Fine-grained Emotion Recognition using Brain-Heart Interplay measurements and eXplainable Convolutional Neural Networks

Guido Gagliardi*[,1,2,3], Antonio Luca Alfeo[1,4], Vincenzo Catrambone[1,4], Mario G.C.A. Cimino[1,4], Maarten De Vos[2], and Gaetano Valenza[1,4]

***Abstract*— Emotion recognition from electro-physiological signals is an important research topic in multiple scientific domains. While a multimodal input may lead to additional information that increases emotion recognition performance, an optimal processing pipeline for such a vectorial input is yet undefined. Moreover, the algorithm performance often compromises between the ability to generalize over an emotional dimension and the explainability associated with its recognition accuracy. This study proposes a novel explainable artificial intelligence architecture for a 9-level valence recognition from electroencephalographic (EEG) and electrocardiographic (ECG) signals. Synchronous EEG-ECG information are combined to derive vectorial brain-heart interplay features, which are rearranged in a sparse matrix (image) and then classified through an explainable convolutional neural network. The proposed architecture is tested on the publicly available MAHNOB dataset also against the use of vectorial EEG input. Results, also expressed in terms of confusion matrices, outperform the current state of the art, especially in terms of recognition accuracy. In conclusion, we demonstrate the effectiveness of the proposed approach embedding multimodal brain-heart dynamics in an explainable fashion.**

I. Introduction

Affective computing focuses on the recognition of emotional and mental states by analyzing heterogeneous information with artificial intelligence (AI) algorithms. Inputs may include audio, video, and electro-physiological signals. In this context, commonly used signals include electroencephalographic (EEG) [1], electrocardiographic (ECG), electromyographic, and electrodermal activity signals [2].

It has been acknowledged that the use of multimodal and multidimensional input may improve emotion recognition accuracy [3], [4]. Indeed, recent studies highlighted the crucial role of the continuous, directional interaction between the central nervous system and the autonomic nervous system [6]; such an interaction is generally referred to functional brain-heart interplay (BHI) [5], [6]. BHI has been exploited to statistically characterize emotional processing [6] in accordance with the so-called circumplex model of affect [7], which is a two-dimensional model describing an emotion through a combination of valence and arousal. While the valence dimension accounts for the pleasantness of an emotion, the arousal dimension quantifies its degree of activation (intensity) [3], [7].

From a methodological perspective, several AI methodologies have been successfully applied to classify emotion perception from multimodal electrophysiological signals: e.g., convolutional neural networks (CNN), both with standard [8] or with transfer-learning methodologies [9]. These networks extract features from input data with convolutions and classify them with dense neural network classifiers [8] or long-short term memory [10] units.

Nevertheless, neuroscientists and domain experts need to understand, and thus validate, the reasoning behind an automatic emotion recognition approach while employing its predictions [11], and would like to infer on the underlying physiological processes in a data-driven fashion. The deep neural network model works as a black box and cannot be easily interpreted unless the model is built especially to exploit speicifc (e.g., neurophysiological) knowledge [12]–[14]. EXplainable Artificial Intelligence (XAI) approaches address this limitation by providing explanations for neural network models. The most used explanation form is the *feature importance* providing the rank of all data attributes by considering the importance of each attribute to the classification. Considering CNN, the most direct way to explore visual patterns hidden inside the neural unit is filter visualization [15]. Widely used methods to this end compute the gradients of the score of a given CNN unit with respect to the input image. These XAI algorithms are equivalent to the features importance ones. In fact, if we consider each input data point as an attribute (or feature), these methods provide the rank of the relative importance of all the input attribute for the prediction. In the affective computing framework, while the role of BHI dynamics in emotional processing has been recognized, to the best of our knowledge no explainable AI algorithm has exploited such a multidimensional input yet in emotion recognition tasks. To overcome this limitation, this study proposes an XAI CNN-based approach for the fine-grained identification of 9 different emotional valence levels,

[1] Department of Information Engineering, University of Pisa, Pisa, Italy.
[2] Dept. of Electrical Engineering, KU Leuven, Belgium
[3] Dept. of Information Engineering, University of Florence, Italy
[4] Bioengineering & Robotics Research Center E. Piaggio, School of Engineering, University of Pisa, Pisa, Italy.
∗ Correspondence to: guido.gagliardi@phd.unipi.it

Work partially supported by (i) the EU Commission - Horizon 2020 Program under GA 101017727 of the project "EXPERIENCE", (ii) the Italian Ministry of Education and Research (MIUR) in the framework of the FoReLab project (Departments of Excellence), (iii) PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme. (iv) the Italian Ministry of University and Research (MUR), in the framework of the "Reasoning" project, PRIN 2020 LS Programme, Project number 2493 04-11-2021. This study uses the MAHNOB Database collected by Prof. Pantic and the iBUG group at Imperial College London and in part collected in collaboration with Prof. Pun and his team at University of Geneva, in the scope of the MAHNOB project financially supported by the ERC under the European Community's 7th Framework programme (FP7/2007-2013)/ERC starting GA 203143.

taking multimodal BHI features as input. More specifically, we propose a spatial features rearrangement of BHI features in accordance with neurophysiology knowledge; then, an XAI gradient-weighted class activation mapping (Grad-CAM) [16] visualization technology is exploited to identify brain regions that are mainly involved in the valence recognition, i.e., regions associated with discriminant features by the underlying CNN mechanism.

The proposed BHI-XAI-CNN approach is tested on publicly available data from the MAHNOB-HCI dataset [17], which is a benchmark dataset for multimodal emotion recognition. Achieved performances are compared against state-of-the-art architectures processing the MAHNOB data for valence recognition, as well as against an EEG-XAI-CNN approach whose multidimensional input comprises EEG power information exclusively.

II. Materials and Methods

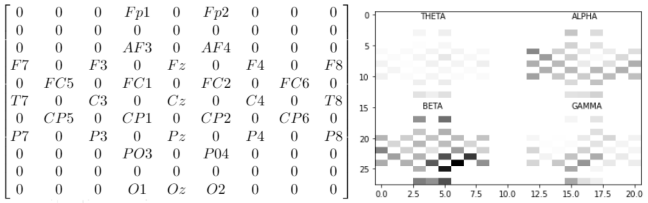
A. Experimental Dataset

The MAHNOB-HCI dataset [17] (available at *https://mahnob-db.eu/hci-tagging/*) comprises recordings gathered from 27 healthy volunteers (age range, 19–40 years; 15 females) undergoing emotional video elicitation. In this study, physiological signals as 32-channel EEG and 3-lead ECG, both sampled at $256Hz$, were retained for further analyses. The experimental data comprise EEG and ECG signals recorded during 20 video trials, extracted from movies, with a duration of 35 to 117 s. Participants were asked to provide subjective ratings on their perceived emotional experience through a 0-8 Likert-type scale for the valence and arousal dimensions. Consequently, each subject associated an emotional elicitation with one out of 9 valence levels. Each participant signed a consent form, and the experiment was approved by the local ethical committee. Further details on the experimental design and data acquisiton can be found in [17].

B. Features Processing and features extraction

1) Unimodal EEG features: The EEG preprocessing procedure aimed to obtain artifact-free signals and comprised frequency filtering, large artifact removal (e.g., eye movements, cardiac-field artifact), interpolation of contaminated channels, and average re-referencing [18]. An extensive description of the applied preprocessing procedure can be found in [6].

Briefly, the EEG power spectral density (PSD) was extracted through Welch's method with a Hanning window. PSD time series were integrated every 2 seconds, without overlapping, within canonical four frequency bands, namely: $\theta \in (4-8]Hz$, $\alpha \in (8-12]Hz$, $\beta \in (12-30]Hz$, and $\gamma \in [30-45]Hz$. EEG and ECG data included in this study referred to 20 videos, and each of which was watched by 27 subjects. Each video (with different length) was segmented in non-overlapping $2s$ windows, and EEG power and BHI features (as described below) were then derived from each of these segments. Consequently, a total number of 20.055 instances, each instance characterized by 32 (channels) $\times$ 4 (frequency bands) features, was available for the explanable classification procedure.



(a) (b)

Fig. 1: (a) Matrix rearrangement of the input features in a 11x9 matrix; (b) Interpretation of the 2x2 block matrix as a grayscale image.

2) Multimodal Brain-Heart Interplay features: Quantitative BHI features were estimated using a synthetic data generation model, which has been extensively described in [6], [19]. At a glance, the model equations were developed to generate synthetic brain and heart rate series. On the brain side, multiple oscillators may generate synthetic EEG signals whose amplitude is modeled as a first-order autoregressive with exogenous cardiac processes; the exogenous term then refers to the functional heart–to-brain interplay. On the cardiovascular side, an integral pulse frequency modulation model generates synthetic heart rate variability (HRV) series, whose embedded parameters are modulated by EEG activity and then quantify the functional brain-to-heart interaction. In this study, features of brain-to-heart interplay defined in the high-frequency range of heartbeat dynamics (between $0.15Hz$ and $0.4Hz$) were retained for further classification purposes.

Time-resolved information of directional BHI series were condensed through the median calculated within non-overlapping 2-second window, independently for each channel. The same calculation has been performed to condense the time-resolved information of EEG power series, thus resulting in the same cardinality of the BHI feature set.

C. Arrangement of spatial features into an image

The 32 input features (i.e. one for each EEG channel) are arranged in a sparse matrix following the standard 10-20 EEG electrode placement scheme (see Fig. 1.a). Elements not corresponding to an electrode location are set to zero. A single matrix like the one represented in Fig. 1.a is built for each of the four frequency bands (i.e., θ, α, β, and γ) and, subsequently, they are combined together to form a 2×2 block matrix (see Fig. 1.b). Between each block, a padding space of 3 null elements is been placed to avoid overlapping between electrodes of different blocks during convolution. Eventually, the matrix is interpreted as a gray-scale image (Fig. 1.b) to be fed as a single instance to the CNN algorithm.

D. CNN Model

The CNN architecture implemented in this study takes the 2×2 block matrix of shape $25 \times 21 \times 1$ as input and processes it with two convolutional layers, with depths of 32 and 64 respectively, and a rectified linear unit (ReLU) activation function. The size of the convolutional filters was set to 3×3 in order to prevent overlapping between electrodes location of different frequency blocks in the 2×2 block

matrix, as mentioned above. Then, a flattening layer was inserted before the final two dense fully connected layers, which included 512 and 9 neurons, respectively. The last level included 9 neurons and a softmax activation function in accordance with the number of classes to be recognized. To prevent overfitting during training, a dropout level has been added after the flattening level.

E. Grad-CAM Saliency Maps

Grad-CAM algorithm [16] provides visual explanations for CNN decisions in classification tasks. Given an input image to a trained model, Grad-CAM produces a coarse localization map that highlights important regions of the image according to the model decision process. Contextualizing, here Grad-CAM highlights the physiological correlates for the specific valence-level predicted.

In this study, the processing of Grad-CAM explanations mainly involved two phases: first, a smoothing filter was applied to the generated activation map to filter out outliers, i.e., pixels that can have intense colour but are isolated and located in regions of low intensity; Second, the matrix input scheme shown in Fig. 1.b was overlaid on the smoothed activation map to isolate the original physiological correlates and provide a clearer representation of the brain activation map. To do so, each null pixel of the input image is manually set as zero.

III. Results

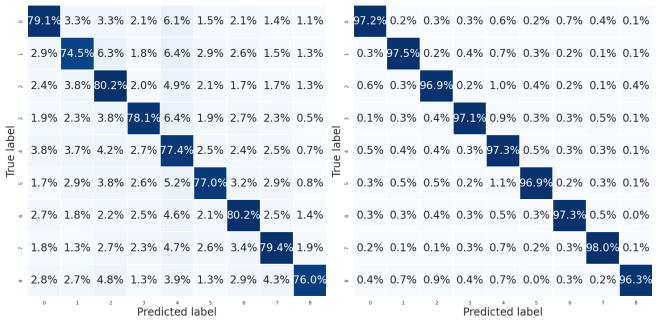
The proposed CNN model has been trained both with BHI features and with EEG power features (for comparison reason) to perform a 9-level valence emotion recognition. The model was tested on a subject-independent 10-fold montecarlo cross-validation scheme.

Tab. I illustrates the aggregated results expressed in terms of average accuracy, precision, recall, and F1 score; the last three metrics were obtained through weighted average [20]. Such results are provided both in case of EEG power input and BHI input. Significantly higher classification performance ($\approx 20\%$) are with the BHI features, and this is consistently observed for all the evaluation metrics.

TABLE I: Average % accuracy, precision, recall, and F1 score during cross-validation with the proposed CNN architecture considering EEG power and BHI-related features over 9 levels valence classification.

Features	Accuracy	Precision	Recall	F1-score
EEG power	78.02%	78.45%	78.09%	78.09%
BHI	97.20%	97.08%	97.12%	97.12%

Fig. 2 shows the average classification results in the case of EEG power and BHI inputs expressed in terms of a precision-confusion matrix. Each term $c_{i,j}$ of the confusion matrix C of dimension 9×9 is equal to the number of input samples belonging to level i and predicted as level j. Levels range from 0 to 8. Eventually, for readability purposes, each value $c_{i,j}$ has been divided by the support of the i-level. The BHI-based architecture achieves its maximum performances with $98.0\% \pm 1.1\%$ precision when predicting level 7, whereas



(a) EEG valence Classification (b) BHI valence Classification

Fig. 2: Precision Confusion Matrices. Valence classification.

the EEG-based one achieves $79.4\% \pm 4.5\%$. The EEG-based architecture achieves its best performance predicting level 6 with $80.2\% \pm 4.86\%$, being outperformed by the BHI-based one which achieves $87.3\% \pm 0.4\%$.

Tab. II shows a comparison analysis between the aforementioned results and previous findings in the literature in valence recognition using the MAHNOB-HCI dataset with different modalities as input. Zhang *et al.* [8] propose a multimodal-CNN approach for emotion recognition based on EEG, electromyography, and electrooculogram, obtaining high accuracy (90.50%) and underlining the effectiveness of considering multiple physiological signals, but it does not take into consideration cardiovascular dynamics. Siddhart *et al.* [10] and Huang *et al.* [9] consider both EEG signals obtaining similar results to our CNN EEG-based approach, Huang *et al.* also consider facial features obtaining lower results in comparison with the others. Those results highlight the impact of introducing BHI features for emotion classification.

TABLE II: Comparison with the state-of-the-art approaches for valence levels recognition on the same dataset.

Method	modality	# levels	Accuracy
Huang [9]	EEG, Facial f.	2	75.21%
Siddharth [10]	EEG	2	80.77%±0.77%
Zhang [8]	EEG, EOG, EMG	2	90.50%
CNN (this study)	EEG	**9**	78.02%±1.55%
CNN (this study)	BHI (EEG, ECG)	**9**	**97.38%±0.50%**

The explanations using the Grad-CAM algorithm were obtained for each input sample and then aggregated on the same predicted class, considering their mean. The resulting global explanations are expressed as topographical maps of features importance (Fig. 3). Dark regions of the image refer to features that are more important for predicting valence, while lighter regions refer to features that are less important, i.e. in dark regions, brain-heart interactions in a specific frequency band have a high influence on the classification results, while in lighter regions they have a low impact. Moreover, in wide dark regions, the emotional information is spread all over the scalp, while in restricted ones the model focuses only on a small portion of the features neglecting the others. In particular, Fig. 3 shows aggregated explanations for level 4. In the α and θ bands, we have large dark regions

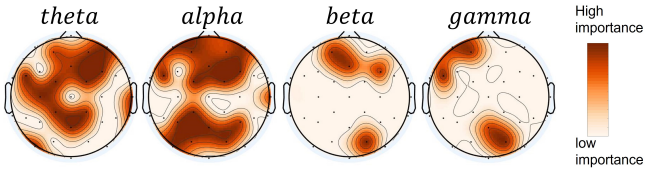


Fig. 3: Topographical representation of the feature importance for predicted valence level 4.

targeting the frontal, temporal and occipital lobes. In β and γ, the dark regions are rather restricted, which means that emotional information is concentrated only in certain parts of the frontal and occipital lobes. Moreover, by comparing the bands, it becomes clear that the BHI features in θ and α are chosen by the model to be more relevant than β and γ.

IV. Discussion

In this study, a novel explainable AI architecture for a 9-level fine-grained recognition of emotional valence has been developed. The architecture employs a CNN with an image-based features rearrangement scheme for measurements gathered from a person's scalp, as well as a Grad-CAM-based methodology to provide explanations for the classification.

The proposed methodology has been tested on the publicly-available MAHNOB-HCI dataset, also by comparing EEG power and BHI feature sets. While the proposed architecture provided satisfactory performance metrics with both feature sets (see Tab. I and Fig. 2), the use of BHI features outperforms the EEG-based approach. Such a significant improvement of approximately 20% is in line with current psychophysiology knowledge highlighting the crucial role of brain-heart interplay in emotional processing [6]. Moreover, this is in agreement with previous findings suggesting that the use of a multimodal and multidimensional input for emotion recognition [3], [4].

The role of brain-heart interaction in the recognition of emotions is also underlined by the comparison between the proposed architecture and the state-of-the-art approaches available in the literature (see Tab. II). These results suggest that, when only the EEG features are used, our model shows similar performance to the literature, but performs a 9-level classification rather than a binary classification; furthermore, the model can effectively leverage BHI and outperforms other approaches that also use multimodal features.

The Grad-CAM visualization technique was applied to extract the prediction explanations from the model in combination with the proposed image rearrangement for the EEG-related measurements, allowing the system to provide easy-to-read explanations that highlight the CNN architecture decision-making process (Fig.3).

Future works will be directed toward testing on additional emotional dimensions, such as the arousal dimension and discrete models of emotion; further BHI features (e.g., heart-to-brain direction, multiple HRV frequency bands) will be tested as well, also considering more challenging cross-validation setups, e.g. leave-one-out. From the XAI point of view, other ways to extract explanations will be also taken into consideration looking in the direction of global or class-wise explanations.

V. Conclusion

Multimodal physiological signals allowing for the calculation of BHI features through ad hoc neurophysiological modelling are suitable inputs for a CNN-based, explainable AI architecture for a high-performance multi-class classification problem. The study highlights the relevance of embedding psychophysiology knowledge (in this case BHI dynamics) in an XAI framework for an automatic emotion recognition task.

References

[1] Suhaimi, N. et al "EEG-based emotion recognition: A state-of-the-art review of current trends and opportunities" Computational intelligence and neuroscience, 2020.

[2] Egger, M. et al "Emotion recognition from physiological signal analysis: A review" Electronic Notes in Theoretical Computer Science, 343, 35-55.

[3] Wang, Y. et al "A systematic review on affective computing: Emotion models, databases, and recent advances" Information Fusion.

[4] Zhang, S. et al "Learning affective features with a hybrid deep model for audio–visual emotion recognition" IEEE Transactions on Circuits and Systems for Video Technology, 28(10), 3030-3043.

[5] Catrambone, V. et al. "Functional linear and nonlinear brain–heart interplay during emotional video elicitation: a maximum information coefficient study." Entropy 21.9 (2019): 892.

[6] Candia-Rivera, D. et al. "Cardiac sympathetic-vagal activity initiates a functional brain–body response to emotional arousal." Proceedings of the National Academy of Sciences 119.21 (2022): e2119599119.

[7] Russell, J. et al "A circumplex model of affect." Journal of personality and social psychology 39.6 (1980): 1161.

[8] Zhang, Y. et al "Emotion recognition using heterogeneous convolutional neural networks combined with multimodal factorized bilinear pooling" Biomedical Signal Processing and Control, 77, 103877.

[9] Huang, Y. et al "Combining facial expressions and electroencephalography to enhance emotion recognition" Future Internet, 11(5), 105.

[10] Siddharth, S. et al "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing" IEEE Transactions on Affective Computing.

[11] Zucco, C., et al. "Explainable sentiment analysis with applications in medicine." 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2018.

[12] Alfeo, A. L. et al. "Concept-wise granular computing for explainable artificial intelligence." Granular Computing (2022): 1-12.

[13] Alfeo, A. L. et al. "Measuring Physical Activity of Older Adults via Smartwatch and Stigmergic Receptive Fields". Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017) (pp. 724-730). PRT.

[14] Alfeo, A. L. et al. "Improving an Ensemble of Neural Networks via a Novel Multi-class Decomposition Schema." Proceedings of the International Conference on E-Business and Telecommunication Networks, (2022).

[15] Zhang, Q. S. et al "Visual interpretability for deep learning: a survey" Frontiers of Information Technology and Electronic Engineering", 19(1), 27-39.

[16] Selvaraju, R. et al "Grad-cam: Visual explanations from deep networks via gradient-based localization" In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[17] Soleymani, M. et al "A multimodal database for affect recognition and implicit tagging" IEEE transactions on affective computing, 3(1), 42-55.

[18] Candia-Rivera, D. et al. "The role of electroencephalography electrical reference in the assessment of functional brain–heart interplay: From methodology to user guidelines." Journal of Neuroscience Methods 360 (2021): 109269.

[19] Catrambone, V., et al. "Time-resolved directional brain–heart interplay measurement through synthetic data generation models" Annals of biomedical engineering 47.6 (2019): 1479-1489.

[20] Grandini, M. et al. "Metrics for multi-class classification: an overview." arXiv preprint arXiv:2008.05756 (2020).