Explainable Evaluation of Emotion Recognition with Low-Cost EEG: Feature Engineering and Interpretability Insights

Tommaso Colafiglio[1,2], Angela Lombardi[2,*], Paolo Sorino[2], Domenico Lofù[2], Danilo Danese[2], Fedelucio Narducci[2], Eugenio Di Sciascio[2] and Tommaso Di Noia[2]

[1] *Department of Computer, Control, and Management Engineering (DIAG), Sapienza University of Rome, Rome, RM, 00185, Italy*
[2] *Department of Electrical and Information Engineering, Politecnico di Bari, Bari, BA, 70125, Italy*

Abstract

Emotion recognition from EEG signals is a core task in affective computing, with growing relevance for real-world applications. In this study, we analyze the NeuroSense Emotion Recognition Dataset, acquired with the Muse 2 brain-computer interface (BCI), a portable, low-cost EEG device with only four electrodes. We implement a complete pipeline that includes signal preprocessing, handcrafted feature extraction (spectral, entropy, autoregressive), and classification using machine learning models under a Leave-One-Subject-Out (LOSO) cross-validation scheme. The models achieve an average accuracy and F1-score of 70% across the four quadrants of Russell's emotional model. To improve transparency, we apply SHAP to evaluate feature importance across subjects and emotional states. The analysis reveals both shared and emotion-specific EEG markers but also highlights a high degree of inter-subject variability in SHAP values. These findings underscore the challenges of generalization in EEG-based emotion recognition and point to the need for adaptive and personalized approaches. This work contributes preliminary but actionable insights toward interpretable, lightweight, and user-aware emotion-aware BCI systems.

Keywords

Emotion Recognition, Brain Computer Interface, Artificial Intelligence, BCI, XAI

1. Introduction

Emotion recognition from EEG signals is gaining momentum in affective computing, with applications in adaptive interfaces, education, and healthcare [1, 2]. Despite the success of high-density EEG datasets, real-world usability remains limited by the complexity and cost of traditional acquisition systems[3, 4].

Recent studies have introduced advanced deep learning architectures [5], domain adaptation strategies [6], and graph-based methods [7]. However, these often require dense montages and lack transparency. In contrast, handcrafted features—especially in low-density setups—offer interpretability and efficiency, yet their role in emotion recognition with portable EEG devices remains underexplored [8].

Furthermore, the integration of eXplainable AI (XAI) is still in its infancy in EEG-based affective systems. SHapley Additive Explanations (SHAP) [9] provide a principled method to assess feature importance in black-box models. Applied to EEG, SHAP can illuminate the neurophysiological basis of emotional states and support trust and personalization in BCI systems [10]. Still, little is known about the consistency of SHAP-derived attributions across subjects, or whether shared EEG features underpin different emotions.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

[*] Corresponding author.

tommaso.colafiglio@poliba.it (T. Colafiglio); angela.lombardi@poliba.it (A. Lombardi); paolo.sorino@poliba.it (P. Sorino); domenico.lofu@poliba.it (D. Lofù); danilo.danese@poliba.it (D. Danese); fedelucio.narducci@poliba.it (F. Narducci); eugenio.disciascio@poliba.it (E. Di Sciascio); tommaso.dinoia@poliba.it (T. Di Noia)

0000-0001-7184-310X (T. Colafiglio); 0000-0003-1815-9522 (A. Lombardi); 0000-0002-9081-2648 (P. Sorino); 0000-0001-6413-9886 (D. Lofù); 0009-0000-5203-1229 (D. Danese); 0000-0002-9255-3256 (F. Narducci); 0000-0002-5484-9945 (E. Di Sciascio); 0000-0002-0939-5462 (T. Di Noia)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

In this work, we analyze the NeuroSense dataset [11], collected with Muse 2 - a low-cost four-electrode EEG device. Our pipeline includes preprocessing, feature extraction, classification via Leave-One-Subject-Out cross-validation, and SHAP-based interpretability analysis.

We address the following research questions:

- **RQ1:** How do engineered features influence EEG-based emotion classification performance?
- **RQ2:** Are feature importance explanations consistent across subjects for the same emotion?
- **RQ3:** Do different emotions rely on shared or distinct EEG features?

To this end, we propose and evaluate a complete pipeline that integrates handcrafted EEG feature extraction, machine learning-based classification, and post-hoc explainability, applied to data collected with a low-cost, four-channel BCI device.

2. Methodology

This study adopts a structured pipeline to evaluate EEG-based emotion recognition using interpretable machine learning techniques. We used the NeuroSense Emotion Recognition Dataset [11], which includes EEG signals from 30 participants recorded with the Muse 2 BCI, a low-cost device with four electrodes (AF3, AF4, TP9, TP10). Each subject was exposed to 40 music video clips, carefully selected to elicit different emotional responses. The stimuli were categorized according to Russell's circumplex model [12], which organizes emotions based on valence (positive/negative) and arousal (high/low activation). EEG data were recorded during both a baseline phase (before stimulus presentation) and an emotion induction phase, ensuring a structured analysis of emotional states. More details on the protocol and characteristics of the dataset can be found in the original work [11]. This section outlines the processing pipeline, including EEG preprocessing and feature extraction, classification strategies, and model interpretability via SHAP.

2.1. Preprocessing and feature engineering

EEG data were preprocessed through bandpass filtering (1–40 Hz), normalization, detrending, and artifact removal via the Artifact Subspace Reconstruction (ASR) method [13]. Signals were segmented into 5-second epochs for both baseline and stimulus periods.

To enhance model performance in this low-density setup, we extracted handcrafted features using TorchEEG[1], grouped into:

- Frequency-domain features (e.g., PSD in delta to gamma bands),
- Entropy-based features (e.g., approximate/sample entropy, fractal dimensions, Hurst exponent) [14],
- Autoregressive coefficients modeling temporal dependencies [15].

2.2. Classification strategy

We employed a Leave-One-Subject-Out (LOSO) cross-validation scheme [16], which simulates real-world deployment by testing generalization on unseen individuals [17]. For each iteration, models were trained on data from 29 participants and tested on the remaining one, repeated over all subjects.

Five classifiers were tested: SVM, Ridge Classifier, Random Forest, Multi-Layer Perceptron (MLP), and K-Nearest Neighbors (KNN). Hyperparameters for each model were optimized through grid search, as detailed in Table 1.

Model performance was assessed using accuracy, precision, recall, and F1-score, calculated independently for each of the four emotional quadrants defined by the circumplex model. This process yielded a total of 120 trained models (30 subjects × 4 emotions).

[1] https://torcheeg.readthedocs.io/en/latest/

Table 1
Hyperparameter Grid for different models

Model	Hyperparameters
SVC	scaler: MinMaxScaler, RobustScaler, StandardScaler C: Logarithmic scale from 10^{-3} to 10^{2} kernel: linear, rbf, sigmoid gamma: auto tol: Logarithmic scale from 10^{-4} to 10^{-1}
RidgeClassifier	scaler: MinMaxScaler, RobustScaler, StandardScaler alpha: 1, 3, 5, to 99
RandomForest	scaler: MinMaxScaler, RobustScaler, StandardScaler n_estimators: 10, 20, 30, 40 max_depth: 5, 15, 25, 35, 45 min_samples_split: 2, 5, 10 min_samples_leaf: 1, 2, 4, 6, 8, 0 criterion: gini, entropy
MLP	scaler: MinMaxScaler, RobustScaler, StandardScaler hidden_layer_sizes: 50, 100, 150, 50 50, 100 50, 100 100 activation: relu alpha: Logarithmic scale from 10^{-4} to 1 learning_rate: adaptive
KNeighbors	scaler: MinMaxScaler, RobustScaler, StandardScaler n_neighbors: 1, 3, 5 to 49 weights: uniform, distance metric: euclidean, manhattan, chebyshev

2.3. Explainability with SHAP

To interpret model decisions, we applied Shapley Additive Explanations (SHAP) [9], a post-hoc technique based on cooperative game theory. SHAP values quantify each feature's contribution to a model's prediction, enabling fine-grained attribution analyses [17]. Given a feature x_i, its Shapley value ϕ_i is defined as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S) \right] \tag{1}$$

where N is the set of all features and $f(S)$ the model output when only features in S are considered.

For each emotional quadrant, we selected the best-performing model based on mean accuracy across subjects. These models were used to compute SHAP values for each test subject and for each trial. This enabled the identification of both shared and emotion-specific feature relevance patterns.

Moreover, to summarize feature importance, SHAP vectors were:

- averaged within each emotion to obtain quadrant-specific attribution profiles,
- ranked to identify the most influential features.

2.3.1. Inter-subject similarity

To evaluate consistency across individuals, Pearson correlation matrices were computed on SHAP vectors for each emotion. The resulting correlations quantified the similarity in feature attribution patterns among subjects and were summarized using boxplots to assess variability.

2.3.2. Cross-emotion comparison

To identify common or emotion-specific features, SHAP value distributions were compared across emotional quadrants. This analysis revealed whether the same features played similar roles in different emotional states or contributed in a differentiated, emotion-dependent manner.

3. Results and discussion

This section presents the results obtained from the EEG-based emotion classification models and the SHAP-based feature attribution analysis, structured according to the research questions outlined in the Section Introduction.

3.1. RQ1: How do engineered features influence EEG-based emotion classification performance?

The impact of feature engineering on classification performance was evaluated by training multiple machine learning models on the extracted feature set. Using LOSO, the best-performing classifier was selected for each Russell's emotional quadrant based on accuracy, precision, recall, and F1-score. The performance metrics of all models are reported in Tables 2.

The results show that SVM outperformed other classifiers in three quadrants (Excited, Sad, Angry), while Random Forest (RF) yielded the highest accuracy for the Relaxed state. The difference in model performance suggests that linear separability in feature space is particularly relevant for these emotions, while decision-tree-based models may better handle the variability present in low-arousal states.

Compared to the previous study introducing the NeuroSense dataset[11], which reported an average accuracy of 75% across the four emotional quadrants, the performance obtained in this work is slightly lower. The previous study employed MiniRocket, an algorithm that applies random convolutional kernels followed by global max pooling, generating a high-dimensional feature space that captures complex temporal patterns in the EEG signals. The extracted features were then classified using SVM, benefiting from the rich and diverse representations learned through the MiniRocket transformation.

In contrast, this work relied on a feature engineering approach, extracting spectral and entropy-based descriptors from the raw EEG signals. Although these features provided meaningful physiological insights into EEG-based emotion classification, their performance was slightly lower than the MiniRocket-based method. This suggests that convolutional feature extraction techniques, such as MiniRocket, may be particularly effective in leveraging hidden temporal dependencies in EEG signals. In contrast, engineered features may be better suited for improving model interpretability and explainability.

Table 2
Average performance of the ML models across the four emotions.

Model	Emotion	Precision	Recall	F1-score	Accuracy
KNeighbors	Excited	0.69	0.70	0.69	0.69
	Relaxed	0.68	0.68	0.67	0.68
	Sad	0.63	0.63	0.62	0.62
	Angry	0.67	0.67	0.67	0.67
RandomForest	Excited	0.70	0.70	0.70	0.70
	Relaxed	**0.70**	**0.70**	**0.70**	**0.70**
	Sad	0.65	0.66	0.65	0.65
	Angry	0.68	0.68	0.68	0.68
MLP	Excited	0.66	0.66	0.66	0.66
	Relaxed	0.70	0.70	0.69	0.69
	Sad	0.64	0.64	0.64	0.64
	Angry	0.68	0.68	0.68	0.68
SVC	Excited	**0.72**	**0.72**	**0.72**	**0.72**
	Relaxed	0.68	0.69	0.69	0.68
	Sad	**0.69**	**0.69**	**0.69**	**0.69**
	Angry	**0.70**	**0.70**	**0.70**	**0.70**

3.2. RQ2: Are feature importance explanations consistent across subjects for the same emotion?

To assess the consistency of feature attributions across subjects for a given emotional state, we computed Pearson correlation coefficients between the SHAP value vectors of all participant pairs. The distribution of these correlation values, summarized in the box plots of Figure 1, provides insight into the variability of feature importance rankings across participants for all the emotions. The results showed generally low inter-subject correlations across all four emotions, with values clustered near zero indicating high variability.

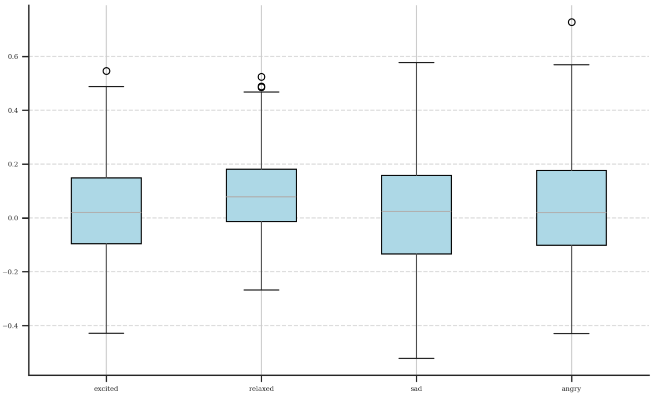

Figure 1: Boxplots of pairwise SHAP value correlations across subjects for each emotional state. Each boxplot summarizes the distribution of Pearson correlation coefficients between SHAP vectors from different participants. Lower correlation values reflect greater inter-subject variability in feature importance.

These findings suggest that the EEG features contributing to emotion classification differ substantially between individuals, regardless of the emotional quadrant. No specific emotion exhibited significantly higher consistency. This inter-subject variability may stem from differences in brain physiology, cognitive processing, or emotional perception.

Overall, the results highlight a key limitation of subject-independent models in EEG-based emotion recognition and support the development of adaptive or personalized approaches that can accommodate individual differences in feature attribution.

3.3. RQ3: Do different emotions rely on shared or distinct EEG features?

To investigate whether specific EEG features contribute similarly across emotions, we compared SHAP-based feature rankings for each emotional quadrant. Figure 2 shows the top 10 most important features for each emotional state, averaged across subjects.

The analysis revealed both shared and emotion-specific neural signatures. Excited states were dominated by spectral features in the theta and alpha bands at frontal electrodes (AF3, AF4), consistent with prior findings linking these rhythms to attentional engagement and arousal [18]. Relaxation was primarily associated with beta-band kurtosis and delta-band entropy at TP9, which may reflect reduced cortical excitability and idling activity [19].

Sad states showed a predominance of entropy-based measures—especially sample and approximate entropy—in delta and alpha bands at frontal and temporal sites. This supports evidence that emotional distress is associated with higher EEG complexity [20]. In contrast, anger was characterized by nonlinear and higher-order features, including delta power and fractal dimension, consistent with increased cortical activation and arousal [21].

From a spatial perspective (reported in Figure 3), the top features for excitement were primarily left-lateralized (AF3), supporting the frontal asymmetry hypothesis [22], whereas relaxed and sad states involved TP9, suggesting stronger temporal-parietal engagement [23]. Anger exhibited a more

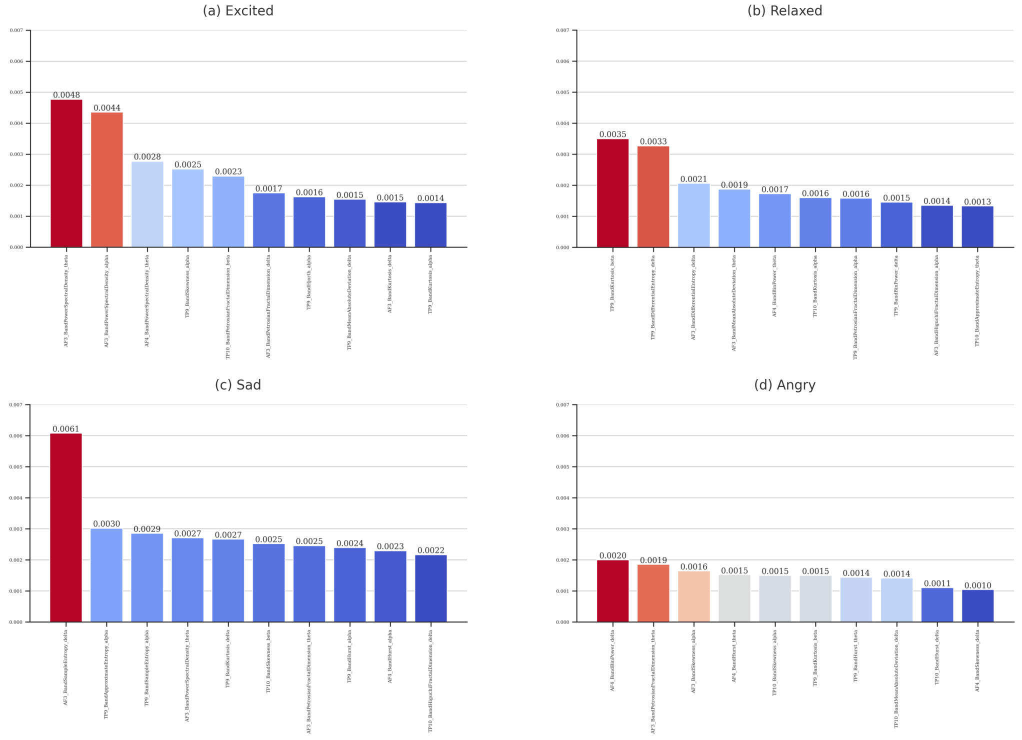

Figure 2: Top-10 SHAP-based features for each emotional state: (a) Excited, (b) Relaxed, (c) Sad, (d) Angry. Each panel shows the most relevant EEG features ranked by mean SHAP values across all subjects.

bilateral distribution, which may reflect motor-preparatory activity tied to action-oriented affective responses [24].

In terms of frequency content, theta and alpha bands were most relevant for high-arousal states (excited, angry), while delta and beta-related entropy measures were dominant for low-arousal states (relaxed, sad). These findings underscore the importance of combining generalizable spectral features with emotion-specific nonlinear descriptors to effectively characterize affective EEG responses.

Overall, the results suggest that no single feature set universally applies across emotions, and that emotion-specific feature selection strategies could enhance classification performance in EEG-based affective computing.

4. Conclusions and future work

This study explored interpretable EEG-based emotion recognition using a low-cost, sparse-electrode device, leveraging engineered features and SHAP analysis. Our results confirmed that spectral and entropy-based features enable competitive classification performance, yet highlighted substantial inter-subject variability in feature relevance.

While some EEG features showed cross-emotion generalizability, others were clearly emotion-specific, reflecting distinct neural mechanisms. These findings emphasize the limitations of subject-independent models and support the use of adaptive, personalized strategies in real-world affective BCI systems.

Future work will focus on dynamic feature selection and online adaptation techniques to improve robustness and user specificity. While the current study was limited to a single dataset, we plan to validate our findings on additional datasets to assess generalizability across acquisition setups. Furthermore, we aim to explore hybrid pipelines that combine engineered and learned representations to better capture complex EEG patterns without sacrificing interpretability. Finally, we intend to deepen the neurophysiological interpretation of EEG markers by involving domain experts and integrating insights from affective neuroscience.

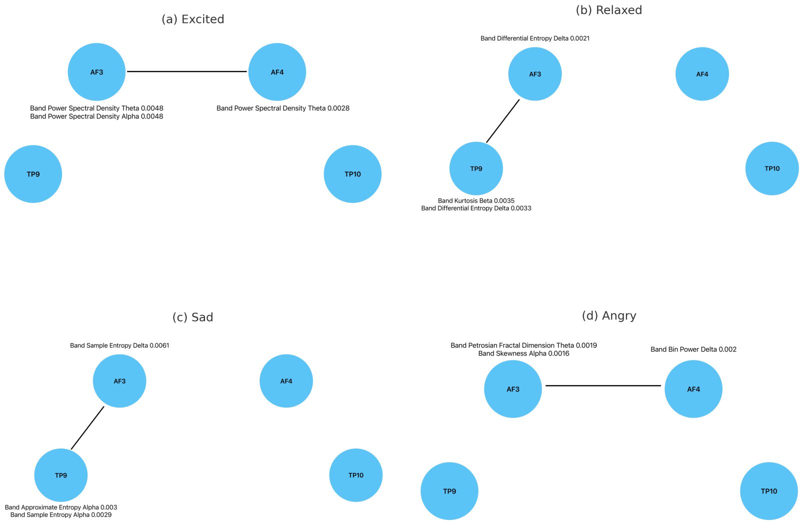
(a) Excited
AF3
AF4
Band Power Spectral Density Theta 0.0048
Band Power Spectral Density Alpha 0.0048
Band Power Spectral Density Theta 0.0028
TP9
TP10
(b) Relaxed
Band Differential Entropy Delta 0.0021
AF3
AF4
TP9
TP10
Band Kurtosis Beta 0.0035
Band Differential Entropy Delta 0.0033
(c) Sad
Band Sample Entropy Delta 0.0061
AF3
AF4
TP9
TP10
Band Approximate Entropy Alpha 0.003
Band Sample Entropy Alpha 0.0029
(d) Angry
Band Petrosian Fractal Dimension Theta 0.0019
Band Skewness Alpha 0.0016
Band Bin Power Delta 0.002
AF3
AF4
TP9
TP10

Figure 3: Spatial distribution of SHAP-based feature importance across emotional states: (a) Excited, (b) Relaxed, (c) Sad, (d) Angry. Each panel highlights the most relevant EEG features and their associated electrode locations.

Acknowledgments

This work is carried out while Tommaso Colafiglio is enrolled in Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with Politecnico di Bari. This work was partially supported by the following projects: ”LIFE: the itaLian system wIde Frailty nEtwork”; DEMETRA: ”Development of an ensemble learning-based, multidimensional sensory impairment score to predict cognitive impairment in an elderly cohort of Southern Italy”.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] W. Ma, Y. Zheng, T. Li, Z. Li, Y. Li, L. Wang, A comprehensive review of deep learning in eeg-based emotion recognition: classifications, trends, and practical implications, PeerJ Computer Science 10 (2024) e2065.

[2] T. Colafiglio, A. Lombardi, T. Di Noia, M. L. N. De Bonis, F. Narducci, A. M. Proverbio, Machine learning classification of motivational states: Insights from eeg analysis of perception and imagery, Expert Systems with Applications (2025) 127076.

[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] 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 (2018) 98–107. doi:10.1109/JBHI.2017.2688239.

[5] N. K. Gunda, M. I. Khalaf, S. Bhatnagar, A. Quraishi, L. Gudala, A. K. P. Venkata, F. Y. Alghayadh,

S. Alsubai, V. Bhatnagar, Lightweight attention mechanisms for eeg emotion recognition for brain computer interface, Journal of Neuroscience Methods 410 (2024) 110223.

[6] P. Yu, X. He, H. Li, H. Dou, Y. Tan, H. Wu, B. Chen, Fmlan: a novel framework for cross-subject and cross-session eeg emotion recognition, Biomedical Signal Processing and Control 100 (2025) 106912.

[7] C. Li, P. Li, Y. Zhang, N. Li, Y. Si, F. Li, Z. Cao, H. Chen, B. Chen, D. Yao, et al., Effective emotion recognition by learning discriminative graph topologies in eeg brain networks, IEEE Transactions on Neural Networks and Learning Systems (2023).

[8] M.-P. Hosseini, A. Hosseini, K. Ahi, A review on machine learning for eeg signal processing in bioengineering, IEEE reviews in biomedical engineering 14 (2020) 204–218.

[9] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[10] N. Sharma, T. K. R. Bollu, Explainable ai methods for interpreting emotions in brain–computer interface eeg data, in: Discovering the Frontiers of Human-Robot Interaction: Insights and Innovations in Collaboration, Communication, and Control, Springer, 2024, pp. 419–436.

[11] 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 (2024).

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

[13] S. Blum, N. S. Jacobsen, M. G. Bleichner, S. Debener, A riemannian modification of artifact subspace reconstruction for eeg artifact handling, Frontiers in human neuroscience 13 (2019) 141.

[14] S. Kesić, S. Z. Spasić, Application of higuchi's fractal dimension from basic to clinical neurophysiology: a review, Computer methods and programs in biomedicine 133 (2016) 55–70.

[15] J. Pardey, S. Roberts, L. Tarassenko, A review of parametric modelling techniques for eeg analysis, Medical engineering & physics 18 (1996) 2–11.

[16] 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.

[17] 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.

[18] M. X. Cohen, Error-related medial frontal theta activity predicts cingulate-related structural connectivity, Neuroimage 55 (2011) 1373–1383.

[19] E. Niedermeyer, F. L. da Silva, Electroencephalography: basic principles, clinical applications, and related fields, Lippincott Williams & Wilkins, 2005.

[20] B. M. Hager, A. C. Yang, J. N. Gutsell, Measuring brain complexity during neural motor resonance, Frontiers in neuroscience 12 (2018) 758.

[21] Y. Liu, O. Sourina, Eeg-based subject-dependent emotion recognition algorithm using fractal dimension, in: 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2014, pp. 3166–3171.

[22] J. J. Allen, P. M. Keune, M. Schönenberg, R. Nusslock, Frontal eeg alpha asymmetry and emotion: From neural underpinnings and methodological considerations to psychopathology and social cognition, 2018.

[23] H. Kober, L. F. Barrett, J. Joseph, E. Bliss-Moreau, K. Lindquist, T. D. Wager, Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies, Neuroimage 42 (2008) 998–1031.

[24] M. Nikolic, P. Pezzoli, N. Jaworska, M. C. Seto, Brain responses in aggression-prone individuals: A systematic review and meta-analysis of functional magnetic resonance imaging (fmri) studies of anger-and aggression-eliciting tasks, Progress in neuro-psychopharmacology and biological psychiatry 119 (2022) 110596.