Explainable Emotion Recognition with EEG Wearable Devices

Tommaso Colafiglio[1], Angela Lombardi[1,*], Paolo Sorino[1], Domenico Lofù[1] and Tommaso Di Noia[1]

[1]Department of Electrical and Information Engineering, Politecnico di Bari, Bari, BA, 70125, Italy

Abstract

Emotion recognition from electroencephalography (EEG) signals is a key task in affective computing, but the interpretability of machine learning models remains a major challenge, especially in wearable low-density setups. This work presents an explainable pipeline for EEG-based emotion recognition using a 14-channel Emotiv headset under a DEAP-inspired protocol. The framework combines handcrafted spectral and nonlinear features with machine learning classifiers and SHAP-based explanations to analyze the neural factors underlying emotion decoding. Results reveal consistent temporal patterns across subjects, interpretable spatial attribution maps, and a strong contribution of nonlinear complexity descriptors. The findings are consistent with previous analyses on the NeuroSense dataset, suggesting that meaningful EEG markers of emotion recognition remain observable across different low-cost acquisition setups.

Keywords

EEG emotion recognition, Explainable AI, Wearable EEG, SHAP, Affective computing

1. Introduction

Emotion recognition from electroencephalography (EEG) signals has become an important research topic in affective computing due to the ability of neural signals to capture implicit responses associated with emotional processing [1]. EEG-based emotion recognition systems enable the development of adaptive human–computer interfaces and provide insights into the neural dynamics underlying affective states [2]. Several benchmark datasets have been introduced to support research in this area, including DEAP [3], SEED [4], and DREAMER [5], which demonstrated that emotional dimensions such as valence and arousal can be decoded from EEG signals using machine learning approaches.

In recent years, there has been increasing interest in the use of wearable and low-cost EEG devices for affective computing applications [6]. Compared to traditional high-density laboratory EEG systems, wearable headsets offer fewer electrodes but enable easier deployment, lower acquisition costs, and the potential to collect neural data in more realistic environments. Several studies have explored emotion recognition using such devices, highlighting both the opportunities and the methodological challenges introduced by reduced spatial resolution and higher signal variability [7].

Alongside predictive performance, an additional challenge concerns the interpretability of machine learning models used for EEG decoding. Modern classification approaches can achieve high accuracy but often provide limited insight into the neural features driving their predictions [8]. This has motivated the adoption of Explainable Artificial Intelligence (XAI) techniques in EEG analysis in order to investigate which signal characteristics, spatial regions, and feature families contribute most strongly to model decisions. Among these approaches, feature attribution methods such as SHapley Additive exPlanations (SHAP) provide a theoretically grounded framework for quantifying the contribution of individual input features to model predictions [9, 10]. When applied to EEG-based emotion recognition, such techniques enable researchers to relate classifier decisions to interpretable neurophysiological patterns,

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

*Corresponding author.
angela.lombardi@poliba.it (A. Lombardi)

including the relevance of specific electrodes, spectral components, or nonlinear descriptors of signal dynamics [11].

Previous work introduced the NeuroSense dataset, which investigates emotion recognition using sparse and low-cost EEG acquisition devices [12, 11]. In that study, explainability analysis revealed consistent spatial and feature-level attribution patterns, suggesting that certain nonlinear signal descriptors and specific scalp regions play a dominant role in emotion decoding.

An open question is whether similar explainability patterns emerge when the analysis is performed on EEG signals acquired with a different wearable headset and acquisition setup. Since wearable devices differ in electrode configuration, signal quality, and hardware characteristics, verifying the reproducibility of attribution patterns is important for assessing the robustness and interpretability of EEG-based emotion recognition models.

In this work, we address this question by analyzing a new EEG dataset acquired using a 14-channel Emotiv headset under a DEAP-inspired affective stimulation protocol. The study follows a feature-based machine learning pipeline combined with SHAP-based explanation analysis in order to investigate the spatial and feature-level contributions underlying emotion decoding. By repeating a similar explainability analysis on data collected with a different wearable EEG device, the study aims to verify whether interpretable patterns observed in previous analyses remain observable under a different acquisition setup. Within this framework, the study addresses the following research questions:

- **RQ1:** How stable are the temporal patterns of emotion-related EEG activity across subjects?
- **RQ2:** Which spatial regions contribute most strongly to emotion decoding in a 14-channel wearable montage?
- **RQ3:** Which feature families (spectral power vs. nonlinear complexity measures) dominate the model explanations?

2. Dataset and protocol

EEG data were collected from 30 healthy participants using a 14-channel Emotiv EEG headset under an affective stimulation protocol inspired by the DEAP experimental design [13]. The recordings were acquired during the same experimental campaign as the NeuroSense dataset and followed the same experimental and ethical protocol approved by the Local Ethical Committee of the University of Bari. In particular, during the experiment, participants viewed a sequence of emotionally evocative audiovisual clips while their EEG activity was recorded. After each stimulus presentation, participants provided self-assessments of their emotional state using the Self-Assessment Manikin (SAM) questionnaire, a widely used non-verbal scale that measures emotional responses along three dimensions: *valence* (pleasantness), *arousal* (intensity of the emotional response), and *dominance* [14]. In line with common practice in affective computing, the analysis focuses on the valence–arousal space defined by Russell's circumplex model of emotion, where emotional states are represented as points in a two-dimensional plane defined by valence and arousal [15]. Continuous ratings were subsequently discretized into classes corresponding to the four quadrants of the circumplex model. The stimulus set was constructed through a structured multi-stage pipeline to ensure emotional validity and balance across the four quadrants of the valence–arousal space. Initially, 120 music videos were selected through a combination of affective tag-based retrieval and manual curation. One-minute emotionally salient highlights were then identified using a methodology aligned with the DEAP stimulus selection procedure. Both video (e.g., motion intensity, color variance, lighting dynamics) and audio features (e.g., MFCCs, energy, pitch, and time–frequency descriptors) were extracted to characterize the emotional content of candidate segments. A Relevance Vector Machine (RVM) model was used to estimate valence and arousal scores for overlapping segments, enabling the identification of the most emotionally informative highlights. Finally, an online subjective evaluation was conducted in order to validate the emotional consistency of the selected clips. Based on these ratings, a final set of 40 stimuli eliciting the strongest and most consistent emotional responses was retained for the experiment. The selected stimuli were balanced

Step	Details
Bad-channel detection	Automatic detection based on amplitude and spatial criteria.
Channel interpolation	Reconstruction of bad channels by spatial interpolation.
High-pass filtering (0.5 Hz)	Removal of slow drifts and DC offsets.
ATAR	Wavelet-based artifact mitigation.
Epoching	Segmentation into 5 s stimulus-locked epochs.
Low-pass filtering (40 Hz)	Reduction of high-frequency noise.
Ocular artifact removal	Regression-based correction using AF3/AF4 as EOG proxies.

Table 1
Preprocessing pipeline for 14-channel recordings.

across the four emotional quadrants of Russell's circumplex model, corresponding to the emotional states of excited, angry, relaxed, and sad.

3. Methodology

The objective of this study is to analyze how neurophysiological EEG features contribute to emotion recognition models and to characterize the spatial and feature-level patterns underlying model predictions. Following preprocessing, feature extraction is performed at the epoch $\times$ channel level in order to preserve spatial information across the 14-channel montage.

3.1. Preprocessing and feature extraction

Raw EEG recordings were preprocessed to improve signal quality and attenuate non-neural artifacts prior to feature computation. Noisy sensors were automatically identified as bad channels based on amplitude thresholds and spatial consistency criteria and subsequently reconstructed via interpolation. The continuous signal was high-pass filtered at 0.5 Hz to remove slow drifts and DC offsets. Transient and nonstationary artifacts were further attenuated using the wavelet-based Artifact Reduction algorithm (ATAR)[1]. The cleaned data were segmented into 5 s stimulus-locked epochs and low-pass filtered at 40 Hz to suppress high-frequency noise components. Ocular contamination was mitigated using a regression-based correction strategy, employing AF3 and AF4 channels as Electrooculogram (EOG) proxies. The complete preprocessing workflow is summarized in Table 1. Features are computed at the epoch $\times$ channel level and subsequently concatenated into subject-wise feature vectors. This representation preserves spatial information across electrodes while maintaining consistent dimensionality across subjects. Both nonlinear and spectral descriptors were extracted in order to capture complementary aspects of EEG dynamics. Nonlinear measures characterize signal irregularity and complexity, whereas spectral power features describe rhythmic neural activity across canonical frequency bands [16, 17, 18, 19, 20]. The complete feature set is summarized in Table 2.

Category	Features	Description
Entropy-based	ApEn, SampEn, PermEn, SVDEn, SpecEn	Signal irregularity and complexity
Hjorth parameters	Mobility, Complexity	Time-domain descriptors
Algorithmic / scaling	LZC, DFA	Sequence complexity and scaling
Fractal dimensions	HFD, KFD, PFD	Fractal complexity measures
Spectral power	δ, θ, α, β_1, β_2, β_3	Normalized band power

Table 2
Extracted nonlinear and spectral features for each epoch and electrode.

[1] https://github.com/Nikeshbajaj/spkit

3.2. Classification strategy and validation protocol

Emotion recognition is formulated as a set of binary classification tasks, following the experimental setup adopted in the NeuroSense dataset [12]. Specifically, for each emotional condition corresponding to a quadrant of Russell's circumplex model, EEG responses elicited by emotional stimuli are discriminated from the baseline condition. Model hyperparameters are optimized using `RandomizedSearchCV` with a maximum of 100 sampled configurations and a 3-fold cross-validation scheme within the training data. Balanced accuracy is used as the optimization metric to mitigate class imbalance. To evaluate subject-independent generalization performance, a Leave One Subject Out (LOSO) cross-validation protocol is employed. At each iteration, data from one participant are held out as an independent test set, while the remaining subjects constitute the training set. Hyperparameter optimization is performed exclusively on the training subjects in order to prevent information leakage. The optimal configuration identified during training is subsequently evaluated on the held-out subject. This procedure is repeated for all participants, and the final performance is computed as the average across all LOSO folds. Table 3 summarizes the evaluated models and the corresponding hyperparameter search spaces.

NimbusQDA	**NimbusSoftmax**
class_prior_alpha, μ_{loc}, μ_{scale}	w_{scale}, b_{scale}, learning rate, num_steps
RidgeClassifier	**Logistic Regression**
α	C, penalty
LDA (LSQR)	**SVC (linear)**
shrinkage	C
SVC (RBF)	**SVC (poly)**
C, γ	C, γ, degree
Random Forest	**Gradient Boosting**
n_estimators, max_depth	n_estimators, learning_rate, max_depth
AdaBoost	**GaussianNB**
n_estimators, learning_rate	var_smoothing
MLPClassifier	**XGBoost**
hidden_layer_sizes, activation, α, lr	n_estimators, max_depth, learning_rate

Table 3
Compact summary of the main tuned hyperparameters for each classifier.

3.3. Region of interest

To identify the most informative temporal regions of interest, model performance was evaluated across consecutive segments of the stimulus period. The EEG signal was divided into stimulus-locked epochs of 5 s, and balanced accuracy was computed independently for each epoch using the classification pipeline described above. For each subject, classification performance was ranked across temporal segments to identify which portions of the EEG response provided the greatest discriminative power for emotion decoding. The temporal window with the highest average performance across subjects was considered the most informative ROI.

3.4. Explainability with SHAP

To interpret the contribution of EEG features to the model predictions, we employed SHapley Additive exPlanations (SHAP) [9]. SHAP is a feature attribution method grounded in cooperative game theory that assigns to each feature a Shapley value representing its marginal contribution to the model output.

SHAP explanations were computed for the best-performing classifier identified during the validation procedure. In more detail, SHAP values were computed for the independent test subject in each LOSO fold and aggregated to obtain spatially interpretable attribution patterns. The aggregation procedure consisted of the following steps:

- SHAP values were computed for each sample in the test set of the held-out subject.
- Feature contributions belonging to the same electrode were summed (preserving sign) in order to obtain channel-level importance scores.

- Channel-wise SHAP values were averaged across samples within each subject.
- Subject-level values were then averaged across all LOSO folds to obtain a global cross-subject attribution profile.
- The resulting 14-dimensional vector was projected onto the Emotiv electrode layout and visualized as a scalp topography.

In addition, feature-level importance was quantified by computing the mean absolute SHAP value across all samples. Features were ranked according to their global contribution, and the top 10 most influential features were visualized using horizontal bar plots. This analysis enables the identification of spatially consistent attribution patterns, highlighting the electrodes and feature families that contribute most strongly to emotion decoding.

4. Results and discussion

This section presents the results obtained from the EEG-based emotion classification models and the SHAP-based feature attribution analysis. The findings are organized according to the research questions introduced in Section 1 and focus on three aspects: temporal stability of the EEG response, spatial relevance of electrodes, and the dominance of specific feature families in the explanations.

RQ1: How stable are the temporal patterns of emotion-related EEG activity across subjects?

Figure 1 reports the balanced accuracy obtained for each emotional condition across consecutive 5 s stimulus-locked EEG epochs. This analysis enables a comparative assessment of the temporal discriminative power of the EEG signal during emotion elicitation.

Across all emotions, the first temporal segment consistently exhibits the highest classification performance. This indicates that the initial 5 s following stimulus onset represent the most informative window for emotion discrimination. The performance progressively decreases in subsequent segments, suggesting that the most salient neural response occurs shortly after the emotional stimulus begins.

These findings are consistent with the observations reported in the NeuroSense dataset study [12], where the early post-stimulus interval was also identified as the most discriminative phase of the EEG response. Despite the differences in acquisition devices and electrode configurations, both studies highlight the importance of early stimulus-driven neural dynamics for emotion recognition. This consistency suggests that temporal patterns of emotion-related EEG activity may represent a robust characteristic of affective processing.

RQ2: Which spatial regions contribute most strongly to emotion decoding in a 14-channel wearable montage?

Figures 2a–2d present the spatial distribution of SHAP values projected onto the scalp topography for each emotional condition. Channel-wise SHAP contributions are mapped according to electrode positions, enabling a spatial interpretation of the EEG regions that contribute most strongly to model predictions. The resulting SHAP topomaps reveal that fronto-temporal and centro-occipital regions consistently emerge as areas with high attribution scores. In particular, frontal electrodes (AF3, AF4, F3, and F4) show strong contributions across multiple emotional conditions, often with asymmetric activation patterns. Such frontal asymmetries have frequently been associated with affective processing and emotional regulation mechanisms in EEG studies. Temporal electrodes (T7, T8) and posterior regions (O1, O2) also exhibit notable contributions to specific emotional states, suggesting that distributed cortical networks participate in emotion decoding. These patterns indicate that even with a low-density wearable montage, it is possible to identify spatially meaningful neural signatures associated with emotional processing. Interestingly, these spatial attribution patterns are broadly consistent with those observed in previous explainability analyses conducted on the NeuroSense dataset [11]. In that work,

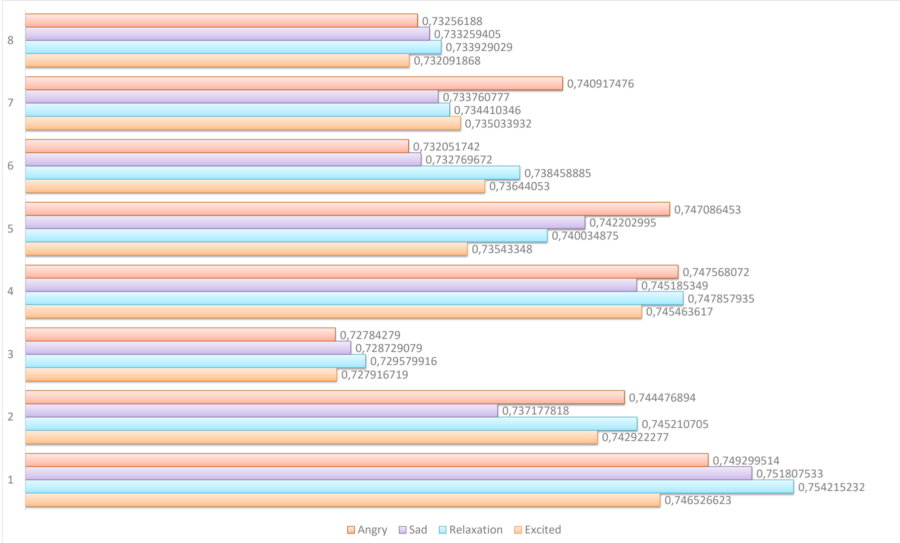
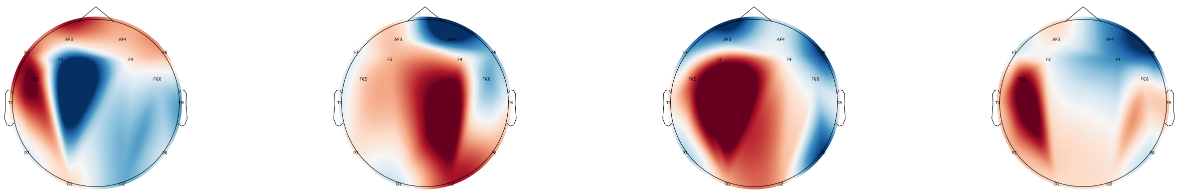


Figure 1: Balanced accuracy across consecutive 5-s EEG epochs for each emotional condition.

frontal and temporal electrodes also emerged as dominant contributors to model explanations, despite the use of a different device and a much sparser electrode configuration. The recurrence of similar spatial patterns across datasets supports the hypothesis that interpretable EEG markers of emotion may remain detectable even when acquisition settings change.



(a) Excited (b) Relaxation (c) Sad (d) Angry

Figure 2: Spatial distribution of SHAP values projected onto the scalp topography for each emotional condition.

RQ3: Which feature families dominate the model explanations?

Figure 3 presents the top-10 features ranked by mean absolute SHAP value for each emotional condition. Across all emotions, nonlinear complexity measures consistently dominate the explanations. Among these descriptors, the Katz Fractal Dimension (KFD) emerges as the most recurrent feature, appearing in the majority of the top-ranked features across electrodes and emotional states. Fractal-based measures capture the irregularity and self-similarity properties of EEG signals, reflecting complex temporal dynamics that are often associated with cognitive and affective processes. More specifically, excited and sad conditions exhibit strong contributions from KFD-based descriptors in frontal, fronto-central, and occipital electrodes. In contrast, the relaxation condition shows increased contributions from temporal and parietal regions, suggesting different neural dynamics associated with low-arousal emotional states.

These findings align with previous analyses performed on the NeuroSense dataset [11], where entropy and nonlinear complexity features also emerged as dominant contributors in SHAP explanations. While the previous study highlighted a combination of entropy-based and spectral features, the present work further emphasizes the relevance of fractal descriptors for emotion decoding in wearable EEG setups.

The results suggest that nonlinear descriptors of signal complexity provide particularly informative representations of emotional EEG activity. The recurrence of similar feature families across different datasets and acquisition devices supports the hypothesis that irregular temporal dynamics constitute a robust neurophysiological signature of affective processing in low-density EEG recordings.

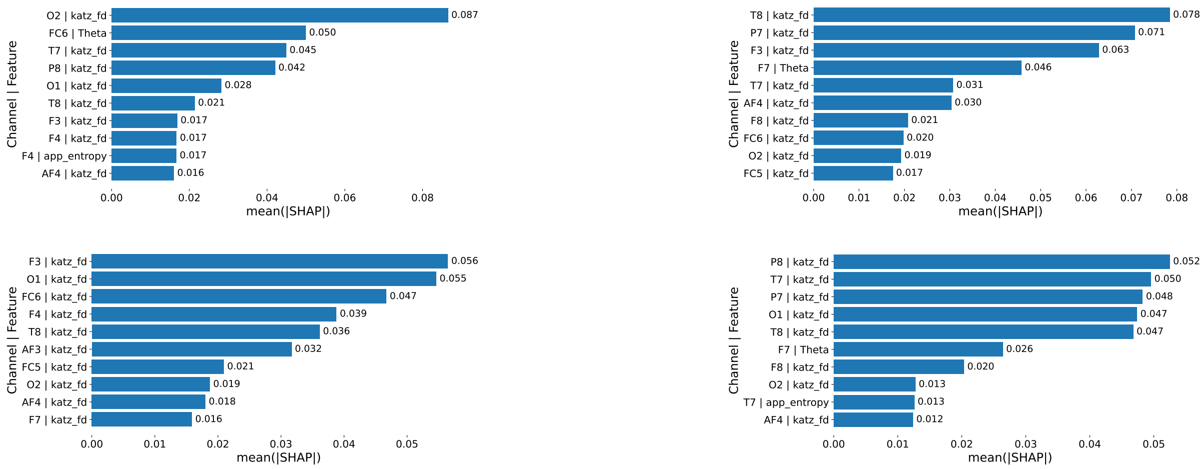


Figure 3: Top-10 SHAP features ranked by mean absolute contribution for each emotional condition: excited (top-left), relaxation (top-right), sad (bottom-left), and angry (bottom-right).

5. Conclusions and future work

This work presented an XAI-oriented pipeline for emotion recognition from EEG signals acquired using a 14-channel wearable Emotiv device. The proposed framework combines handcrafted spectral and nonlinear descriptors with machine learning models and SHAP-based explainability in order to investigate the neurophysiological factors underlying emotion decoding. The analysis highlights the presence of stable and interpretable attribution patterns in wearable EEG-based emotion recognition models. In particular, spatial relevance patterns and feature-level explanations show consistent structures across emotional conditions, with nonlinear descriptors emerging as key indicators of affective neural dynamics. These observations align with previous findings obtained on the NeuroSense dataset, suggesting that meaningful EEG markers of emotion recognition can remain observable even across different low-cost acquisition setups.

Future work will investigate the stability of these explainability patterns across additional EEG datasets and acquisition devices. Furthermore, we plan to explore channel reduction strategies to identify minimal electrode configurations capable of preserving both predictive performance and interpretable attribution patterns in wearable EEG systems.

Acknowledgments

This work was partially supported by the project DEMETRA: "Development of an ensemble learning-based, multidimensional sensory impairment score to predict cognitive impairment in an elderly cohort of Southern Italy" (CUP D99J22001970006) Missione 6/componente 2/Investimento: 2.1 "Rafforzamento e potenziamento della ricerca biomedica del SSN", funded by European Commission – NextGenerationEU.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar, P. Marttinen, Eeg based emotion recognition: A tutorial and review, ACM Computing Surveys 55 (2022) 1–57.

[2] Y. Chen, Y. Peng, J. Tang, T. Camilleri, K. Camilleri, W. Kong, A. Cichocki, Eeg-based affective brain–computer interfaces: Recent advancements and future challenges, Journal of Neural Engineering 22 (2025) 031004.

[3] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis using physiological signals, IEEE Transactions on Affective Computing 3 (2011) 18–31.

[4] W.-L. Zheng, B.-L. Lu, Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks, IEEE Transactions on autonomous mental development 7 (2015) 162–175.

[5] S. Katsigiannis, N. Ramzan, Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices, IEEE journal of biomedical and health informatics 22 (2017) 98–107.

[6] C. He, Y.-Y. Chen, C.-R. Phang, C. Stevenson, I.-P. Chen, T.-P. Jung, L.-W. Ko, Diversity and suitability of the state-of-the-art wearable and wireless eeg systems review, IEEE Journal of Biomedical and Health Informatics 27 (2023) 3830–3843.

[7] H. A. Hamzah, K. K. Abdalla, Eeg-based emotion recognition datasets for virtual environments: A survey, Applied Computational Intelligence and Soft Computing 2024 (2024) 6091523.

[8] O. Presacan, J. Ojha, A. Yazidi, E. Monteiro, P. G. Lind, A comprehensive review of explainable ai in deep learning algorithms for eeg analysis, ACM Transactions on Computing for Healthcare (2025).

[9] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Curran Associates, Inc., 2017.

[10] M. L. N. De Bonis, G. Fasano, A. Lombardi, C. Ardito, A. Ferrara, E. Di Sciascio, T. Di Noia, Explainable brain age prediction: a comparative evaluation of morphometric and deep learning pipelines, Brain Informatics 11 (2024) 33.

[11] T. Colafiglio, A. Lombardi, P. Sorino, D. Lofù, D. Danese, F. Narducci, E. Di Sciascio, T. Di Noia, Explainable evaluation of emotion recognition with low-cost eeg: Feature engineering and interpretability insights (2025).

[12] T. Colafiglio, A. Lombardi, P. Sorino, E. Brattico, D. Lofù, D. Danese, E. Di Sciascio, T. Di Noia, F. Narducci, Neurosense: a novel eeg dataset utilizing low-cost, sparse electrode devices for emotion exploration, IEEE Access 12 (2024) 159296–159315.

[13] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, DEAP: A database for emotion analysis using physiological signals, IEEE Transactions on Affective Computing 3 (2012) 18–31.

[14] M. M. Bradley, P. J. Lang, Measuring emotion: the self-assessment manikin and the semantic differential, Journal of behavior therapy and experimental psychiatry 25 (1994) 49–59.

[15] J. A. Russell, A circumplex model of affect., Journal of personality and social psychology 39 (1980) 1161.

[16] B. Hjorth, Eeg analysis based on time domain properties, Electroencephalography and Clinical Neurophysiology 29 (1970) 306–310.

[17] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, A. L. Goldberger, Mosaic organization of dna nucleotides, Physical Review E 49 (1994) 1685–1689.

[18] T. Higuchi, Approach to an irregular time series on the basis of the fractal theory, Physica D: Nonlinear Phenomena 31 (1988) 277–283.

[19] M. J. Katz, Fractals and the analysis of waveforms, Computers in Biology and Medicine 18 (1988) 145–156.

[20] A. Petrosian, Kolmogorov complexity of finite sequences and recognition of different preictal eeg patterns, in: Proceedings of the IEEE Symposium on Computer-Based Medical Systems, 1995.