XAI in Affective Computing: a Preliminary Study

Elena Sajno, Alessio Rossi, Stefano De Gaspari, Maria Sansoni, Giulia Brizzi and Giuseppe Riva

June 12, 2023

XAI in Affective Computing: A preliminary study

Elena Sajno[a,b,1], Alessio Rossi[c], Stefano De Gaspari[a,b], Maria Sansoni[d], Giulia Brizzi[e], and Giuseppe Riva[a,e]

[a]*HTLAB, Università Cattolica del Sacro Cuore, Milan, Italy*
[b]*Department of Computer Science, University of Pisa, Pisa, Italy*
[c]*Institute of Information Science and Technologies (ISTI), National Research Council of Italy (CNR), Pisa, Italy*
[d]*Department of Psychology, Università Cattolica del Sacro Cuore, Milan*
[e]*Applied Technology for Neuro-Psychology Laboratory, IRCCS Istituto Auxologico Italiano, Milan, Italy*
ORCiD ID: Elena Sajno https://orcid.org/0000-0002-9621-8981,
Alessio Rossi https://orcid.org/0000-0002-6400-5914,
Stefano De Gaspari https://orcid.org/0009- 0006-6083-2134,
Maria Sansoni https://orcid.org/0000-0002-5189-7159,
Giulia Brizzi https://orcid.org/0009-0000-7472-742X,
and Giuseppe Riva https://orcid.org/0000-0003-3657-106X

Abstract. Affective computing is a rapidly growing field that aims to understand human emotions through Artificial Intelligence. One of the most promising ways to achieve this goal is the use of physiological data (e.g. electrocardiogram - ECG) and Machine Learning (ML) algorithms to classify affective states. ECG correlates, such as Heart Rate Variability (HRV) and its features, are reported as viable indicators in both dimensional approaches, especially for valence, and in detecting discrete emotions. In this preliminary study, we used the ECG data from the open-source HCI Tagging Database, which includes physiological data and self-referred feedback from 30 subjects who watched videos designed to elicit different emotions. The subjects evaluated their reactions using a three-dimensional affective space defined by arousal, valence, and dominance levels and reported the emotions they felt. To classify the affective states, we trained and tested different classification algorithms on the HRV features, using as labels, each self-reported feedback (i.e., valence, arousal, dominance, and emotions). The results showed that HRV features, when combined with normalization methods and ML algorithms, were effective in recognizing emotions as experienced by individuals. In particular, the study showed that Decision Tree was the best-performing algorithm for predicting emotions based on HRV data. Additionally, an Explainable AI (XAI) model provided insights into the weight of these features in the ML discrimination phases. Overall, the study highlights the potential of HRV as a valid and unobtrusive source for detecting emotional states.

Keywords. Machine Learning, ECG, HRV, Emotion recognition, Explainable AI (XAI)

[1] Corresponding Author: Elena Sajno, elena.sajno@phd.unipi.it.

1. Introduction

Affective Computing is an approach defined by Picard [1] that links computing and human emotions; one of its branches aims at the detection of emotional states by interpreting with Machine Learning (ML) physiological signals. Affective Computer models typically follow a standard workflow [2–4], in which the subjects are presented with different emotional stimuli, while their EEG or peripheral signals are recorded. These data are used to populate a database, with each condition paired with a label, extrapolated from the personal evaluation provided by the subject. Data are preprocessed, and features, typical of the used signal, are computed. In some cases, normalization and further feature selections or aggregations are performed. A classification ML model is afterward trained to predict the corresponding label from the physiological data.

The labeling of the emotional state is defined in accordance with the selected emotion model, based on discrete emotions or continuous dimensions [3]. The first, initially proposed by Ekman [5] assumes that emotions are distinct, measurable, and universal. The idea of a continuous spectrum of emotions arises instead from Russell and his Circumplex Model of Affect [6], in which emotions are distributed in a cartesian space created around two perpendicular axes of pleasantness and activation. This idea would then be repurposed on a 3-axis space, Valence (from positive to negative), Arousal (from calm to excited), and Dominance (from submissive/controlled to Dominant) [7,8]. A viable solution for assessing these dimensions is the Self-Assessment Manikins (SAM), three graphical 9 or 5-point scales representing the affective experiences [8]. After data are correctly labeled in accordance with subjective perception, a classifier is trained to predict the emotional label from the physiological data: many algorithms are tested, and their performances compared. Literature reports a wide use of Support Vector Machine (SVM), K-nearest neighbors (KNN), Random Tree (RT), Decision Tree (DT), and more complex frameworks, like Deep Learning [3].

When a Database for Affective Computing is created, multiple sources of data are usually collected [3,9–11]: the most frequent signals are Electroencephalogram (EEG), Electrocardiogram (ECG), Electrodermal activity (EDA), Respiration patterns, Electromyography (EMG) of emotion-linked facial muscles, Skin Temperature, and Eye Gaze movements. In this work, we focus on ECG and, in particular, on Hearth Rate Variability (HRV) feature: these signals are found as a valid source for emotional [12,13] or arousal and valence levels discrimination [13,14]. ECG sensors are nowadays quite common (e.g., in smartwatches) and are particularly unobtrusive to the user [15].

Heart Rate Variability (HRV) measures the variability of the time intervals between heartbeats, which is related to changes in neuro-cardiac functions and is influenced by the autonomic nervous system [16]. Increases in HRV are associated with activation of the Parasympathetic Nervous System (SNP), while a decrease in HRV is linked to increased sympathetic activation. Furthermore, the different HRV frequency components can be linked to the individual's activation [17].

The aim of this preliminary study is the detection of patterns in HRV data that permits to predict the level of Valence, Arousal, and Dominance and the different underlying emotions. We also hypothesize that HRV features, when combined with different normalization methods and ML algorithms, would be more effective in comparison with other peripheral signals at recognizing emotions as experienced by the individuals. The use of an Explainable AI (XAI) model would provide additional insights into the weight of these features in the discrimination of emotional states. Additionally, the study aims to identify the best-performing ML algorithm and normalization method for each classification task, as well as providing performance results.

2. Methods

MAHNOB-HCI database [9], licensable in Open mode [18], was used in this study. MAHNOB-HCI collects multimodal responses (EEG, ECG, EDA, Respiration pattern, and Skin Temperature) of 27 subjects to 20 emotionally connected videos. Participants were asked to evaluate them by kind of Emotion, Valence, Arousal, Dominance, and Predictability levels.

A dataset was created with the ECG data paired with subject-given values (Arousal, Valence, and Dominance evaluated on a 9-point Likert Scale, and which emotion was stirred in them). Each row of the dataset refers to a specific emotion (both on discrete and continuous dimensions) perceived by a subject, paired with the ECG data of the baseline and of the emotion-stimulation task. Before extracting ECG information from the HR time series, outliers, and ectopic beats were removed from the signal and linear interpolation was computed to replace outlier values. Time and frequency HRV features were extracted from the corrected ECG time series.

Three different normalization conditions were tested: no normalization, normalization before baseline subtraction, and normalization after baseline subtraction. The data normalization was performed by ranging the data from 0 to 1. For each of the three normalization conditions, five supervised learning algorithms were applied for classification, with a 67/33 train-test proportion: Decision Tree, Random Forest, KNeighbors, Support Vector Machine [19], and XGBoost [20]. The following HRV features were used as input features: Min HR, NNI 50, RMMSSD, SDNN, HF, LF, LF/HF ratio, Total Power', Max HR, Mean HR, and Median NNI. The algorithms were trained to predict three different levels of valence, arousal, and dominance, (1-3 low, 4-6 medium, 7-9 high in the 3-step layout, as in [9,21]). Additionally, the algorithms were trained to predict which emotion (Neutral, Anger, Disgust, Fear, Joy, Happiness, Sadness, Surprise, Amusement, Anxiety) was being experienced. Each model also underwent a further Features Selection through a Recursive Feature Elimination with Cross-Validation (RFECV) to select which and how many features should increase the functionality of the model and the dummy value, through a Dummy Classifier, as a benchmark against randomness [19].

To understand the contribution of each feature to the classification tasks, a SHapley Additive exPlanations (SHAP) analysis was performed, both globally and locally, using the SHAP library [22]. SHAP values represent the contribution of each feature to the output of a model, and they help in understanding the decision-making process of the model. This approach is obtained through a model-agnostic explanator, i.e. functions on different kinds of models, without needing to be specifically set [23].

3. Results

In Table 1 are reported the results of the best-performing algorithms for each classification: Arousal, Valence, Dominance, and Emotions. Random Forest and K-Nearest Neighbors perform best for Arousal levels and Emotion discriminations, on not-normalized data and without features-selection, while Decision Tree obtains the best results for Valence levels, on non-normalized data, and on Dominance levels, on data normalized after the subtraction of the baselines, both after an RFECV Features Selection

Table 1. Performance results are divided into target experiences or emotions. The best performance is reported and, in brackets, the Dummy Classifier results. The best-performing algorithm is reported alongside the kind of normalization (NN = no normalization, BN = normalization applied before subtracting the baseline, AN= normalization applied after having subtracted the baseline) and feature selection (YF= yes, NF= no) on which the results were obtained.

Target	Best Performance	Best performing Algorithm
Arousal (3 levels)	51% (28%)	Random Forest and K-nearest neighbors (NN, NF)
Valence (3 levels)	51% (32%)	Decision Tree (NN, YF)
Dominance (3 levels)	46% (26%)	Decision Tree (AN, YF)
Emotion (9 classes)	27% (12%)	Random Forest and K-nearest neighbors (NN, NF)

In Table 2 more detailed results of the best performing algorithm for the recognition of distinct emotions are reported: the states Neutral, Joy/Happiness, Sadness, and Amusement are detected with more than 20% accuracy, Disgust, and Anxiety show a worse performance, while Anger, Fear, and Surprise are never detected. These results are usually influenced by the number of Supports (or examples) of that category: the emotions less represented in the dataset also have weak performance.

Table 2. Detailed performance results for the Emotion detection, obtained through Random Forest. Results are divided by emotion. Precision is the percentage of samples that are positive, based on predictions, Recall is the proportion of positive samples that the predictions successfully capture, F-score is equal to the harmonic mean of recall and precision, and Support is the number of samples in the specific class [24]

Emotion	Precision	Recall	F1-score	Support
Neutral	0.28	0.39	0.32	18
Anger	0.00	0.00	0.0	3
Disgust	0.14	0.11	0.1	9
Fear	0.00	0.00	0.00	5
Joy, Happiness	0.30	0.2	0.27	16
Sadness	0.38	0.66	0.48	12
Surprise	0.00	0.00	0.00	4
Amusement	0.20	0.17	0.18	17
Anxiety	0.16	0.20	0.18	5
Accuracy	0.27	0.27	0.27	0.27
Macro average	0.16	0.20	0.17	89
Weighted average	0.22	0.27	0.24	89

Explanations of the result have been reached with Shap. Figure 1 reports the global explanations (features importance) for Arousal, Valence, and Dominance levels and the different emotions. HR mean appears as an important feature for Arousal, Valence, and Emotions discrimination, LF/HF ratio for Dominance, and Emotions, total ECG power for Arousal and Dominance, LF for Arousal and Valence, while RMSSD seems to be particularly influent in the discrimination of different emotions.

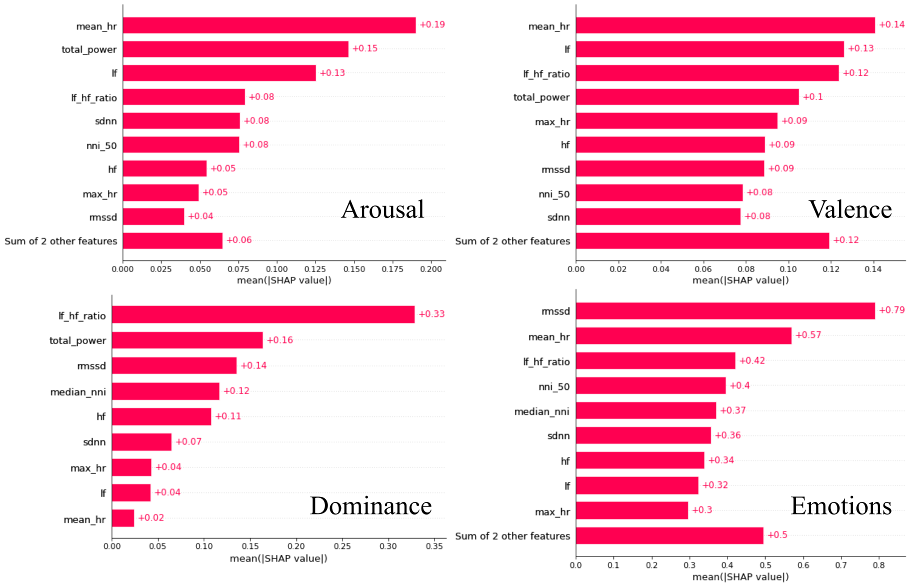

Figure 1. Global SHAP explanation for the result of the best performer algorithm for Arousal, Valence, Dominance, and Emotions. The features are classified for influence to reach the prediction.

For local explanations, a specific example is instead selected, and the weight of the feature is calculated, differentiating for positive or negative influences. Some local explanations for correctly detected emotions are reported in Figure 2, offering some insight into the weight of the features.

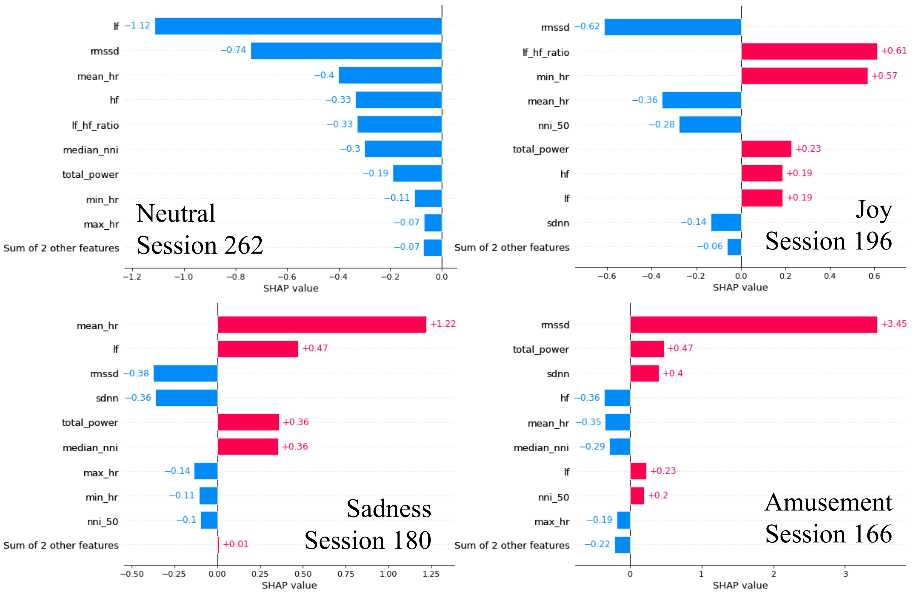

Figure 2. Local SHAP explanation for same selected Sessions of correctly detected emotions. The features are classified for influence to reach the prediction, both in positive and in negative directions.

4. Discussion and conclusions

The main finding of this preliminary study is that Valence, Arousal, and Dominance levels are characterized by different HR responses. Conversely, only a few emotions seem to affect the HR responses, but this result is deemed due to the low sample size in each emotion. Future research with a larger dataset should be performed to better understand the role of HR in emotion detection.

The results obtained in this study are in line with results reported on the same database. Compared to the results in Soleymani et al., the use of only HRV features reaches better results than including all the peripheral physiology (the reported performance is 46% for Arousal, and 45.5% for Valence) [9]. Ferdinando et al. considered also the ECG signals alone: they raised the performance from 42.6% to 64.1% for Valence and 47.7% to 69.6% for Arousal by applying multiple kinds of Dimensionality Reductions [21]: these techniques, however, render the interpretability of the model more complex, as they modify the input data and make them less recognizable. ECG seems a good signal for detecting Arousal, Valence, and Dominance, especially considering the ease of use and the scarce invasiveness to the subject.

With regard to emotion detection, the performance appears low but higher than the baseline (+15%accuracy): this can partially be explained by the number of examples for each emotion that reduces the training process accuracy. In fact, emotions with the highest sample size can be accurately distinguished and recognized (Table 2). If literature, with different databases, sometimes reports really high results, it needs to be noted how often the discrimination is performed just between two classes, without including results from a dummy classifier that indicates the validity of the prediction [3,4].

The local explanation provided in Figure 2 allows us to evaluate the decision-making process of the ML algorithm to assign an emotion to an example. For example, the low normalized difference between baseline and emotion state in each HRV feature indicates a Neutral emotion. This result indicates that the Neutral emotion does not affect the HRV responses. Differently, the probability to perceive Joy increases as LF-HF ratio, minimal HR, total power, HF, and LF differences increase, while it decreases as the other HRV feature differences increase. Similar results were detected for the other emotions (i.e., Sadness and Amusement) indicating these emotions activate both sympathetic and parasympathetic nervous systems.

In conclusion, the study evaluates the feasibility of using different activation of sympathetic and parasympathetic nervous systems detected by HRV analysis to accurately detect valence, arousal, and dominance levels, showing promising results. Furthermore, even if the accuracy of detecting emotion is low, the results of this work are encouraging, and suggest that with adequate sample size, it will be possible to create an ML model which accurately detects emotions. These preliminary findings are a first step that could lead to the creation of more complex applications aimed at detecting a variety of mental and emotional states. [25,26].

References

[1] Picard RW. Affective Computing. MIT Press; 2000.

[2] Zheng W-L, Lu B-L. Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. IEEE Trans Auton Ment Dev 2015;7:162–75. https://doi.org/10.1109/TAMD.2015.2431497.

[3] Bota PJ, Wang C, Fred ALN, Plácido Da Silva H. A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals. IEEE Access 2019;7:140990–1020. https://doi.org/10.1109/ACCESS.2019.2944001.

[4] Wang Y, Song W, Tao W, Liotta A, Yang D, Li X, et al. A systematic review on affective computing: emotion models, databases, and recent advances. Inf Fusion 2022;83-84:19–52. https://doi.org/10.1016/j.inffus.2022.03.009.

[5] Ekman P. An argument for basic emotions. Cognition and Emotion 1992;6:169–200. https://doi.org/10.1080/02699939208411068.

[6] Russell JA. A circumplex model of affect. J Pers Soc Psychol 1980;39:1161.

[7] Mehrabian A, Russell JA. An approach to environmental psychology 1974;266.

[8] Bradley MM, Lang PJ. Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. J Behav Ther Exp Psychiatry 1994;25:49–59. https://doi.org/10.1016/0005-7916(94)90063-9.

[9] Soleymani M, Lichtenauer J, Pun T, Pantic M. A Multimodal Database for Affect Recognition and Implicit Tagging. IEEE Transactions on Affective Computing 2012;3:42–55. https://doi.org/10.1109/T-AFFC.2011.25.

[10] Koelstra S, Muhl C, Soleymani M, Lee J-S, Yazdani A, Ebrahimi T, et al. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. IEEE Transactions on Affective Computing 2012;3:18–31. https://doi.org/10.1109/T-AFFC.2011.15.

[11] Sajno E, Bartolotta S, Tuena C, Cipresso P, Pedroli E, Riva G. Machine learning in biosignals processing for mental health: A narrative review. Front Psychol 2022;13:1066317. https://doi.org/10.3389/fpsyg.2022.1066317.

[12] Scheirer J, Fernandez R, Klein J, Picard RW. Frustrating the user on purpose: a step toward building an affective computer. Interact Comput 2002;14:93–118. https://doi.org/10.1016/S0953-5438(01)00059-5.

[13] Valenza G, Citi L, Lanatá A, Scilingo EP, Barbieri R. Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics. Sci Rep 2014;4:4998. https://doi.org/10.1038/srep04998.

[14] Girardi D, Ferrari A, Novielli N, Spoletini P, Fucci D, Huichapa T. The way it makes you feel predicting users' engagement during interviews with biofeedback and supervised learning. 2020 IEEE 28th International Requirements Engineering Conference (RE), IEEE; 2020, p. 32–43.

[15] Hasnul MA, Aziz NAA, Alelyani S, Mohana M, Aziz AA. Electrocardiogram-Based Emotion Recognition Systems and Their Applications in Healthcare—A Review. Sensors 2021;21:5015. https://doi.org/10.3390/s21155015.

[16] Shaffer F, Ginsberg JP. An Overview of Heart Rate Variability Metrics and Norms. Front Public Health 2017;5:258. https://doi.org/10.3389/fpubh.2017.00258.

[17] Camm AJ, Malik M, Bigger JT, Breithardt G, Cerutti S, Cohen RJ, et al. Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology 1996.

[18] Koelstra S. HCI tagging database - home n.d. https://mahnob-db.eu/hci-tagging/ (accessed March 24, 2023).

[19] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research, 12 2011:2825–30.

[20] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: Association for Computing Machinery; 2016, p. 785–94. https://doi.org/10.1145/2939672.2939785.

[21] Ferdinando H, Seppänen T, Alasaarela E. Enhancing emotion recognition from ECG signals using supervised dimensionality reduction. Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods, SCITEPRESS - Science and Technology Publications; 2017. https://doi.org/10.5220/0006147801120118.

[22] Lundberg S, Lee S-I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017, 30 2017.

[23] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A Survey of Methods for Explaining Black Box Models. ACM Comput Surv 2018;51:1–42. https://doi.org/10.1145/3236009.

[24] Müller AC, Guido S. Introduction to Machine Learning with Python: A Guide for Data Scientists. "O'Reilly Media, Inc."; 2016.

[25] Sajno E, Beretta A, Novielli N, Riva G. Follow the Flow: A Prospective on the On-Line Detection of Flow Mental State through Machine Learning 2022. https://doi.org/10.31234/osf.io/9z5pe.

[26] Sajno E, Riva G. Follow the Flow: Artificial Intelligence and Machine Learning for Achieving Optimal Performance. Cyberpsychol Behav Soc Netw 2022;25:476–7. https://doi.org/10.1089/cyber.2022.29251.ceu.