eng.* **71** 3111–22
- Yu J, Li C, Lou K, Wei C and Liu Q 2022 Embedding decomposition for artifacts removal in EEG signals *J. Neural Eng.* **19** 026052
- Zanetti R, Aminifar A and Atienza D 2025 EEG glasses for real-time brain electrical activity monitoring *Sci. Rep.* **15** 43574
- Zanetti R, Arza A, Aminifar A and Atienza D 2022 Real-time EEG-based cognitive workload monitoring on wearable devices *IEEE Trans. Biomed. Eng.* **69** 265–77
- Zhang H, Wei C, Zhao M, Liu Q and Wu H 2021 A novel convolutional neural network model to remove muscle artifacts from EEG *ICASSP 2021 - 2021 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 1265–9
- Zhang H, Zhao M, Wei C, Mantini D, Li Z and Liu Q 2021 EEGdenoiseNet: a benchmark dataset for deep learning solutions of EEG denoising *J. Neural Eng.* **18** 056057