AI in medicine

MDPI

Article

A Hybrid Anomaly Detection Framework for Reliable Physiological Signal Extraction in Multimodal Wearable Sleep Monitoring

Feiya Xiang [1,†], Geet Khatri [1,†], Alec Brewer [1], Emily Garceau [1], Kirstie M. K. Queener [2,3], Mauro Caballero Victorio [4], Parvez Ahmmed [1], James Reynolds [1], Vladimir Aleksandrovich Pozdin [4], Michael Daniele [1,2,3], Alper Bozkurt [1] and Edgar Lobaton [1,*]

1 Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695, USA; fxiang@ncsu.edu (F.X.); gkhatri@ncsu.edu (G.K.); egarcea@ncsu.edu (E.G.); jlreyno4@ncsu.edu (J.R.); mdaniel6@ncsu.edu (M.D.); aybozkur@ncsu.edu (A.B.)
2 Lampe Joint Department of Biomedical Engineering, North Carolina State University, Raleigh, NC 27695, USA; kmqueene@ncsu.edu
3 Lampe Joint Department of Biomedical Engineering, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
4 Department of Electrical and Computer Engineering, Florida International University, Miami, FL 33174, USA; maucabal@fiu.edu (M.C.V.); vpozdin@fiu.edu (V.A.P.)
* Correspondence: ejlobato@ncsu.edu
† These authors contributed equally to this work.

Abstract

Wearable sleep monitoring systems provide a scalable and low-burden alternative to laboratory-based polysomnography, but overnight physiological recordings collected from wearable sensors are frequently corrupted by poor skin contact, sensor displacement, flatline behavior, saturation, abrupt autoscaling, outliers, and non-physiological noise. These low-quality segments can prevent reliable extraction of clinically relevant biomarkers, including heart rate, heart rate variability, pulse rate, oxygen saturation, and electrodermal activity, thereby limiting the robustness of downstream sleep stage classification and sleep apnea prediction. In this work, we present a multistage hybrid anomaly detection framework for signal quality assessment in a custom-designed multimodal wearable sleep monitoring platform. The proposed framework is shared across modalities: each sensor stream is segmented into windows, screened using signal processing algorithms for obvious sensor failures, and then analyzed using an autoencoder trained on normal windows to detect subtler morphology-level deviations from normal physiological patterns. On chest ECG recordings collected over 10 overnight sessions, the hybrid fusion detector achieves an accuracy of 0.9067, AUROC of 0.9068, F1 score of 0.8739, precision of 0.9470, and recall of 0.8202, outperforming rule-based detection and autoencoder-based detection alone. Additional experiments on PPG, EDA and ExG recordings show consistently high F1 scores of 0.9431, 0.9567 and 0.8728, respectively. These results demonstrate that hybrid anomaly detection can serve as an effective quality-control layer for multimodal wearable sleep monitoring and provide a foundation for future work on robust sleep stage and apnea prediction under real-world sensing conditions.

Keywords: wearable sleep monitoring; anomaly detection; signal quality assessment; ECG; PPG; EDA; autoencoder; physiological signals; sleep stage prediction; sleep apnea

Check for updates

Academic Editor: Nuno M. Garcia

Received: 31 May 2026
Revised: 28 July 2026
Accepted: 7 August 2026
Published: 12 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Polysomnography (PSG) is the clinical reference standard for sleep assessment [1]. A standard sleep study records multiple physiological signals, including electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), respiratory effort, airflow, oxygen saturation, and electrocardiography (ECG), to assess sleep architecture and diagnose sleep disorders such as obstructive sleep apnea. Although PSG provides comprehensive physiological measurements, it is costly and usually requires overnight monitoring in a clinical sleep laboratory. This limits large-scale and long-term deployment, and the unfamiliar laboratory setting may also alter normal sleep behavior.

Wearable sleep monitoring systems provide a practical alternative to traditional PSG by enabling long-term physiological monitoring outside clinical environments [2,3]. However, wearable physiological signals are often noisier than clinical PSG recordings. During overnight use, sensors may shift, skin contact may weaken, and body motion can introduce artifacts or missing segments. In multimodal wearable systems, signal quality can differ substantially across modalities, subjects, recording sessions, and sensor locations [4].

The AI-DReAMS system, short for AI-Driven, Resilient, and Adaptive Monitoring of Sleep, was developed as a reconfigurable multimodal wearable platform for sleep monitoring. The system includes distributed head, chest, and wrist devices with shared environmental sensors and physiological sensing modalities. The head device records bioelectrophotonic signals related to sleep stage physiology, the chest device records cardiac optical and electrical signals, and the wrist device records pulse oximetry and electrodermal activity. The broader goal of this platform is to enable robust and adaptive sleep monitoring through multimodal sensing, sensor fusion, and AI-based interpretation.

A central challenge in this setting is that downstream sleep analytics are only as reliable as the physiological signals used to compute them. Low-quality cardiac signals can corrupt heart rate and heart rate variability features [5–7]. Low-quality photoplethysmography (PPG) or pulse oximetry can corrupt pulse rate and oxygenation estimates [8]. Low-quality electrodermal activity (EDA) can obscure autonomic activity [9]. Device-level anomalies, including battery or memory issues, can compromise the interpretation of entire recording intervals. If these corrupted windows are used directly for sleep stage classification or apnea prediction, they can degrade model performance and reduce trustworthiness.

This manuscript focuses on anomaly detection as a quality control layer for wearable sleep monitoring. The objective is not to perform final sleep stage or apnea prediction in this paper. Instead, the goal is to identify low-quality physiological windows from which reliable biomarkers cannot be extracted. These windows can then be removed, down-weighted, or flagged before downstream feature extraction and prediction. A subsequent study will evaluate how anomaly filtering affects robust sleep stage classification and apnea prediction performance [10–12].

Physiological anomaly detection in wearable sleep data is difficult because artifacts are heterogeneous. Some anomalies, such as flatlines, saturation, abrupt autoscaling, and isolated outliers, are straightforward to detect using signal processing rules. Other anomalies preserve plausible amplitude ranges but distort the underlying morphology or spectral structure. Fully supervised labeling of all artifact types is also impractical for large overnight multimodal datasets. A robust quality control system should therefore combine interpretable signal processing algorithms for common failure modes with data-driven models capable of detecting subtler out-of-distribution patterns [13].

In this work, we propose and evaluate a multistage hybrid anomaly detection framework. Importantly, the anomaly detection pipeline is shared across sensors at the conceptual level. Each sensor stream is segmented into windows; obvious failures are detected using rule-based signal processing criteria; remaining windows are analyzed by an autoencoder

trained on normal windows; and the final decision is obtained by fusing the rule-based and autoencoder-based outputs. The implementation details, such as preprocessing, window length, thresholds, and residual score, are modality-specific, but the pipeline itself is the same across sensors.

The main contributions of this work are:

1. We introduce a unified multistage hybrid anomaly detection framework that applies across sensor modalities using rule-based detection, signal preprocessing, and autoencoder-based residual analysis.
2. We evaluate the framework on chest ECG, PPG, EDA and ExG recordings and show that the hybrid fusion detector substantially outperforms rule-based and autoencoder-based detection alone.
3. We provide a structured foundation for future robust sleep stage and apnea prediction models that can account for unreliable sensor windows.

The rest of the article is organized as follows. Section 2 describes the study protocol, wearable devices, anomaly taxonomy, and semi-automated labeling process. Section 3 presents the proposed hybrid anomaly detection framework, including rule-based screening, autoencoder-based residual analysis, decision fusion, and evaluation design. Section 4 reports anomaly detection results across ECG, PPG, EDA, and ExG recordings. Section 5 discusses the implications, limitations, and future use of anomaly scores for robust sleep stage and apnea prediction. Section 6 concludes the paper.

2. Materials

This section describes the data sources and annotation process used to develop and evaluate the proposed anomaly detection framework. We first summarize the AI-DReAMS wearable platform developed at North Carolina State University, Raleigh, NC, USA and the sleep study recordings used in this work. We then define the sensor anomaly categories and describe the semi-automated labeling procedure used to generate window-level annotations for model training and evaluation.

2.1. Study Protocol

Data were collected as part of the AI-DReAMS sleep monitoring study, which included both home-based and clinical recordings. In the home study phase, 8 participants wore the AI-DReAMS devices overnight in their own homes for two nights. In the clinical phase, data were collected in collaboration with Sleep Works LLC, Raleigh, NC, USA, a sleep clinic in Raleigh.

2.2. Devices for Data Collection

Physiological data were collected using the AI-DReAMS platform, a custom battery-powered multimodal wearable system developed for overnight sleep monitoring [14,15]. The platform consists of three patch devices placed on the head, chest, and wrist. Each device uses a shared main board with location-specific sensing boards, allowing the same core hardware architecture to support different electrophysiological, photonic, and environmental sensors depending on body location.

All three devices include shared contextual sensors, including a three-axis accelerometer for motion and body orientation, a skin temperature sensor, and a microphone for sound-level monitoring. The head device records ExG channels, including two EEG channels, one EOG channel, and one EMG channel. It also records 10 near-infrared spectroscopy (NIRS) channels from the forehead. The chest device records ECG for cardiac electrical activity and green light PPG for blood volume changes. The wrist device records pulse oximetry (POX) using red and infrared LEDs and electrodermal activity (EDA) from the forearm.

Each device stores sensor data locally using onboard flash memory, so continuous Bluetooth Low Energy communication is not required during overnight recordings. Sensor packets include sensor identifiers and timestamps, enabling synchronization across devices and modalities during post-processing. The offloaded data is stored as sensor-specific files and used for labeling, preprocessing, anomaly detection, and downstream analysis.

2.3. Anomaly Taxonomy and Semi-Automated Labeling

Sensor anomalies are categorized according to common failure modes observed in the wearable recordings:

- Flatline: A region with little or no signal variation, potentially caused by disconnection, sensor failure, or loss of contact.
- Autoscaling or wraparound: An abrupt transition from high to low values, or vice versa, often caused by overflow or scaling issues.
- Saturation: A region where the signal reaches or remains near the maximum or minimum representable sensor value.
- Outlier or spike: An isolated sample or short region with signal magnitude far outside the local distribution.
- Noise: A region with low signal-to-noise ratio or non-physiological fluctuations.
- Big jump: A large adjacent sample difference that is inconsistent with local signal variation.
- Other: Anomalies that do not fit the above categories, including complex morphology distortions and sensor failures.

In this study, a *window* refers to a contiguous segment of a single sensor channel used as the basic unit for labels, training, and evaluation. The window duration and hop size are modality-specific because sensors have different sampling rates and physiological time scales. Therefore, anomaly labels are assigned at the single-channel window level rather than jointly across modalities.

Figure 1 illustrates representative examples of the anomaly types. To standardize labeling, anomalous windows were defined based on physiological usability rather than a universal amplitude threshold. A window was labeled as anomalous when corruption prevented reliable extraction of the relevant physiological information. For example, ECG windows were labeled anomalous when QRS/R-peak morphology was no longer reliably identifiable, PPG windows when pulse morphology was obscured, and EDA windows when nonphysiological fluctuations dominated the phasic response [16,17].

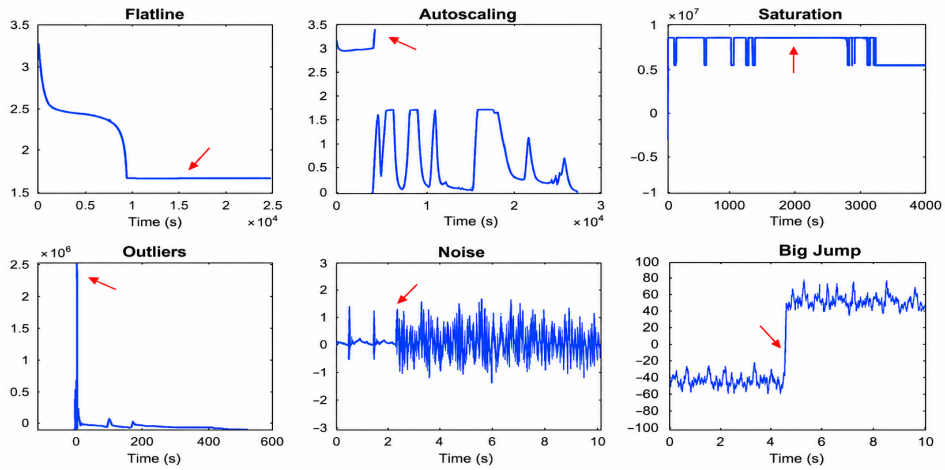


Figure 1. Representative examples of common anomaly types across different physiological modalities. Flatline and autoscaling are shown using wrist EDA signals, saturation and outlier using chest PPG signals, noise using chest ECG, and big jump using head ExG. Red arrows indicate the anomalous regions.

To reduce manual annotation burden, a semiautomated labeling pipeline (Figure 2) was used to generate window-level labels. The automatic labeler first applied signal processing algorithms to identify common artifact types. These automatic labels were used as initial prelabels. The labeled time series were then visually inspected, and missed detections or false detections were manually corrected. The reviewed labels were converted into the final binary window-level annotations used for model training and evaluation.

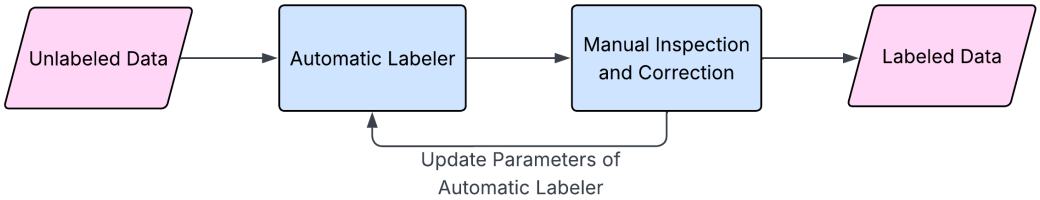


Figure 2. The semi-automated pipeline for labeling anomalies.

The automatic labeling algorithms include flatline detection, saturation detection, autoscaling detection, and outlier detection. Flatline detection identifies regions whose variance remains below a predefined threshold for longer than a minimum duration. Saturation detection uses the same principle but additionally requires the signal level to be close to the sensor's maximum or minimum value. Autoscaling detection identifies consecutive samples whose difference is approximately equal to the full observed signal range. Outlier detection used a Hampel-filter-style local median deviation criterion, which is a standard robust method for detecting impulsive outliers in time-series data [18,19].

Table 1 summarizes the annotated windows used in this study. The number of windows differs across modalities because annotation availability and usable recording duration varied across sensors and study phases. For ExG, the totals for all 4 channels combined are shown. Another reason the ExG counts are much higher than those for other modalities is its small window size (2.048 s) and high sampling rate (250 Hz).

Table 1. Summary of annotated windows used for anomaly detection experiments. ECG, PPG and EDA are single-channel, whereas ExG has 4 channels.

Modality	Recordings/Nights	Total Windows	Normal Windows	Anomalous Windows
ECG	10 nights	6380	3060	3320
PPG	6 nights	3839	2905	934
EDA	9 nights	30,338	15,993	14,345
ExG	7 nights	290,092	166,361	123,731

Table 2 shows the breakdown of annotated windows by type for wrist EDA data. The overwhelming majority of anomalies in EDA data are flatlines. This is important to note because flatlines are different from many other types of anomalies like outliers and big jumps in the sense that they are in the same range as normal data and do not have any high-frequency components. This has implications for the performance of the purely autoencoder-based approach, as discussed further in Section 4.3.

Table 2. Summary of annotated windows for wrist EDA. Note that flatlines represent the vast majority of anomalies in EDA data.

Normal	Flatline Only	Autoscaling Only	Flatline & Autoscaling	Total
15,993	14,308	36	1	30,338

3. Methods

The proposed anomaly detection pipeline is designed to be shared across sensors. The goal is to avoid building a completely separate method for each modality. Instead, all sensor streams follow the same four-step framework:

1. *Windowing:* The continuous signal is divided into fixed-length or overlapping windows.
2. *Rule-based screening:* Obvious sensor failures are detected using interpretable signal processing rules, such as flatline, saturation, autoscaling, spike, big jump, or outlier criteria.
3. *Autoencoder-based residual analysis:* Windows that are not confidently rejected by the rule-based stage are reconstructed using an autoencoder trained on normal windows. Reconstruction residuals are converted into anomaly scores.
4. *Decision fusion:* The final anomaly decision is the logical OR of rule-based detection and autoencoder-based detection.

This unified pipeline is motivated by the complementary failure modes of rule-based and autoencoder-based detection. Rule-based detection is reliable for known artifacts that have clear definitions, such as flatlines and saturation. However, rule-based methods cannot anticipate every possible artifact. Autoencoders can detect windows outside the normal training distribution [20], but they may fail on anomalies that are simple to reconstruct, such as flatlines. As shown in Figure 3, the hybrid decision combines the strengths of both components.

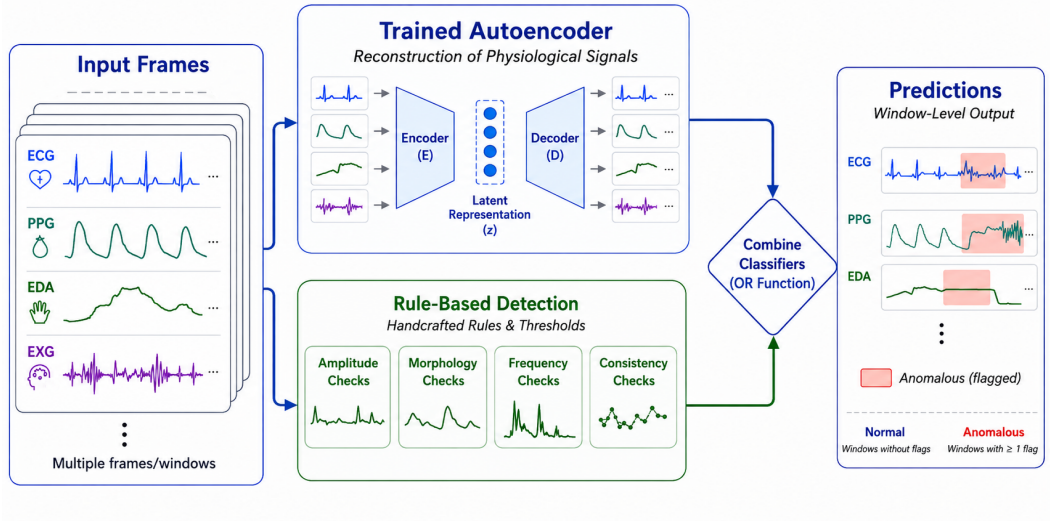


Figure 3. Hybrid anomaly detection framework for multimodal wearable sleep monitoring. Windowed physiological signals are processed by two complementary branches: a rule-based detector for known sensor failures and an autoencoder-based detector for morphology level deviations from normal data. The two outputs are fused using an OR rule to produce window-level anomaly predictions before downstream sleep analysis.

The following sections describe the three main components of the proposed anomaly detection framework. Section 3.1 describes the rule-based screening component, which identifies common sensor failures using signal processing criteria. Section 3.2 describes the autoencoder-based residual analysis component, which detects windows that deviate from normal physiological morphology. Section 3.3 describes the decision fusion strategy used to combine the two components into a final window-level anomaly prediction.

3.1. Rule-Based Screening

The first component of the proposed framework uses interpretable rule-based screening to identify common sensor failures before applying data-driven anomaly detection. This component is designed to capture artifacts that have clear signal processing definitions,

such as flatlines, saturation, spikes, large jumps, low-amplitude windows, and abnormal frequency content. Because these failure modes are often sensor-specific, the rule-based criteria are adapted to each modality while preserving the same overall screening principle.

For ECG, rule-based detection removes obviously invalid windows using peak-to-peak amplitude, standard deviation, local flatness, and low-amplitude block criteria. In addition, ECG windows are screened using a morphology-related frequency ratio. Specifically, the ratio between QRS-band energy and low-frequency energy is computed as

$$R_{\mathrm{ECG}} = \frac{E_{5\text{–}15\ \mathrm{Hz}}}{E_{0.5\text{–}3\ \mathrm{Hz}} + 10^{-12}}, \tag{1}$$

where $E_{5\text{–}15\ \mathrm{Hz}}$ corresponds to the frequency range containing most of the QRS complex energy [21], while $E_{0.5\text{–}3\ \mathrm{Hz}}$ corresponds to the fundamental cardiac rhythm (approximately 30–180 bpm). The band energy is computed as the sum of squared one-sided FFT magnitudes within the corresponding frequency range. Before spectral analysis, each ECG analysis window is mean-centered and transformed using a one-sided real FFT with a rectangular window (i.e., no tapering) and no zero padding. The FFT length equals the analysis window length, resulting in a frequency resolution of f_s/N, where f_s is the sampling frequency and N is the number of samples in the analysis window. Consequently, normal ECG signals exhibit a relatively stable balance between these two bands, whereas motion artifacts or waveform distortion reduce the relative QRS energy and produce abnormally low values of R_{ECG}. Windows with ratios below the threshold estimated from the validation data are therefore flagged as anomalous.

For PPG, rule-based screening focuses on flatline, spike, and big-jump detection. The strict setting used in the current reported experiment uses a spike threshold of 1000.0, a flatline standard-deviation threshold of 0.001, and a big-jump factor of 10.0. Additional threshold sweeps are used to inspect false positives and select operating points. For ECG and PPG, anomaly thresholds are selected from the distribution of scores on normal training windows to target a 1% false positive rate.

For EDA, rule-based screening detects flatlines. As shown in Figure 4, the original EDA signal is first decomposed into tonic and phasic components, following common EDA preprocessing practice for separating slow baseline activity from faster phasic responses [17,22]. The phasic component is used for subsequent autoencoder-based detection because it preserves abrupt artifacts while reducing slow baseline variation.

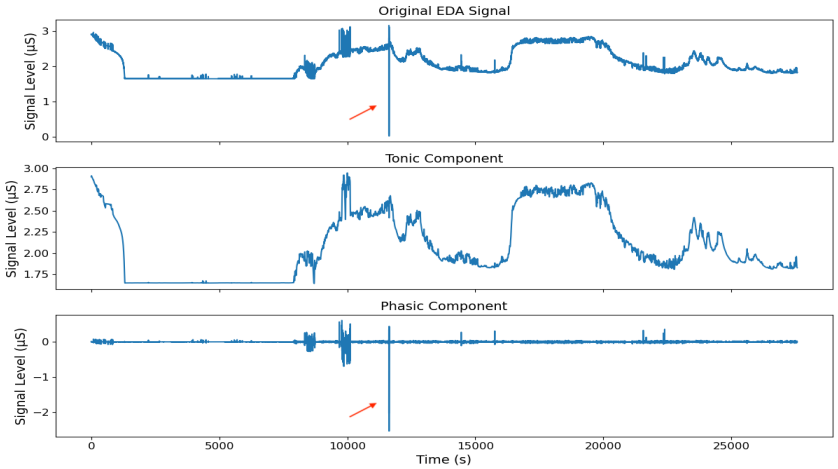


Figure 4. The tonic and phasic components of an EDA signal. The phasic component preserves information about abrupt anomalies such as outliers (shown with a red arrow) and autoscaling, making it suitable for anomaly detection.

For ExG, rule-based screening detects flatlines.

Table 3 summarizes the modality-specific implementation details used within the shared anomaly detection framework.

Table 3. Modality-specific implementation details within the shared anomaly detection pipeline.

Component	**Chest ECG**
Input signal	Chest channel C_ecg
Sampling rate	100 Hz
Preprocessing	Median/MAD robust normalization
Rule screening	Flatline; peak-to-peak amplitude; standard deviation
Frequency rule	QRS-to-low-frequency energy ratio
Autoencoder	1D convolutional autoencoder trained on normal windows only
Fusion rule	Rule-based OR autoencoder-based detection
Component	**Chest PPG**
Input signal	Chest channel C_grn
Sampling rate	100 Hz
Preprocessing	Median/MAD robust normalization
Rule screening	Flatline, spike, and big-jump detection
Frequency rule	Optional morphology screening
Autoencoder	1D convolutional autoencoder trained on normal windows only
Fusion rule	Rule-based OR autoencoder-based detection
Component	**Wrist EDA**
Input signal	Wrist EDA
Sampling rate	18.75 Hz
Preprocessing	Tonic–phasic decomposition (phasic component used for autoencoder); z-score
Rule screening	Flatline
Frequency rule	Not used
Autoencoder	CNN autoencoder trained on normal windows only
Fusion rule	Rule-based OR autoencoder-based detection
Component	**Head ExG**
Input signal	Head ExG
Sampling rate	250 Hz
Preprocessing	Z-score
Rule screening	Flatline
Frequency rule	Not used
Autoencoder	CNN autoencoder trained on normal windows only
Fusion rule	Rule-based OR autoencoder-based detection

3.2. Autoencoder-Based Residual Analysis

Autoencoder-based anomaly detection models learn typical signal patterns from normal data and detect anomalous windows through reconstruction mismatch, without requiring explicit labels for every artifact type [13,23].

Let $x[n]$ denote a single-channel sensor signal. The signal is divided into L frames, each with frame length M, hop size H, and window function $w[n]$. The l-th windowed frame is

$$x_l[n] = x[n + lH]w[n], \quad n \in \{0, \ldots, M-1\}, \quad l \in \{0, \ldots, L-1\}. \tag{2}$$

Each frame has a binary label $y_l \in \{0, 1\}$, where 0 denotes no anomaly and 1 denotes anomaly. A frame was labeled according to the most frequent annotation among its samples. Given an autoencoder reconstruction $\hat{x}_l[n]$, the predicted label is

$$\hat{y}_l = \mathbb{1}\{\epsilon(x_l, \hat{x}_l) > \eta\}, \tag{3}$$

where ϵ is a reconstruction-based anomaly score and η is a decision threshold.

For ECG and PPG, the autoencoder-based component uses a one-dimensional convolutional architecture trained on normal windows only. The ECG autoencoder uses encoder channels

$$1 \to 16 \to 32 \to 64 \to 128 \to 256$$

with a mirrored decoder. It is trained for 100 epochs using Adam [24] with learning rate 10^{-3} and batch size 64. The loss combines raw-domain and derivative-domain Smooth L1 losses:

$$\mathcal{L} = 0.2\,\mathrm{SmoothL1}(\hat{x}, x) + 0.8\,\mathrm{SmoothL1}(\Delta\hat{x}, \Delta x),$$

where $\Delta x_t = x_t - x_{t-1}$. The derivative-domain term is weighted more heavily to preserve local ECG morphology and sharp transitions.

For EDA, the CNN autoencoder is trained on normal phasic EDA frames. The EDA experiments use 9 trials totaling 57.61 h. Frames have length 128 samples, corresponding to approximately 6.83 s, with hop size 16. The input frames are normalized using the z-score. The CNN autoencoder has two encoder layers with filter sizes 32 and 16, respectively, and two decoder layers with filter sizes 16 and 32, respectively. The stride is 2, and the kernel size is 7. The total number of parameters is 9505. The CNN is trained using Adam with learning rate 5×10^{-4}, batch size 256, and 100 epochs with an early stopping patience of 10.

For ExG, the 4 channels are processed independently. The experiments use 7 trials with 41.26 h per channel (165.04 h in total). The frame length is 128 samples (0.512 s), and the hop size is 16. For each channel, the input frames are normalized using the z-score. The CNN has the same architecture and hyperparameters as the one for EDA, except its early stopping patience is 5. This is to reduce training time considering the much larger number of windows, as shown in Table 1.

For some sensors, mean absolute error or mean squared error is sufficient as anomaly scores. For morphology-sensitive physiological signals, additional residual statistics may be more informative, including the 95th percentile absolute residual, derivative-domain residual metrics, input–reconstruction correlation, normalized residual energy, residual entropy, and high-frequency residual fraction. These metrics distinguish normal windows with small reconstruction attenuation from anomalous windows with structural mismatch [25].

3.3. Decision Fusion

The final component combines the rule-based and autoencoder-based outputs into a single window-level anomaly prediction. Let $g(x_l)$ denote the rule-based detector, where $g(x_l) = 1$ indicates that window x_l is flagged by at least one rule-based criterion. Let $\hat{y}_l^{\mathrm{AE}}$ denote the autoencoder-based prediction. The final hybrid prediction is defined as

$$\hat{y}_l^{\mathrm{hybrid}} = \mathbb{1}\{g(x_l) = 1 \vee \epsilon(x_l, \hat{x}_l) > \eta\}. \tag{4}$$

Equivalently, a window is classified as anomalous if it is flagged by either the rule-based detector or the autoencoder-based detector. This OR-based fusion is used because the two components capture complementary failure modes. Rule-based detection is effective for known artifacts with clear signal processing definitions, such as flatlines, saturation, or large jumps. Autoencoder-based detection captures subtler morphology-level deviations from the normal training distribution. Combining the two allows the framework to detect both obvious sensor failures and less structured physiological signal abnormalities.

The same fusion rule is used for ECG, PPG, EDA, and ExG. The modalities differ only in preprocessing, rule thresholds, and residual score selection, while the final decision logic remains shared across sensor streams. This design supports a unified anomaly detection framework for multimodal wearable sleep monitoring while still allowing each modality to use signal-appropriate features and thresholds.

3.4. Evaluation Design

The proposed anomaly detection framework was evaluated using leave-one-subject-out (LOSO) cross-validation. In each fold, all windows from one subject were held out as the test set. From the remaining subjects, approximately 40% of the windows were randomly reserved as a validation set for threshold selection, while only normal windows from the remaining data were used to train the autoencoder.

For each fold, the anomaly detection thresholds were selected using the validation set to achieve a target false positive rate (FPR) of 0.01 on normal validation windows (except for ExG, as further discussed in Section 4.4). This conservative operating point was selected because false positive anomaly decisions would unnecessarily discard usable physiological windows and reduce the temporal coverage available for downstream feature extraction and prediction. Since the proposed detector serves as a signal quality control layer, it is intended to remove high-confidence artifacts while retaining most physiologically usable data. Using a fixed target FPR also provides a consistent criterion for threshold selection across LOSO folds.

After threshold selection, the thresholds were fixed and applied to the held-out test subject. Performance was evaluated using accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUROC). The scalar performance metrics were macro-averaged across all LOSO folds so that each subject contributed equally to the final results. Confusion matrices were obtained by aggregating predictions from all held-out test subjects and then applying row normalization.

In the context of signal quality control, precision reflects the ability to avoid discarding usable windows, whereas recall reflects the ability to prevent corrupted windows from entering downstream analysis. The preferred balance between these objectives may depend on the downstream task; for example, HRV extraction may require stricter cardiac signal quality control than coarse pulse rate estimation, while apnea prediction requires reliable oxygenation and respiratory surrogates.

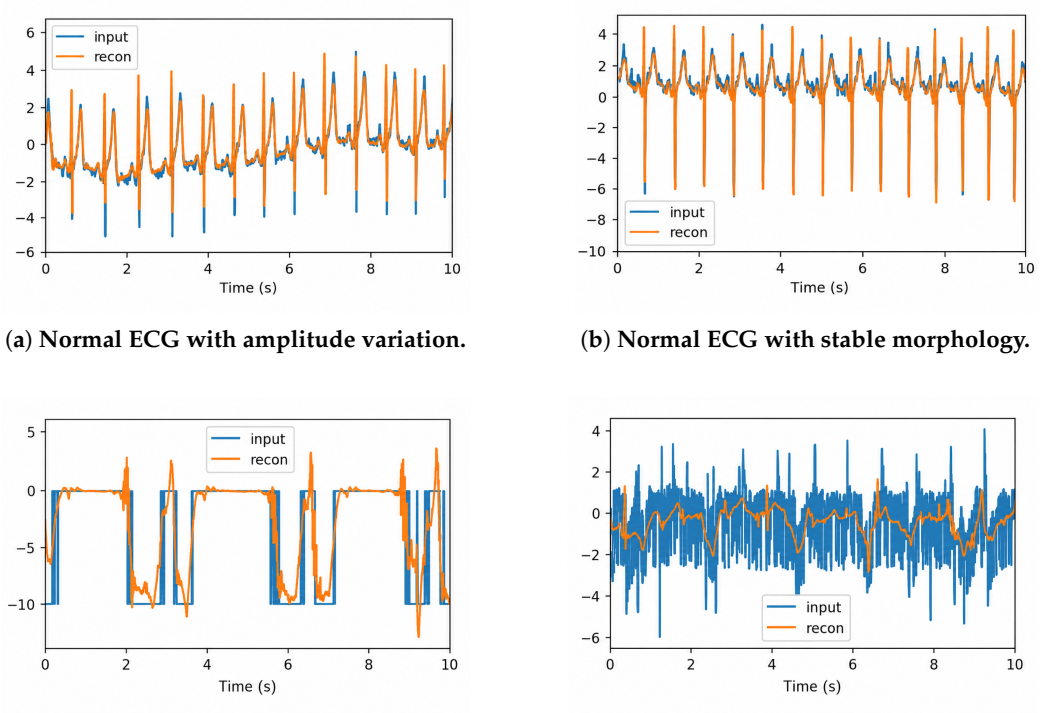
4. Results

This section evaluates the proposed hybrid anomaly detection framework across ECG, PPG, EDA, and ExG recordings. We first present the ECG results in detail to illustrate reconstruction behavior, residual score selection, and hybrid fusion performance. We then report results for PPG, EDA, and ExG, followed by an overall multimodal comparison and an evaluation of the proposed framework across clinical and home recording environments.

4.1. ECG Anomaly Detection Performance

We first evaluated the ECG anomaly detector qualitatively by examining representative reconstruction examples. Figure 5a,b show normal ECG windows that passed the rule-based detection component. For these usable windows, the autoencoder preserved the overall waveform structure. Although sharp peaks were sometimes mildly attenuated, the reconstructed signals remained temporally aligned with the original inputs, indicating that the model learned the dominant morphology of normal ECG windows.

In contrast, anomalous ECG windows that passed the rule-based detection component often exhibited structural mismatch, excessive noise, or morphology inconsistent with the normal training distribution. As shown in Figure 5c,d, these windows were reconstructed poorly by the autoencoder, resulting in lower reconstruction similarity and larger anomaly scores. These qualitative examples support the use of autoencoder-based residual analysis as a complementary component of the proposed pipeline: the rule-based detector removes obvious signal failures, while the autoencoder-based detector identifies subtler windows whose morphology deviates from normal physiological patterns.



(**a**) **Normal ECG with amplitude variation.** (**b**) **Normal ECG with stable morphology.**

(**c**) **Anomalous ECG with saturation.** (**d**) **Anomalous ECG with noise.**

Figure 5. Representative ECG windows before and after autoencoder reconstruction. (**a**,**b**) Normal ECG windows that passed rule-based screening and were reconstructed accurately by the autoencoder despite normal physiological variability. (**c**,**d**) Representative anomalous ECG windows that also passed rule-based screening but exhibit morphology-level deviations, resulting in substantially larger reconstruction errors. The blue curve denotes the original ECG signal, and the orange curve denotes the corresponding autoencoder reconstruction.

To select the autoencoder-based anomaly score, several reconstruction-residual metrics were compared using the validation subset from each LOSO fold. For each fold, AUROC was computed on the validation subset, and the reported values represent the macro-average across all folds.

As shown in Table 4, negative reconstruction correlation achieved the highest AUROC among the evaluated residual scores. This suggests that morphology alignment between the input and reconstruction is more informative than pointwise reconstruction error alone. Therefore, negative reconstruction correlation was selected as the final ECG autoencoder-based anomaly score.

Table 4. AUROC comparison of candidate autoencoder residual scores on validation subset. Bold indicates the best-performing residual score.

Residual Score	AUROC (Mean ± SD)
MSE	0.9214 ± 0.0187
MAE	0.9239 ± 0.0179
Median absolute error	0.9046 ± 0.0213
95th percentile absolute residual	0.9293 ± 0.0165
Negative reconstruction correlation	$\mathbf{0.9430 \pm 0.0142}$
Spectral flatness	0.7148 ± 0.0351
Residual entropy	0.7312 ± 0.0327

To quantify the separability between normal and anomalous ECG windows after autoencoder reconstruction, we analyzed the distribution of the selected residual score and

the negative reconstruction correlation. Figure 6 illustrates the score distribution and ROC curve from one representative LOSO fold for visualization purposes. Normal windows are concentrated near the lower end of the anomaly score range, whereas anomalous windows are shifted toward higher scores. The selected threshold separates the two groups while maintaining the target false positive rate. Similar score distributions were consistently observed across the remaining folds, while the quantitative results reported below are macro-averaged over all LOSO folds.

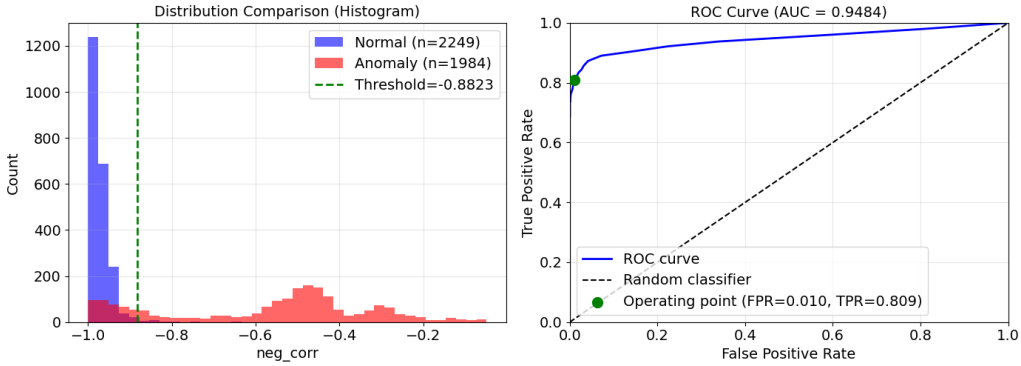


Figure 6. Distribution and ROC analysis of the selected ECG residual score on validation subset.

Following the LOSO evaluation protocol described in Section 3.4, Table 5 reports the macro-averaged performance on the held-out test subjects.

Table 5 compares the rule-based detector, autoencoder-based detector, hybrid fusion detector, and the Orphanidou SQI baseline under the LOSO evaluation protocol. The rule-based detector achieves high precision but relatively low recall, indicating that conservative thresholds effectively identify obvious signal failures while missing many subtle anomalies. The autoencoder-based detector substantially improves recall by capturing morphology-level abnormalities that are difficult to detect using handcrafted rules. By combining both components through OR-based decision fusion, the proposed hybrid detector achieves the best overall balance across the evaluation metrics. We further compared it with the widely used Orphanidou SQI, a representative handcrafted ECG quality assessment algorithm based on signal morphology and beat consistency [26]. Compared with the Orphanidou SQI, the proposed hybrid detector achieves higher accuracy, AUROC, F1 score, and precision, while the SQI provides slightly higher recall. The hybrid detector also requires lower inference latency.

Table 5. ECG anomaly detection performance on held-out test data. Bold values indicate the best performance for each evaluation metric.

Method	Accuracy	AUROC	F1	Precision	Recall	Latency (ms/Window)
Rule-Based Detection	0.6645	0.6780	0.4182	0.8181	0.3359	1.4984
Autoencoder-Based Detection	0.8158	0.7899	0.7529	**0.9716**	0.6405	1.5515
Hybrid Fusion Detector	**0.9067**	**0.9068**	**0.8739**	0.9470	0.8202	3.0773
Orphanidou SQI	0.9011	0.8937	0.8460	0.8314	**0.8638**	4.3930

The confusion matrices in Figure 7 further illustrate the complementary behavior of the three detection strategies. The rule-based detector produces relatively few false positives but misses a considerable number of anomalous windows. The autoencoder-based detector improves sensitivity by identifying morphology-level deviations, while the hybrid fusion detector combines the strengths of both approaches, yielding the most balanced performance across unseen subjects. These results demonstrate that integrating

interpretable rule-based screening with autoencoder-based residual analysis provides a robust and subject-independent solution for ECG signal quality assessment.

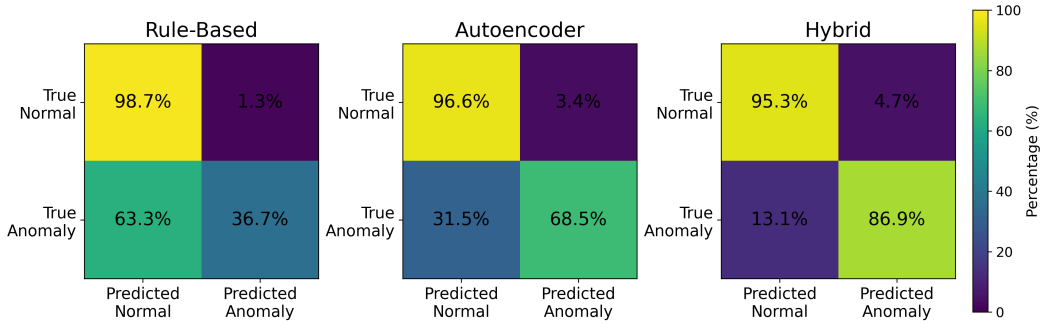


Figure 7. Row-normalized confusion matrices computed from the aggregated predictions of all held-out test subjects.

4.2. PPG Anomaly Detection Performance

We next applied the same hybrid anomaly detection framework to PPG recordings using the LOSO protocol described in Section 3.4. The PPG implementation follows the same overall pipeline, including rule-based detection of spikes, flatlines, and large jumps, autoencoder-based residual analysis, and final decision fusion. Figure 8 illustrates the selected anomaly score distribution from one representative validation fold, whereas Table 6 reports macro-averaged performance across all held-out test subjects.

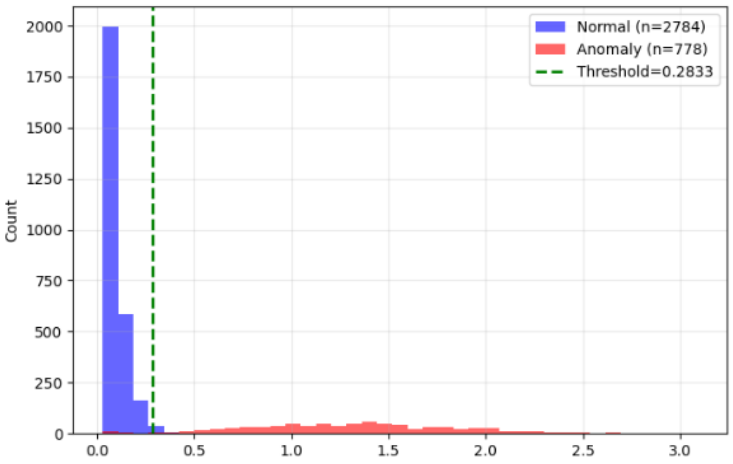


Figure 8. Distribution of the selected PPG anomaly scores for normal and anomalous chest PPG windows. The dashed green line indicates the selected threshold. Normal windows are concentrated below the threshold, whereas anomalous windows are shifted toward larger scores.

Table 6. PPG anomaly detection performance. Bold values indicate the best performance.

Method	Accuracy	AUROC	F1	Precision	Recall	Latency (ms/Window)
Rule-Based Detection	0.8184	0.8037	0.6542	0.7719	0.5685	1.3861
Autoencoder-Based Detection	0.8661	0.8138	0.7843	0.7715	0.7975	1.8126
Hybrid Fusion Detector	**0.9649**	**0.9833**	**0.9431**	**0.9291**	**0.9578**	3.0292
Elgendi SQI	0.8904	0.8522	0.6593	0.6258	0.7113	5.6793

Table 6 shows that the proposed hybrid fusion detector achieves the best overall performance, with an accuracy of 0.9649, AUROC of 0.9833, F1 score of 0.9431, precision of 0.9291, and recall of 0.9578, outperforming individual detectors across all evaluation metrics.

For comparison, we included a conventional morphology-based PPG SQI baseline implemented following the signal quality assessment methodology described by Elgendi [27]. Unlike the proposed hybrid framework, this baseline relies on handcrafted signal quality features without learning-based reconstruction. The hybrid detector achieves better performance and lower inference latency than the SQI baseline. Figure 9 further shows that the autoencoder improves anomaly recall relative to rule-based screening, while hybrid fusion reduces both missed anomalies and false-positive rejections. These preliminary results suggest that the same framework can be extended beyond ECG to optical physiological signals. This is important for wearable sleep monitoring because PPG and pulse oximetry are used to derive pulse rate and oxygenation-related features, which are relevant for downstream sleep stage and apnea analysis.

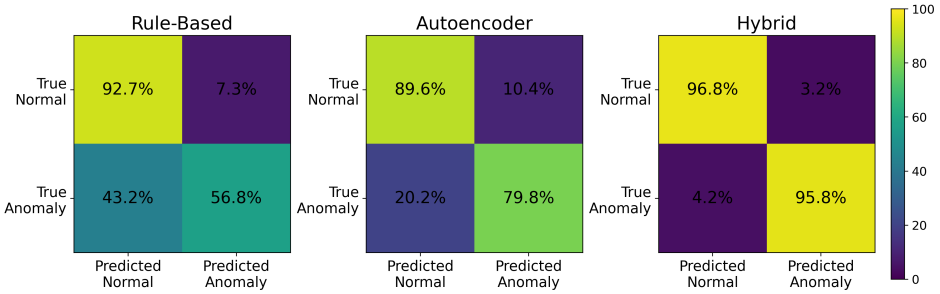


Figure 9. Confusion matrices of the PPG anomaly detection methods, computed from the aggregated predictions of all held-out test subjects.

4.3. EDA Anomaly Detection Performance

EDA experiments further illustrate why hybrid detection is preferable to either rule-based or autoencoder-based detection alone.

Figure 10 shows histograms of autoencoder reconstruction errors for wrist EDA after one training fold. When flatline windows are included in the validation set, anomalous and non-anomalous samples are not clearly separated because flatline segments are simple to reconstruct and therefore produce low reconstruction error. When flatlines are excluded, anomalous windows become much more separable from non-anomalous windows. This result supports the use of rule-based flatline detection together with autoencoder-based detection for other anomaly types.

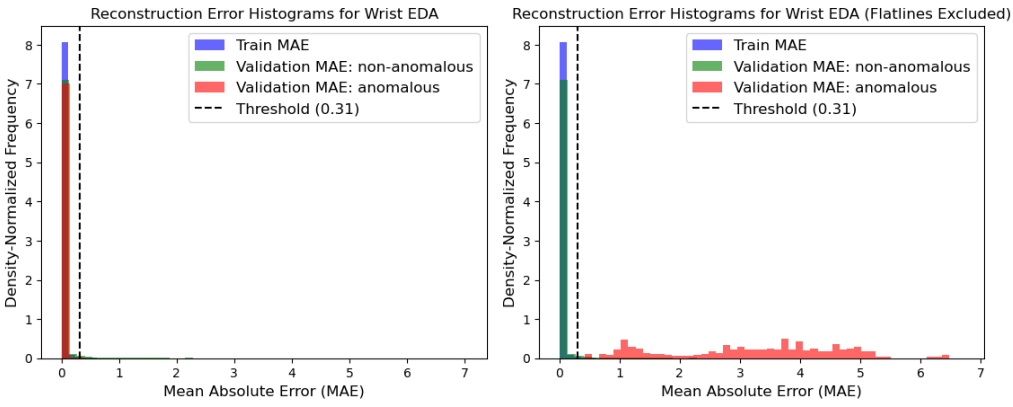


Figure 10. Histograms of autoencoder reconstruction errors for wrist EDA after one training fold. Plot (**a**) includes flatline anomalies in the validation set, while plot (**b**) excludes flatline anomalies.

Table 7 shows the EDA results. Instead of using a simple average across LOSO folds, the accuracy, F1 score, precision and recall were computed from the sum of confusion

matrices for all folds. This is because some of the sessions are imbalanced, i.e., some have very few non-anomalous samples, and some have very few anomalous samples. This causes issues with computation of precision or recall (and, in turn, F1 score) for these sessions. AUROC, on the other hand, was calculated by taking a simple average because each fold has a different threshold. The final confusion matrix for each method is shown in Figure 11.

Table 7. EDA anomaly detection performance aggregated from LOSO folds. Note that the baseline method uses a much larger window, which is part of the reason why its latency is higher. Bold values indicate the best performance for each metric.

Method	Accuracy	AUROC	F1	Precision	Recall	Latency (ms/Window)
Rule-Based Detection	**0.9658**	0.8505	**0.9628**	**0.9921**	0.9351	0.0340
Autoencoder-Based Detection	0.5213	0.3908	0.0170	0.2951	0.0087	0.0346
Hybrid Fusion Detector	0.9596	**0.8884**	0.9567	0.9709	**0.9428**	0.0532
Power Ratio (Baseline)	0.8559	0.6605	0.8565	0.9346	0.7905	0.3224

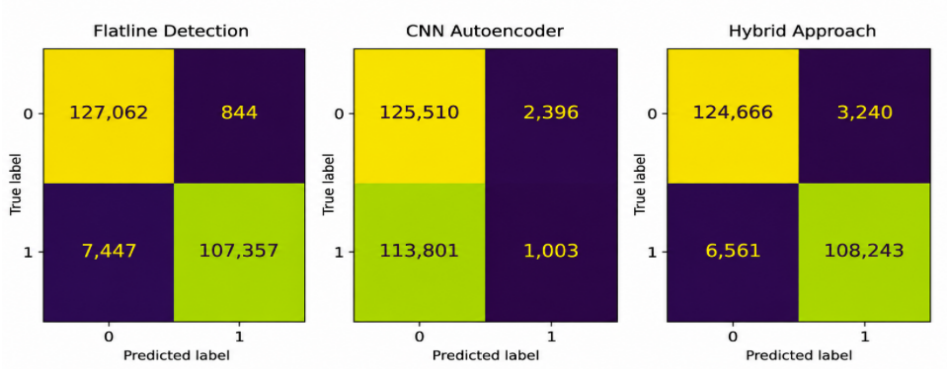


Figure 11. Confusion matrices for the EDA anomaly detection methods, computed from the aggregated predictions of all held-out test subjects.

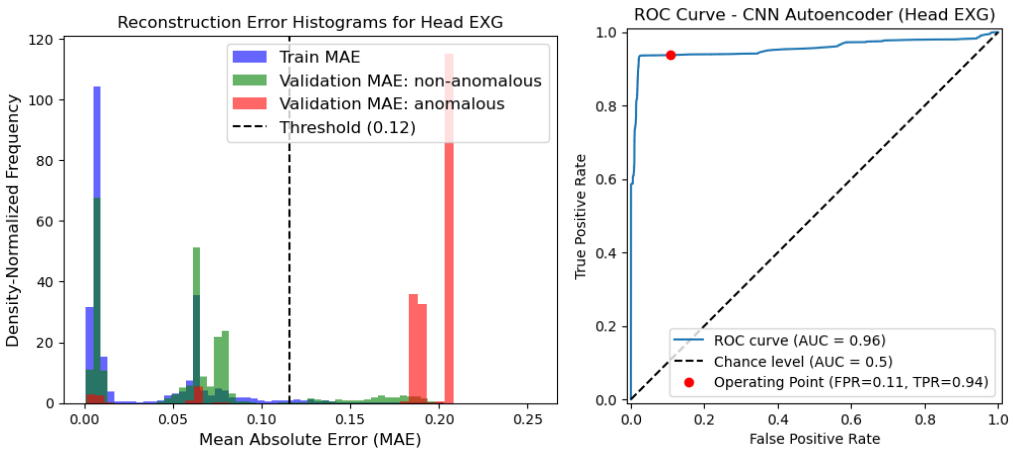
Since flatlines make up the majority of anomalies in EDA data (see Table 2), rule-based screening (flatline detection) alone performs well. The autoencoder on its own has poor performance, as discussed previously. The hybrid fusion detector has the best AUROC and recall, but slightly trails rule-based screening in terms of accuracy, F1 score and precision. Moreover, because the number of non-flatline anomalies is very small (see Table 2), the difference between rule-based screening and hybrid fusion detection is highly sensitive to the performance on the specific non-flatline anomalies.

The baseline method in Table 7 computes the Welch periodogram for each window, and then uses the ratio of the power in the noise band (frequencies > 0.4 Hz) to the total power as the signal quality index [28]. A high frequency resolution is needed to calculate this ratio accurately, so this method requires a much larger window size compared to the one used for the autoencoder-based approach. We used a window size of 8192 samples (436.91 s) for the baseline method. Also, for the baseline, we set the threshold for each fold to the 95th percentile of the training errors as opposed to the 99th percentile for the autoencoder-based approach. This was based on the observation that in terms of the power ratio metric, the normal and anomalous samples in the validation set were not very well separated.

4.4. ExG Anomaly Detection Performance

For ExG, a higher target false positive rate of 0.04 was used based on validation data from one fold for channel 1. Figure 12 shows the error histograms and ROC curve for

autoencoder-based detection for channel 4 data. Both plots in the figure show that normal and anomalous samples are separated in terms of the reconstruction error.



(**a**) (**b**)

Figure 12. Performance of CNN-based autoencoder for channel 4 of head EXG data after one training fold. Plot (**a**) shows the error histograms. Plot (**b**) shows the ROC curve. The chance level shows the performance of a random classifier.

The ExG results for all four channels are reported in Table 8. The hybrid fusion detector has the highest accuracy, F1 score and recall across all four channels. When averaged across channels, the hybrid fusion detector has an accuracy of 0.8888, an F1 score of 0.8728, a precision of 0.8528, and a recall of 0.8962. In terms of the hybrid fusion F1 score, the variation across channels is relatively small (between 0.8556 and 0.8949). The autoencoder on its own achieves a performance close to that of the hybrid approach. Rule-based screening using flatline detection alone does not result in a comparable performance because, unlike EDA, flatlines in ExG data do not make up the majority of anomalies.

Table 8. ExG anomaly detection performance aggregated from LOSO folds. Note that for the first 2 channels (EEG), the baseline method uses a much larger window, which is part of the reason why the average latency is higher. Bold values indicate the best performance for each metric.

Channel	Method	Accuracy	AUROC	F1	Precision	Recall	Latency (ms/Window)
1 (EEG)	Rule-Based	0.8235	0.6653	0.7058	**1.0000**	0.5454	0.0420
	Autoencoder-Based	0.8804	0.6830	**0.8556**	0.8058	0.9119	0.0550
	Hybrid Fusion	**0.8805**	0.6830	**0.8556**	0.8058	**0.9120**	0.0730
	Baseline	0.8184	**0.7157**	0.7652	0.8910	0.6705	0.2410
2 (EEG)	Rule-Based	0.8235	0.7337	0.7432	**1.0000**	0.5914	0.0480
	Autoencoder-Based	0.8784	0.6448	0.8679	0.8176	0.9248	0.0500
	Hybrid Fusion	**0.8787**	0.6448	**0.8683**	0.8177	**0.9256**	0.0690
	Baseline	0.7918	**0.7360**	0.7452	0.9227	0.6250	0.2520
3 (EOG)	Rule-Based	0.8097	0.7295	0.7228	**1.0000**	0.5659	0.0580
	Autoencoder-Based	0.9081	0.7441	0.8946	0.8988	0.8906	0.0490
	Hybrid Fusion	**0.9083**	**0.7442**	**0.8949**	0.8988	**0.8911**	0.0780
	Baseline	0.7075	0.5459	0.6338	0.7029	0.5770	1.1750
4 (EMG)	Rule-Based	0.7932	**0.7198**	0.6994	**1.0000**	0.5377	0.0460
	Autoencoder-Based	0.8876	0.7054	0.8720	0.8890	0.8556	0.0500
	Hybrid Fusion	**0.8879**	0.7057	**0.8723**	0.8891	**0.8562**	0.0720
	Baseline	0.7412	0.6338	0.6683	0.7826	0.5831	0.6110

The baseline uses different metrics for different modalities in the ExG device. For channels 1 and 2 (EEG), a similar metric as the one in Section 4.3 is used, except that the ratio compares the power in the 0.5–80 Hz band to the total power. The reason for choosing this band is that much of the information in EEG signals, including gamma waves, is in this band [29]. This metric is also similar to the ones used in [30]. To calculate this metric accurately, a high frequency resolution is needed, so a longer window of length 4096 samples (16.384 s) is used. For channel 3 (EOG), the signal is first cleaned using a bandpass filter with range 0.25–7.5 Hz, based on the default EOG cleaning method in the NeuroKit2 toolkit [31]. Then, the signal-to-noise ratio between the cleaned signal and the residual noise signal is computed. The metric for channel 4 (EMG) is similar, except that the signal is cleaned using a high-pass filter with a cutoff frequency of 100 Hz, based on the default EMG cleaning method in the NeuroKit2 toolkit [31]. For channels 3 and 4, the window size is the same as that for the autoencoder-based approach.

4.5. Multimodal Summary

Table 9 summarizes the anomaly detection results across the four evaluated physiological modalities. Although ECG, PPG, EDA and ExG differ in morphology, sampling rate, and artifact characteristics, they are processed using the same conceptual framework: rule-based detection for known sensor failures, autoencoder-based residual analysis for subtler deviations, and hybrid decision fusion.

The hybrid fusion framework performs best (or close to best) across the evaluated modalities, with F1 scores of 0.8739 for ECG, 0.9431 for PPG, 0.9567 for EDA, and 0.8728 (average) for ExG. For ECG, PPG, EDA, and ExG, the hybrid detector achieves the highest recall among the compared methods, indicating that fusion improves the ability to identify corrupted windows at the selected operating points. These improvements indicate that rule-based screening and reconstruction-based anomaly detection capture complementary characteristics of signal degradation. Their combination consistently provides a better balance between precision and recall than either individual detector.

Table 9. Anomaly detection performance across ECG, PPG, EDA, and ExG. For ExG, the results are averaged across 4 channels. Performance metrics are macro-averaged across LOSO folds. Bold values indicate the best performance within each modality.

Modality	Method	Accuracy	F1	Precision	Recall
ECG	Rule-Based Detection	0.6645	0.4182	0.8181	0.3359
	Autoencoder-Based Detection	0.8158	0.7529	**0.9716**	0.6405
	Hybrid Fusion Detector	**0.9067**	**0.8739**	0.9470	**0.8202**
PPG	Rule-Based Detection	0.8184	0.6542	0.7719	0.5685
	Autoencoder-Based Detection	0.8661	0.7843	0.7715	0.7975
	Hybrid Fusion Detector	**0.9649**	**0.9431**	**0.9291**	**0.9578**
EDA	Rule-Based Detection	**0.9658**	**0.9628**	**0.9921**	0.9351
	Autoencoder-Based Detection	0.5213	0.0170	0.2951	0.0087
	Hybrid Fusion Detector	0.9596	0.9567	0.9709	**0.9428**
ExG	Rule-Based Detection	0.8125	0.7178	**1.0000**	0.5601
	Autoencoder-Based Detection	0.8886	0.8726	0.8528	0.8957
	Hybrid Fusion Detector	**0.8888**	**0.8728**	0.8528	**0.8962**

To further evaluate robustness under different acquisition environments, Table 10 compares the performance of the proposed hybrid detector on clinical and home recordings. For ECG, the detector achieves an F1 score of 0.9158 in the clinical setting and 0.8927 in the home setting. Clinical recordings provide higher precision (0.9813 versus 0.8473), whereas

home recordings achieve higher recall (0.9433 versus 0.8586), indicating a trade-off between precision and sensitivity under more challenging real-world acquisition conditions. A similar trend is observed for PPG. The hybrid detector achieves an F1 score of 0.9633 for clinical recordings and 0.8825 for home recordings, with clinical recordings yielding higher accuracy (0.9777 versus 0.9250) and precision (0.9671 versus 0.8222), while home recordings maintain a comparable recall (0.9522 versus 0.9595). For these three modalities, the proposed framework maintains high anomaly recall across both clinical and home recordings. The lower precision and F1 scores observed in the home environment indicate an increased false positive burden under less controlled acquisition conditions.

Table 10. Clinical versus home performance of the proposed Hybrid Fusion Detector across different sensing modalities. ExG results are shown for channel 1 only. Bold values indicate the better performance between the two environments for each modality.

Modality	Environment	Positive Rate	Accuracy	F1	Precision	Recall
ECG	Clinical	0.5224	**0.9065**	**0.9158**	**0.9813**	0.8586
	Home	0.5093	0.8913	0.8927	0.8473	**0.9433**
PPG	Clinical	0.3053	**0.9777**	**0.9633**	**0.9671**	**0.9595**
	Home	0.2957	0.9250	0.8825	0.8222	0.9522
EDA	Clinical	0.3430	0.9085	0.8583	0.9152	0.8081
	Home	0.5557	**0.9823**	**0.9843**	**0.9737**	**0.9950**
ExG	Clinical	0.0322	0.5764	0.0424	0.0229	0.2912
	Home	0.5852	**0.6725**	**0.7801**	**0.6426**	**0.9925**

For EDA, the trend is reversed, as the hybrid detector achieves a better F1 score for home recordings (0.9843) compared to clinic recordings (0.8583). This is likely because, in the labeled EDA dataset, there are six home sessions and three clinic sessions, so there is more training data available for the model trained on home recordings. For ExG, the clinical performance is significantly worse for the clinic recordings. This is because the anomalous windows represent a very small fraction of all windows (0.0322), i.e., the clinic dataset is highly imbalanced. Also, there are fewer clinic sessions in the ExG dataset (3) compared to home sessions (4).

Overall, these results show that combining rule-based detection with autoencoder-based residual analysis provides a robust quality control strategy across heterogeneous wearable sensor streams.

5. Discussion

This study presents a unified multistage hybrid anomaly detection framework for reliable extraction of physiological signals in multimodal wearable sleep monitoring. The key design choice is that the pipeline is shared across sensors rather than written as separate independent detectors. Each modality follows the same conceptual structure: window the signal, detect known failures using interpretable rules, detect subtler deviations using autoencoder-based residual analysis, and combine both decisions.

The chest ECG results show that rule-based detection alone is not sufficient. It achieves high precision but low recall, indicating that it removes obvious corrupted windows but misses many subtle morphology-level anomalies. Autoencoder-based residual analysis improves recall and AUROC, and the hybrid fusion detector achieves the strongest overall performance. The qualitative reconstruction examples explain this result: normal windows maintain input–reconstruction similarity, while anomalous windows exhibit structural divergence that is captured by residual metrics.

The EDA results reveal a complementary failure mode. Autoencoder-based detection performs poorly when flatline anomalies are included because flatlines are simple to reconstruct. In this case, rule-based flatline detection is essential. Once flatlines are handled algorithmically, autoencoder-based residual analysis becomes more useful for other anomaly types. This explains why the hybrid EDA detector achieves the best F1 score.

Together, the ECG, PPG, EDA, and ExG results support the same conclusion: no single anomaly detector is sufficient for wearable sleep data. Rule-based algorithms are effective for known sensor failures, while autoencoders are useful for detecting deviations from learned normal morphology. The hybrid design is therefore better aligned with the heterogeneous artifact structure of real overnight wearable recordings.

In practical wearable deployments, anomaly scores may be used not only as binary exclusion masks but also as reliability indicators for multimodal fusion. This is particularly important when one modality is temporarily corrupted while other sensors remain usable. Future work may further incorporate uncertainty-aware anomaly detection to distinguish clear signal corruption from ambiguous or noisy physiological patterns [32,33].

In addition, this study represents an initial retrospective validation using the currently available AI-DReAMS dataset; future work will evaluate the proposed framework on larger cohorts and external validation data to further assess its generalizability.

This work has several limitations. First, the current experiments evaluate modality-specific anomaly detection but do not yet quantify the downstream impact on sleep stage classification or apnea prediction. Second, although the pipeline is conceptually shared across sensors, each modality still requires signal preprocessing and threshold selection.

6. Conclusions

This work introduced a unified hybrid anomaly detection framework for identifying low-quality physiological windows in multimodal wearable sleep recordings. The framework combines interpretable signal processing rules for common sensor failures with autoencoder-based residual analysis for subtler deviations from normal physiological morphology. Across ECG, PPG, EDA, and ExG recordings, the hybrid fusion detector achieved strong performance while preserving a shared pipeline with modality-specific preprocessing, thresholds, and residual scores. These findings support anomaly detection as a quality control layer for robust wearable sleep analytics and future sleep stage and apnea prediction studies.

Author Contributions: Conceptualization, F.X., G.K. and E.L.; methodology, F.X., G.K. and E.L.; software, F.X. and G.K.; validation, F.X. and G.K.; formal analysis, F.X. and G.K.; data curation, A.B. (Alec Brewer), E.G., K.M.K.Q., M.C.V., P.A., J.R. and V.A.P.; writing—original draft preparation, F.X. and G.K.; writing—review and editing, all authors; supervision, E.L., M.D., A.B. (Alper Bozkurt) and V.A.P.; project administration, E.L., M.D. and A.B. (Alper Bozkurt); funding acquisition, E.L., M.D. and A.B. (Alper Bozkurt). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the United National Science Foundation (NSF) grant numbers IIS-2037328, ECCS-2124002, ECC-1160483 (ERC for ASSIST) and in part by the National Institute of Health (NIH), National Cancer Institute (NCI) under Grant 1R01CA297854-01.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of North Carolina State University (protocol 25393, approved on 15 September 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are not publicly available due to privacy and ethical restrictions associated with human-subject physiological recordings. De-identified

derived results may be available from the corresponding author upon reasonable request and subject to institutional approval. The source code will be made publicly available upon publication of this manuscript.

Acknowledgments: The authors thank the study participants and collaborators involved in the AI-DReAMS data collection.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI-DReAMS	AI-Driven, Resilient, and Adaptive Monitoring of Sleep
AE	Autoencoder
AUROC	Area Under the Receiver Operating Characteristic Curve
ECG	Electrocardiography
EDA	Electrodermal Activity
EEG	Electroencephalography
EOG	Electrooculography
EMG	Electromyography
ExG	Electrophysiological Signal
HRV	Heart Rate Variability
NIRS	Near-Infrared Spectroscopy
PPG	Photoplethysmography
POX	Pulse Oximetry
PSG	Polysomnography

References

1. Markun, L.C.; Sampat, A. Clinician-Focused Overview and Developments in Polysomnography. *Curr. Sleep Med. Rep.* **2020**, *6*, 309–321. [CrossRef] [PubMed]
2. Imtiaz, S.A. A Systematic Review of Sensing Technologies for Wearable Sleep Staging. *Sensors* **2021**, *21*, 1562. [CrossRef] [PubMed]
3. Kwon, S.; Kim, H.; Yeo, W.H. Recent advances in wearable sensors and portable electronics for sleep monitoring. *iScience* **2021**, *24*, 102461. [CrossRef] [PubMed]
4. Retamales, G.; Gavidia, M.E.; Bausch, B.; Montanari, A.N.; Husch, A.; Goncalves, J. Towards Automatic Home-Based Sleep Apnea Estimation Using Deep Learning. *npj Digit. Med.* **2024**, *7*, 144. [CrossRef] [PubMed]
5. Chan, J.; Miller, A.C.; Fox, E.B. Representing and Denoising Wearable ECG Recordings. *arXiv* **2020**, arXiv:2012.00110. [CrossRef]
6. Tan, H.; Lai, J.; Liu, Y.; Song, Y.; Wang, J.; Chen, M.; Yan, Y.; Zhong, L.; Feng, Q.; Yang, W. Neural architecture search for real-time quality assessment of wearable multi-lead ECG on mobile devices. *Biomed. Signal Process. Control* **2022**, *74*, 103495. [CrossRef]
7. Mondal, A.; Manikandan, M.S.; Pachori, R.B. Automatic ECG Signal Quality Assessment Using Convolutional Neural Networks and Derivative ECG Signal for False Alarm Reduction in Wearable Vital Signs Monitoring Devices. *Biomed. Signal Process. Control* **2025**, *108*, 107876. [CrossRef]
8. Gao, R.; Miller, C.S.; Lin, B.T.W.; Schwarz, C.W.; Jones, M.L.H. Signal Quality Assessment and Reconstruction of PPG-Derived Signals for Heart Rate and Variability Estimation in In-Vehicle Applications: A Comparative Review and Empirical Validation. *Sensors* **2025**, *25*, 7556. [CrossRef] [PubMed]
9. Kong, Y.; Hossain, M.B.; Peitzsch, A.; Posada-Quintero, H.F.; Chon, K.H. Automatic Motion Artifact Detection in Electrodermal Activity Signals Using 1D U-Net Architecture. *Comput. Biol. Med.* **2024**, *182*, 109139. [CrossRef] [PubMed]
10. Phan, H.; Mikkelsen, K. Automatic sleep staging of EEG signals: Recent development, challenges, and future directions. *Physiol. Meas.* **2022**, *43*, 04TR01. [CrossRef] [PubMed]
11. Soleimani, R.; Barahona, J.; Chen, Y.; Bozkurt, A.; Daniele, M.; Pozdin, V.; Lobaton, E. Advances in Modeling and Interpretability of Deep Neural Sleep Staging: A Systematic Review. *Physiologia* **2023**, *4*, 1–42. [CrossRef]
12. Abd-alrazaq, A.; Aslam, H.; AlSaad, R.; Alsahli, M.; Ahmed, A.; Damseh, R.; Aziz, S.; Sheikh, J. Detection of Sleep Apnea Using Wearable AI: Systematic Review and Meta-Analysis. *J. Med. Internet Res.* **2024**, *26*, e58187. [CrossRef]
13. Choi, K.; Yi, J.; Park, C.; Yoon, S. Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines. *IEEE Access* **2021**, *9*, 120043–120065. [CrossRef]

14. Queener, K.M.K.; Brewer, A.; Sun, H.; Twiddy, J.; Pozdin, V.A.; Sode, K.; Bozkurt, A.; Daniele, M. Self-Adhesive Conductive Elastomers for Gel-Free Biopotential Recording. *Adv. Electron. Mater.* **2026**, *12*, e00004. [CrossRef]
15. Queener, K.M.K.; Ahmmed, P.; Victorio, M.; Twiddy, J.; Dehn, A.; Brewer, A.; Lobaton, E.J.; Bozkurt, A.; Pozdin, V.A.; Daniele, M. Conformal Micropatterned Organic-Metal Electrodes for Physiological Recording. In Proceedings of the 2023 IEEE SENSORS, Vienna, Austria, 29 October–1 November 2023. [CrossRef]
16. Li, Q.; Mark, R.G.; Clifford, G.D. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter. *Physiol. Meas.* **2008**, *29*, 15–32. [CrossRef] [PubMed]
17. Taylor, S.; Jaques, N.; Chen, W.; Fedor, S.; Sano, A.; Picard, R. Automatic Identification of Artifacts in Electrodermal Activity Data. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Milan, Italy, 25–29 August 2015; pp. 1934–1937. [CrossRef] [PubMed]
18. Hampel, F.R. The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.* **1974**, *69*, 383–393. [CrossRef]
19. Pearson, R.K.; Neuvo, Y.; Astola, J.; Gabbouj, M. Generalized Hampel filters. *EURASIP J. Adv. Signal Process.* **2016**, *2016*, 87. [CrossRef]
20. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **2009**, *41*, 15. [CrossRef]
21. Pan, J.; Tompkins, W.J. A real-time QRS detection algorithm. *IEEE Trans. Biomed. Eng.* **1985**, *32*, 230–236. [CrossRef] [PubMed]
22. Cacioppo, J.T.; Tassinary, L.G.; Berntson, G. *Handbook of Psychophysiology*; Cambridge University Press: Cambridge, UK, 2007.
23. Nawaz, A.; Khan, S.S.; Ahmad, A. Ensemble of Autoencoders for Anomaly Detection in Biomedical Data: A Narrative Review. *IEEE Access* **2024**, *12*, 17273–17289. [CrossRef]
24. Kingma, D.P.; Lei, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
25. Rassam, M.A. Autoencoder-Based Neural Network Model for Anomaly Detection in Wireless Body Area Networks. *IoT* **2024**, *5*, 852–870. [CrossRef]
26. Orphanidou, C.; Bonnici, T.; Charlton, P.H.; Clifton, D.; Vallance, D.; Tarassenko, L. Signal-quality indices for the electrocardiogram and photoplethysmogram. *Physiol. Meas.* **2015**, *36*, 1741–1764.
27. Elgendi, M. Optimal Signal Quality Index for Photoplethysmogram Signals. *Bioengineering* **2016**, *3*, 21. [CrossRef] [PubMed]
28. Hossain, M.B.; Kong, Y.; Posada-Quintero, H.F.; Chon, K.H. Comparison of electrodermal activity from multiple body locations based on standard EDA indices' quality and robustness against motion artifact. *Sensors* **2022**, *22*, 3177. [CrossRef] [PubMed]
29. Nayak, C.S.; Anilkumar, A.C. Normal EEG Waveforms. In *StatPearls [Internet]*; StatPearls Publishing: Saint Petersburg, FL, USA, 2025.
30. Radüntz, T. Signal quality evaluation of emerging EEG devices. *Front. Physiol.* **2018**, *9*, 98. [CrossRef] [PubMed]
31. Makowski, D.; Pham, T.; Lau, Z.J.; Brammer, J.C.; Lespinasse, F.; Pham, H.; Schölzel, C.; Chen, S.H.A. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behav. Res. Methods* **2021**, *53*, 1689–1696. [CrossRef] [PubMed]
32. Yong, B.X.; Brintrup, A. Bayesian autoencoders with uncertainty quantification: Towards trustworthy anomaly detection. *Expert Syst. Appl.* **2022**, *209*, 118196. [CrossRef]
33. Wiessner, P.; Bezirganyan, G.; Sellami, S.; Chbeir, R.; Bungartz, H.J. Uncertainty-Aware Time Series Anomaly Detection. *Future Internet* **2024**, *16*, 403. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.