Machine Learning (2025) 114:273
https://doi.org/10.1007/s10994-025-06921-y

Model-driven validation of visual explanations for multimodal emotion recognition

Guido Gagliardi[1,2,3] · Antonio Luca Alfeo[4] · Vincenzo Catrambone[1,5] · Mario G. C. A. Cimino[1,5] · Maarten De Vos[2,6] · Gaetano Valenza[1,5]

Received: 25 March 2025 / Revised: 16 September 2025 / Accepted: 13 October 2025 /
Published online: 10 November 2025

Abstract
AI-based emotion recognition approaches may benefit from the integration of multimodal data, but their explainability and validation is still a critical challenge. Indeed, the limited neurophysiological understanding of novel multimodal features, e.g. brain-heart interaction, can be insufficient to assess whether the AI-extracted physiological insights (i.e., the model explanations) accurately reflect the real underlying physiological processes. To validate the explanations obtained by an AI-based model in this context, we introduce a novel framework that autonomously identifies the optimal explanations for a black-box model used in emotion recognition. Our approach leverages a convolutional neural network to process BHI features, which are derived from EEG and HRV data and rearranged as images. A model-agnostic methodology is employed to extract local explanations, which are then dynamically evaluated to select the most accurate for representing specific emotional states. The effectiveness of the proposed framework is evaluated across multiple classification tasks, including up to 9-level arousal and valence emotion classification, as well as nine discrete emotions classification, using the MAHNOB-HCI and DEAP datasets. The system achieved remarkable accuracy levels, consistently reaching approximately 97–98% across all tasks. Furthermore, our dynamic selection framework revealed that Integrated Gradients outperformed other state-of-the-art explainable AI approaches in reliably capturing global explanations.

Keywords Affective computing · Brain-heart interplay · EEG · HRV · EXplainable artificial intelligence · Integrated gradients

Editors: Riccardo Guidotti, Anna Monreale, Dino Pedreschi.

Extended author information available on the last page of the article

1 Introduction

Emotion recognition has gained significant attention in the fields of affective computing and computational neuroscience due to its potential applications in various domains such as healthcare, marketing, and human-computer interaction (Calvo & D'Mello, 2014). Over the years, several techniques have been proposed to distinguish human emotions from the analysis of physiological signals (Saganowski et al., 2022; Gagliardi et al., 2023b, 2023a). To enhance recognition performance while minimizing the burden of developing an emotion recognition model from scratch, artificial intelligence (AI) approaches are increasingly being exploited.

Quantitative EEG has been widely used to identify emotional states from physiological signals in previous studies (Khosrowabadi et al., 2010; Calvo & D'Mello, 2014), while recent approaches suggest that the combined information of EEG and cardiovascular dynamics, has immense potential for emotion recognition (Gagliardi et al., 2023b).

Emotional processing and regulation have long been associated with the dynamics of the heartbeat and the autonomic nervous system (ANS). The ANS, which comprises the sympathetic and parasympathetic branches, is responsible for regulating various involuntary physiological processes, including cardiovascular function, in response to emotional stimuli (Thayer et al., 2012). Heart rate variability (HRV), an index of the beat-to-beat variation in heart rate, is considered a reliable marker of ANS activity, reflecting the balance between the sympathetic and parasympathetic systems (Valenza et al., 2018).

In clinical practice and affective computing, there is a pressing need for highly accurate and precise emotion recognition models. Current systems, primarily relying on EEG data, cannot often provide detailed and granular insights into emotional states, typically distinguishing only between binary levels of arousal and valence. To improve precision and granularity, it is essential to develop models capable of identifying a broader range of emotional states (Gagliardi et al., 2023b).

When combined with AI algorithms, we hypothesise that a quantitative assessment of functional brain-heart interactions (BHI) in the frequency bands commonly used in EEG and HRV analysis will be more effective in discerning emotions than using EEG alone.

The integration of multimodal information, particularly through the combination of heart rate variability (HRV) and electroencephalogram (EEG) signals, has emerged as a rapidly growing area of research in emotion recognition (Pillalamarri & Shanmugam, 2025; Wei et al., 2023; Torres-Valencia et al., 2014). For instance, in Barra et al. (2017), the authors employed peak features extracted from electrocardiogram (ECG) data together with spectral features from EEG, demonstrating notable improvements in recognition performance. Similarly, Guo et al. (2020) introduced a voting-based strategy that leverages EEG, cardiac, and speech features to determine emotions based on the most informative modality. Extending this line of work, Kumar and Kumar (2025) explored the joint use of EEG, ECG, and Galvanic skin responses (GSR).

A broader review of multimodal approaches for emotion recognition in individuals with autism spectrum disorder is provided in Said et al. (2026), which emphasizes the importance of integrating speech, physiological, and visual signals within comprehensive multimodal frameworks. More recently, deep-learning architectures have become the dominant strategy for modality fusion. For example, Lu et al. (2025) introduced a deep learning model designed for competitive environments, fusing EEG, ECG, and facial video recordings to

predict valence and arousal. In Lian et al. (2025), facial images and EEG signals were fused at the decision level to enhance feature complementarity and robustness to interference. Likewise, Kim (2025) proposed an incremental quality-assessment strategy for EEG, EMG, and ECG, while (Wang and Wang, 2025) developed an attention-based CNN with GRU layers to fuse EEG and ECG signals. Finally, Hosseini et al. (2024) presented a comprehensive deep model combining CRNNs for facial features, ConvLSTMs for spatio-temporal video data, and Conv1D networks for physiological signals (EEG, EOG, ECG, GSR).

Overall, deep-learning-based modality fusion constitutes a powerful and effective technique for leveraging complementary information across diverse modalities in complex tasks such as emotion recognition. Nevertheless, these architectures often reduce interpretability, making it difficult to determine precisely how modalities are fused and which specific features are being exploited by the network.

In summary, these studies underscore the significance of incorporating brain-heart information into emotion recognition systems; however, they overlook the consideration of the directional flow of information, whether it is from brain-to-heart or heart-to-brain, within this context (Sedehi et al., 2025).

Acknowledging the established role of directional and functional BHI in emotions (Candia-Rivera et al., 2022), we observe that while these features hold significant promise, their complexity and novelty have impeded their full integration into explainable emotion recognition systems. Specifically, the limited neurophysiological understanding of these intricate interactions hinders our ability to validate AI-derived explanations. To address this gap, we propose a novel approach to emotion recognition that leverages functional BHI data, aiming not only to enhance performance but also to advance the interpretability of these complex features.

Advancing recognition performance through complex feature inputs and black-box models compromises interpretability, thereby constraining the system's applicability in practical clinical environments. Within such high-stakes contexts, ensuring trust between end-users and AI systems is crucial. However, this trust cannot be realized if clinicians receive suboptimal explanations that inadequately highlight the model's decision-making processes.

This issue is emphasized because different XAI methodologies may produce divergent explanations, such as different rankings of the importance of the features. Considering the relative novelty of such features, it is challenging for the current literature to physiologically validate such importance ranks. Also, these rankings may vary in their representation of the true influence of features on the model's output, and thus, their *fidelity* to the model. Furthermore, global explanations offer a broad overview of how the model makes local decisions based on specific input elements. Consequently, we need to evaluate the extent to which different global rankings correspond to each specific decision made by the model on each input sample, that is, the global explanation *sensitivity*.

In many biomedical applications, such as emotion recognition based on brain-heart interactions domain experts may lack the means to directly validate the model's explanations, due to the complexity and limited understanding of the underlying physiological mechanisms. In such cases, relying solely on human interpretability is insufficient. Our work addresses this gap by proposing a model-driven approach that evaluates the reliability of explanations using quantitative quality metrics. By assessing fidelity and sensitivity, we aim to identify the most trustworthy explanation method, thus increasing confidence in the model's decisions even when human validation is not possible.

In this study, we extracted global physiological insights for each emotional state from the neural network using different state-of-the-art explainable AI algorithms such as Integrated Gradients, DeepLIFT, Expected Gradients, and Grad-CAM. Subsequently, we introduced a novel methodology to assess and visualize the quality of the explanations obtained, with a focus on fidelity and sensitivity. Ultimately, we determined the most suitable explanation for each emotional category.

Our methodology streamlines the interpretation process by automatically identifying the best global explanation, thus ensuring that domain experts are not burdened with validating suboptimal or biased explanations. Rather than bypassing human expertise, this approach provides clinicians and researchers with a vetted, high-quality explanation to review, ultimately strengthening the validation process.

By dynamically selecting and examining the most suitable global explanations from a range of local explainers, experts can confirm alignment with established neurophysiological knowledge and uphold the stringent standards demanded by critical applications.

To summarize, the main contributions of this study consist of:

- Employing BHI features for emotion recognition via neural networks, distinguishing up to 9 arousal and valence levels and nine categorical emotions.
- Using a novel explainable AI framework, we extract global physiological insights from the neural network and select the explanation quality, thereby boosting the interpretability and clinical application of the system.

2 Related work

2.1 Functional brain-heart interplay and emotional states

Emotion regulation, which involves the activation of neural circuits to modulate emotional responses, has been linked to the interaction between the central autonomic network (CAN) and the ANS (Valenza et al., 2020, 2019). CAN includes brain structures such as the amygdala, insula, prefrontal cortex, and hypothalamus, which play a pivotal role in the processing of emotional information and orchestrating autonomic responses to emotions (Valenza et al., 2020, 2019).

The amygdala, in particular, is a key component of CAN and has been implicated in the detection and evaluation of important emotional stimuli, as well as the initiation of autonomic responses to these stimuli (LeDoux, 2000). In addition, the amygdala is known to have direct and indirect connections with other brain regions within CAN, such as the insula and prefrontal cortex, which are involved in the integration of sensory information, emotional awareness, and the modulation of ANS activity during emotional regulation (Thayer et al., 2012).

Furthermore, the interaction between heartbeat dynamics and emotional processing is not unidirectional. While the CAN modulates ANS activity in response to emotions, the ANS provides feedback to the central nervous system (CNS) through afferent biochemical, mechanical, and electrical pathways, leading to the continuous adjustment of emotional regulation processes (Benarroch, 2008). This bidirectional communication between the CNS

and the ANS is commonly defined as functional brain-heart interplay (BHI) and highlights the intricate relationship between emotional processing, regulation, and heartbeat dynamics.

The estimation of functional BHI involves several methodological and technical challenges, such as nonlinearity, multimodal (in most cases EEG and HRV) and multivariate (EEG channels) variables, directionality (the brain-to-heart interaction does not correspond to the heart-to-brain interaction), and various dynamical features (Catrambone & Valenza, 2021).

The quantitative assessment of functional BHI involves several methodological and technical challenges (Catrambone & Valenza, 2021). In fact, methodological approaches for the quantification of functional BHI may focus on the directionality of the phenomenon (Catrambone et al., 2021), as well as on non-linear interactions (Faes et al., 2015), and a brain-wise assessment (i.e. estimates BHI at a whole-brain level) (Catrambone & Valenza, 2023); additionally, an ad-hoc physiologically-inspired model based on the BHI synthetic data generation model (SDG) has been developed (Catrambone, 2019).

2.2 Explainable AI approaches for emotion recognition

A recent survey (Suhaimi et al., 2020) highlights a wide variety of classifiers being used in emotion recognition, with convolutional neural networks (CNN), support vector machine (SVM), random forest (RF), K-nearest neighbour (KNN) being the most promising, achieving recognition performances above 97%.

AI models like CNN operate as black boxes, making it challenging for domain experts to interpret and trust their outcomes. To this end, the field of eXplainable AI (XAI) aims to develop algorithms that provide insights into the decision-making processes of AI models (Alfeo et al., 2022; Gagliardi et al., 2023c).

Transformer-based attention mechanisms have been recently utilized to selectively emphasize and amalgamate complementary input features within emotion recognition tasks (Di et al., 2020). This approach permits the fusion layer to effectively integrate information from diverse sources (EEG + video) at an intermediate phase, thereby improving the model's capacity to acquire a unified representation of the input data. Furthermore, the attention mechanism enables a neural network to learn adaptable fusion weights across multiple modalities, resulting in improved multimodal fusion and emotion identification (Ahmed et al., 2023).

Typically, these methodologies are employed to integrate EEG signals with facial video data (Choi et al., 2020), leveraging attention mechanisms that enable the model to dynamically assess the significance of features across each modality, thus facilitating the selective integration of the most pertinent information. Analogous mechanisms have been applied to combine data from EEG, unprocessed eye images, and ocular movements (Lan et al., 2020). This methodology further permits the evaluation of each modality's contribution to emotion recognition by rendering attention weights visible. By offering a level of interpretability that elucidates the interaction between different input features or modalities during the construction of their fusion representation, these mechanisms confer a mechanistic insight into the model's behaviour, such that they elucidate a particular aspect of the model by exploiting its internal structure. Hence, these approaches furnish a model-dependent and non-comprehensive elucidation of the model's decision-making system.

Post-hoc XAI algorithms, instead, are model-agnostic methods that explain the decision-making approach of a black-box model, relying only on the model's input–output behaviour. Those approaches are generally divided into local XAI algorithms, which consider a specific input query and its respective output, i.e. they explain why the model predicted an output when a specific input was provided; and global XAI algorithms instead which explain black-box models by explaining their decision-making among a group of different input examples, such as the target classes.

When it comes to CNNs, popular local XAI algorithms involve computing the gradients of the score attributed to the input image. These XAI algorithms treat each pixel as an input feature and rank them based on their relative importance for recognition. For instance, gradient-weighted class activation mapping has proven valuable in identifying brain regions that play a crucial role in CNN-based emotion recognition (Selvaraju et al., 2017).

In our recent studies, we proved the capability of CNN models to recognize emotions using EEG signals in subject-dependent binary classification tasks (Gagliardi et al., 2023a). We gained preliminary insights about the superior performance that could be obtained by such CNN models if trained with brain-to-heart high-frequency features compared to EEG features in a valence classification task (Gagliardi et al., 2023b), and we extracted class-wise global explanations using a traditional XAI method, i.e., GradCAM, to explain such CNN.

The capability of gradient-based methodologies, such as GradCAM (Selvaraju et al., 2020), in visualizing neural network activation maps encounters challenges posed by the issue of saturation (Sundararajan et al., 2017).

In the neural network context, saturation refers to the state in which a neuron predominantly outputs values close to the asymptotic ends of the activation function. Hence, since their output value is predominantly constant, their gradients are always 0. Consequently, traditional methods for gradient computation are progressively being surpassed by approaches that rely on the accumulation of gradients along a path from a designated reference point to the input image, such as integrated gradients (Sundararajan et al., 2017) or expected gradients (Erion et al., 2021), or by approaches that rely on the difference between the input elements and the reference points, such as DeepLIFT (Shrikumar et al., 2017).

Despite their effectiveness, the accurate determination of the reference points in both cases remains strictly related to the specific application, and methodologies for generalizing input-specific activation maps across distinct classes (i.e., transforming local explanations into global explanations) are still undergoing investigation. Furthermore, since most of the properties of such explainable AI algorithms are axiomatic (Sundararajan et al., 2017), a practical and numerical comparison between different explanation ranks extracted by different algorithms is troublesome. For this reason, approaches to estimate the quality of local explanations have been recently proposed.

These approaches measure how well a local explanation describes the decision-making system of a black-box model for a given input, hence they measure the degree of *fidelity* of the rank to the black-box. For example, insertion and deletion tests (Petsiuk et al., 2018) measure the drop or the increase in the model confidence when inserting or deleting important input features following the rank provided by a local explanation algorithm. That said, standardized approaches to estimate the best global explanation rank are not yet established.

Another critical aspect addressed by these approaches is the *sensitivity* of explanations, i.e., their susceptibility to variations when input samples undergo minor perturbations (Yeh et al., 2019). The underlying hypothesis is that an explanation should exhibit robustness to

small input perturbations, as the model's decision-making should ideally remain consistent for minor changes to an input sample.

Sensitivity is crucial in the context of global explanations. Global explanations typically rely on averaging local explanations from limited groups of samples. This aggregation raises concerns about how well global explanations handle small variations within these groups. It's essential to understand whether global explanations remain reliable and consistent when there are slight changes in the elements of these sample groups. This aspect is critical for assessing the applicability and trustworthiness of global explanation methods across different datasets and real-world scenarios.

That said, considering fidelity and sensitivity alone might not be sufficient. An approach that always predicts the same explanation could have very low sensitivity but be entirely unfaithful to the model's behavior. Conversely, a highly faithful explanation might be too sensitive to accurately represent the model's behavior across a group of samples. Therefore, it is essential to balance both fidelity and sensitivity to ensure that global explanations are reliable and representative of the model's overall decision-making process.

In this study, we proposed three tests to calculate and visualize the quality of explanations in terms of fidelity and sensitivity. Additionally, we developed an approach to dynamically select the optimal global explainer based on a combined optimization of these two metrics.

3 Materials

For this study, we used two publicly available datasets, namely MAHNOB-HCI and DEAP, both collecting physiological signals from a group of healthy volunteers who underwent emotional video elicitation. The evaluation of emotion perception was based on the widely recognized circumplex model of affect, which defines emotions in a two-dimensional space: arousal, associated with the intensity of the feeling, and valence, associated with the pleasantness of the feeling. Both arousal and valence were quantified using a Likert-type scale ranging from 0 to 9. Before the experiment, all participants provided informed consent, as documented in the original paper that presented the dataset.

3.1 DEAP dataset

The DEAP dataset comprises physiological data from 32 healthy participants (age range: 19–27 years; 16 females) (Koelstra et al., 2011). In this study, we focused on the 32-channel EEG signals sampled at 512 Hz. The DEAP dataset is freely accessible at https://www.eecs.qmul.ac.uk/mmv/datasets/deap/.

The experimental protocol involved 40 emotional video trials extracted from famous music videos. Following a 2-minute resting state, emotional videos of varying arousal and valence levels were presented for 60 seconds each (Koelstra et al., 2011).

3.2 MAHNOB-HCI dataset

The MAHNOB-HCI dataset includes physiological data from 27 healthy participants (age range: 19–40 years; 15 females) (Soleymani et al., 2011). The dataset consists of 32-channel

EEG signals sampled at 256 Hz, and varying trial sizes resulted from different numbers of volunteers participating.

The experimental protocol for the MAHNOB-HCI dataset encompassed 20 emotional video trials sourced from famous movies. After a 30-second initial resting state, emotional videos of different duration (ranging from 35 to 177 seconds) were presented, each eliciting distinct arousal and valence responses (Soleymani et al., 2011).

The study was approved by the local ethical committee of the University of Pisa.

3.3 EEG and ECG signal preprocessing

To obtain artifact-free, robust, and reliable signals for the classification task, we implemented a standard processing procedure that involved various steps. For the EEG preprocessing, these steps included frequency filtering in the range [0.5; 45] Hz, rejection of large artefacts through a wavelet-enhanced independent component analysis method, removal of eye movements and cardiac-field artifacts, interpolation of contaminated channels, and average re-referencing (Candia-Rivera et al., 2021). The processing procedure was carried out using MATLAB R2018b (MathWorks) and the Fieldtrip Toolbox (Oostenveld et al., 2011), further details on the preprocessing pipeline may be found in Candia-Rivera et al. (2022). To extract the EEG power spectral density (PSD), we employed Welch's method with a Hanning window. A sliding time window of 2 s in length with 50% overlap was used. The PSD time series were then integrated within four frequency bands: theta (θ) ranging from 4 to 8 Hz, alpha (α) ranging from 8 to 12 Hz, beta (β) ranging from 12 to 30 Hz, and gamma (γ) ranging from 30 to 45 Hz.

The ECG preprocessing involved: bandpass filtering in the range (0.5–45 Hz) using a Butterworth filter of order 4; heartbeats identification through an automated process (Candia-Rivera et al., 2021); misdetections' correction employing a point-process algorithm (Citi et al., 2012). The heart-rate variability (HRV) series were then constructed as inter-beat interval duration time course, and evenly resampled at 4 Hz using spline interpolation. The HRV series has been derived by the blood volume pulse in the DEAP dataset. The HRV PSD was computed using a smoothed pseudo-Wigner-Ville distribution (Orini et al., 2011), filtered in the low-frequency ($LF : [0.04 - 0.15]Hz$) and high-frequency ($HF : [0.15 - 0.4]Hz$) ranges to quantify the sympathovagal and parasympathetic activity from the ANS, respectively.

Further details on the EEG and ECG preprocessing procedures can be found in Candia-Rivera et al. (2022).

3.4 Brain heart interplay estimation

To quantify the functional BHI, we employed the physiologically-inspired SDG model (Catrambone, 2019). In this model, the EEG series represents a multi-oscillator model, where the amplitudes are generated by a first-order exogenous autoregressive (ARX) model. The exogenous term in the model denotes information transfer from the heart to the brain. Conversely, the HRV series represented as RR intervals are modeled using an extended integral and pulse frequency modulation (IPFM) model (Catrambone, 2019). The control function of sympathetic and vagal activity quantifies information transfer from the brain to the heart (Catrambone, 2019).

To estimate directional BHI, we calculated a BHI measure for each combination of EEG bands and HRV frequency components. The underlying concept of this model is that the electrophysiological signals of the brain and heart are not mutually exclusive. The term "functional coupling" is used to formalize these interactions. To illustrate, a positive value of $C_{\alpha \rightarrow HF}$ indicates a positive effect of the EEG alpha band on the HRV-PSD series in the HF range, indicating a linearly proportional increase. The directional BHI quantification is achieved by estimating the control function term for the IPFM model's heartbeat dynamics (i.e., brain-to-heart interaction) and the exogenous term of the ARX model representing the EEG dynamics (i.e., heart-to-brain communication).

An easy-to-use MATLAB implementation of the BHI model is freely available (Catrambone, 2019), and detailed descriptions of the inverse model formulation and derivation of the entire BHI biomarker suite can be found in previous studies (Catrambone, 2019, 2021; Candia-Rivera et al., 2022).

In this study, the directional BHI indices that were retained for further analysis are listed in Table 1.

4 Methods

This section details the state-of-the-art methods exploited to generate local explanations from a black-box model and our novel proposed methodology for extracting global explanations from such local explainers. In order to implement these approaches in BHI-based emotion recognition tasks using the aforementioned datasets, we initially preprocessed both EEG and ECG (HRV) signals and subsequently extracted the BHI features. Thereafter, we trained a simple CNN model, as referenced in "Appendices B. and C.", utilizing an image rearrangement of these features, as specified in "Appendix A.". Upon completion of training and subsequent parameter freezing, we applied our explanation methodology to elucidate the overarching decision-making processes of the model.

The overall schema highlighting the main contributions of the proposed approach can be found in Fig. 1.

4.1 Explanation methods

Our dynamic selection of the optimal global class-wise explanation explainer selection is model-agnostic and explainer-agnostic. Its first step requires selecting a set of local explainers to calculate the global explanations.

To extract class-wise global explanations from our network we exploited 4 different state-of-the-art local explanation algorithms: integrated gradients (IG) (Sundararajan et al., 2017), expected gradients (EG) (or GradientSHAP) (Erion et al., 2021), DeepLIFT (DLIFT) (Shrikumar et al., 2017) and GradCAM (GC) (Selvaraju et al., 2020).

IG and EG are gradient-based methods that assign importance scores to input features by integrating the gradients of the model's output over a straight path from a reference point

Table 1 BHI indices utilized in this study

Index	From	Band	To	Band
$Brain_j \rightarrow Heart_{B_C}$	Brain	$\theta, \alpha, \beta, \gamma$	Heart	LF, HF
$Heart_{B_C} \rightarrow Brain_j$	Heart	LF, HF	Brain	$\theta, \alpha, \beta, \gamma$

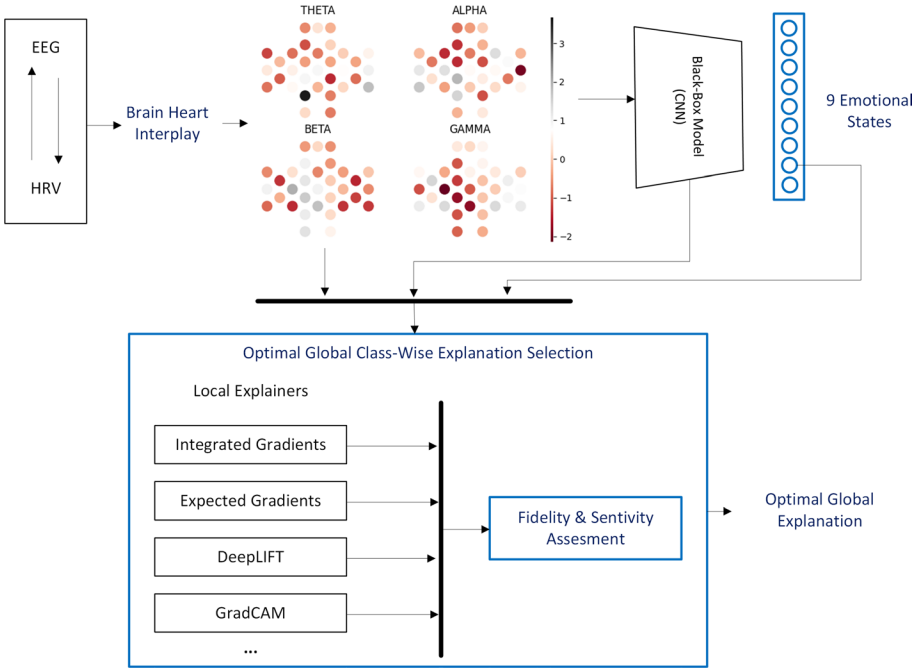

Fig. 1 Overview of the proposed architecture for emotion recognition and explanation optimization. The system integrates EEG and HRV data to capture the brain-heart interplay and feeds them into a CNN-based black box model to classify up to nine emotional states. The innovative aspect of this study, highlighted in blue, is a model-driven approach to select the optimal class-wise explanation of an emotion classifier employing multimodal physiological features of brain-heart interplay. The figure also illustrates an instance of reorganizing the features associated with the 32 EEG channels and four frequency bands, as proposed in Gagliardi et al. (2023a) for EEG signals and in Gagliardi et al. (2023b) for BHI signals

to the input element. IG allows for a user-defined reference point, providing flexibility and control in the explanation process. In contrast, EG autonomously computes the reference point by sampling from the data distribution, ensuring adaptability to different data contexts.

DLIFT compares the activation of each neuron against a reference point activation, attributing contributions to each input feature based on the difference w.r.t. to this reference point.

Finally, GC has been selected to compare our results to the one obtained by us in Gagliardi et al. (2023b). GC creates explanations by utilizing the gradients of a target class flowing into the final convolutional layer of a CNN to generate a coarse activation map.

In our case study, each input is a pixel within an image created by rearranging the features extracted from the physiological signals. This image-like representation allows us to associate different frequency bands and brain areas with specific regions or patterns within the image, facilitating the identification of relevant factors influencing the network's decision-making process.

To define the IG reference points (or baselines) we considered the IG formulation, in which the importance of the *i*-pixel of our input image, i.e. Fig. 1, is computed as:

$$\mathrm{IG}(x)_i = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F(x_i' + \alpha \times (x_i - x_i'))}{\partial x_i} \, d\alpha \tag{1}$$

where x_i represents the i-th feature of the input vector, x'_i is the i-th feature of the baseline input, and F is the model's output function. The integral is taken over the path from the baseline input x'_i to the actual input x_i, and $\frac{\partial F(x'_i+\alpha\times(x_i-x'_i))}{\partial x_i}$ represents the partial derivative of the model's output concerning the i-th image pixel.

IG integrates the gradients from a baseline (reference point) to the input to quantify how each feature influences the model's output. Using 0 as the reference point, representing a blank input image ensures that pixels unrelated to electrode positions receive zero importance scores. This approach captures the model's sensitivity to variations in signal dynamics, reflecting the importance of features based on changes in signal characteristics.

Testing samples were used as query elements to compute the local explanations and training samples were used to simulate the data distribution for the approaches that autonomously compute the reference points (i.e. for EG and DLIFT). Global class-wise explanations were derived from local explanations by integrating local instances and considering their mean for each emotional class.

4.2 Proposed global explanations selection framework

In this subsection, we introduce our proposed approach for selecting global explanations. The method involves generating a set of global explanations using various local explanation algorithms and evaluating their fidelity and sensitivity through three tests: insertion, deletion, and sensitivity.

For each global explanation, these tests produce curves that display the test results as their parameters vary, allowing for qualitative assessment by visualizing the curves. Quantitatively, the area under each curve is calculated to provide a numerical score representing the quality of the test. This combined approach of qualitative visualization and quantitative scoring ensures a comprehensive evaluation of the fidelity and sensitivity of each global explanation.

4.2.1 Insertion and deletion tests

We evaluated the fidelity of each class-wise global explanation by proposing two metrics. Our global deletion test involves ranking features based on their importance from the global explanation and systematically setting them to zero in the input, one at a time. After each deletion, the model's accuracy is recorded, producing a deletion curve that shows how accuracy changes as more important features are removed. The deletion score is calculated as the normalized area under the deletion curve (AUDC), with a lower AUDC indicating higher fidelity.

The global insertion test follows a similar procedure but in reverse: starting with a blank input, features are added in order of importance. The model's accuracy is recorded after each insertion, producing an insertion curve. The insertion score, calculated as the area under the insertion curve (AUIC), reflects the explanation's quality, with a higher AUIC indicating that the explanation correctly identifies the most important features for the model's prediction.

We used zero-deletion, which has both physiological and numerical justification since non-electrode positions are zero and the data is standardized to zero mean and unit variance.

4.2.2 Sensitivity test

The sensitivity of local explanation refers to the property of local explanation methods to produce similar explanations from similar input samples(Yeh et al., 2019). Typically, sensitivity is estimated by comparing explanations derived from minimally perturbed inputs or very similar input examples. In our global class-wise explanation context, sensitivity refers to the consistency of a global explanation with minimal changes in the distribution of input samples.

To measure the global class-wise explanation sensitivity, S, we compute the distance between the global explanation $\mathbb{E}[\phi(M, X)] = E$ and the global explanation $\mathbb{E}[\phi(M, X^*)] = E^*$, where $\phi(\cdot)$ is the local explainer and X^* is a partition of the input samples X from the same class. The sensitivity is calculated as:

$$S = \text{dist}(E, E^*)$$

Using cosine distance, S becomes:

$$S = 1 - \frac{E \cdot E^*}{||E|| \cdot ||E^*||}$$

The value of S can vary depending on the choice of partition X^* and its size. To correctly estimate S, we consider its expected value $\mathbb{E}[S]$ over possible partitions X^* of the same size. Sensitivity is expected to be higher for small partitions and lower for larger partitions. To reduce the impact of small partitions that differ significantly from X, we weight S by the inverse distance between $\mathbb{E}[X]$ and $\mathbb{E}[X^*]$:

$$S = \text{cosine}(E, E^*) \cdot (1 - \text{cosine}(\mathbb{E}[X], \mathbb{E}[X^*]))$$

This ensures that sensitivity decreases faster as the partition size increases, indicating that E better approximates all values E^*.

The sensitivity score can be obtained similarly to the insertion/deletion curves by calculating the area under the sensitivity curve (AUSC).

Exploiting a minimization strategy our final quality test (QT) can be written as:

$$QT = \alpha * AUDC + \beta * (1 - AUIC) + \gamma AUSC$$

where α, β and $\gamma \in \mathrm{R}^+$, $\alpha + \beta + \gamma = 1$, are the relative weights used to balance the importance of one test to the other. In our experiments, all the weights are settled to 0.33.

By utilizing this technique, we can dynamically select the best global explanations in terms of both fidelity and stability for each target class, enabling a more comprehensive interpretation of the model's behaviour and providing valuable information about the features and regions in the input that significantly influence the model's predictions (Sundararajan et al., 2017).

5 Results

This section presents the experimental results of the study. First, we evaluate the performance of the CNN across various emotion classification tasks. This experiment aims to identify the most effective classifier for analysis with the proposed XAI approach, as the best-performing architecture provides the most reliable representation of the underlying input–output relationships.

Subsequently, the selected architecture is integrated with the model-driven evaluation framework to determine the optimal class-wise explanations.

Experiments were conducted on a 24-core AMD Epyc 7402 CPU (2.8 GHz), 1 NVIDIA A100-SXM4 (40 GB GPU), and 238 GiB RAM. Gradient computations used the PyTorch 2.1.2 autograd functionality.

5.1 Emotion recognition performance

The classification tasks encompassed three multi-class classification tasks, using a 10-fold cross-validation validation schema, aiming to categorize emotions into nine levels of valence, arousal, and categorical emotions, and two binary classification tasks to discriminate between high and low levels of arousal and valence. Additional results with different validation schemas are provided in the appendix.

Table 2 presents the performances of the classification tasks. The BHI features, including HF-to-Brain, Brain-to-HF, LF-to-Brain, and Brain-to-LF, are evaluated in terms of accuracy (average and standard deviation).

The standard deviation of the results is consistently below 1%, despite the validation schema employed. Comparing the average accuracy obtained with BHI and PSD features, their difference is greater than three standard deviations, hence we have no statistical overlap between the results, assuming gaussian distribution (except for the arousal multiclass experiment on DEAP, where the standard deviations overlap between the HF to Brain and Brain to LF).

Across both the MAHNOB-HCI and DEAP datasets, the HF-to-Brain feature consistently outperforms other BHI features in emotion classification tasks. In the MAHNOB-HCI dataset, it achieves the highest accuracy in both arousal (96.92% ± 0.30%) and valence (96.79% ± 0.28%) classification, surpassing Brain-to-HF, LF-to-Brain, and Brain-to-LF features. Similarly, in the DEAP dataset, HF-to-Brain demonstrates strong discriminative power, though in binary valence classification, LF-to-Brain performs slightly better (96.56% ± 0.31%) compared to HF-to-Brain (93.37% ± 14.89%).

In contrast, EEG-PSD-based analysis yields lower accuracies, ranging from 73.23% to 89.31% in MAHNOB-HCI and from 37.74% to 66.44% in DEAP, further highlighting the superior performance of BHI features in emotion classification.

These findings underscore the superior performance of BHI-based analysis over EEG-based approaches in emotion classification. In particular, the HF-to-Brain feature consistently outperforms other BHI features and EEG-PSD analysis, effectively capturing more relevant information for emotion recognition. Given its strong predictive capability, this model was selected to investigate the physiological relationship between BHI and emotional states, as it provides the most accurate approximation of the underlying phenomenon.

Table 2 Accuracy results of the proposed CNN model in various emotion classification tasks (binary: arousal and valence, multiclass: categorical emotions, arousal, valence) and a 10-cross fold validation schema

		Arousal [1,9]		Valence [1,9]		Emotion [0, 8]		Arousal [0, 1]		Valence [0, 1]	
MAHNOB-HCI											
		acc	std	acc	std	acc	std	acc	std	acc	std
EEG	PSD	73.23%	0.91%	74.69%	0.88%	74.72%	0.87%	88.12%	0.80%	89.31%	0.70%
BHI	**HF→Brain**	**96.92%**	0.30%	**96.79%**	0.28%	**96.98%**	0.48%	**98.27%**	0.32%	**98.31%**	0.26%
	LF→Brain	93.92%	0.61%	93.67%	0.57%	93.77%	0.85%	96.59%	0.44%	96.92%	0.54%
	Brain→HF	96.72%	0.63%	97.07%	0.45%	96.56%	0.42%	98.07%	0.20%	98.28%	0.36%
	Brain→LF	95.15%	0.64%	95.35%	0.77%	95.14%	0.60%	97.43%	0.35%	97.36%	0.40%
DEAP											
EEG	PSD	39.46%	0.70%	37.24%	0.49%			66.22%	0.83%	66.44%	0.69%
BHI	**HF→Brain**	**96.80%**	**0.26%**	**96.36%**	**0.17%**			**98.27%**	**0.19%**	93.37%	14.89%
	LF→Brain	92.51%	0.50%	92.02%	0.49%			97.22%	0.32%	**96.56%**	**0.31%**
	Brain→HF	87.22%	24.11%	86.83%	23.72%			93.43%	13.37%	92.49%	15.03%
	Brain→LF	96.72%	0.22%	92.52%	0.42%			96.21%	0.23%	92.08%	0.29%

Best mean values per classification task are marked in bold

5.2 Model-driven evaluation of the explanations

Figure 3 shows the quality test results derived from the insertion, deletion, and sensitivity tests. In Fig. 3, GradCAM consistently underperforms the other XAI algorithms across all tests: in the insertion test, the GradCAM curve is consistently below all other curves, indicating a slower improvement in model performance when important features are reintroduced; in the deletion tests, the GradCAM curve is above all the others, reflecting a slower decline in performance when key features are removed or perturbed.

The area under the curve (AUC) results indicate GradCAM's poor performance in quality tests, with higher AUC in deletion and sensitivity tests, and lower AUC in insertion tests, underscoring its high infidelity.

Figure 5 displays the selection scores obtained with QT across multiple tasks on the DEAP (top row) and MAHNOB (bottom row) datasets, focusing on arousal, valence, and, where applicable, emotions classification. Each heatmap cell shows a percentage score for a particular class label and explanation method, with darker cells indicating stronger performance (i.e., lower percentage values). Overall, IG demonstrates consistently high scores across most classes, often outperforming or closely matching DL and EG. However, the optimal method may depend on the target class. For instance, in the multi-class arousal classification task on MAHNOB, IG leads in most classes, yet DL surpasses it for the 0-class (21.4%) and GC exceeds it for the 2-class (38.1%). These results underscore that, while IG frequently excels, no single explanation method is universally optimal for every class or task.

Figure 4 compares topographical activation maps from IG (left) and GC (right) for binary arousal classification on the MAHNOB dataset across four frequency bands. These maps, shown as topographic plots, illustrate EEG electrode data in 2D. Darker brain regions indicate higher importance of BHI HF-to-Brain features in the CNN model's decision, while lighter areas show lower or neutral importance.

Although IG achieves a better QT (i.e. lower), 6%, compared to GC, 22.2%, the visual differences between their corresponding activation maps are not always evident.

In Beta and Gamma bands, both methods identify frontal and central regions as key for distinguishing arousal levels. In Theta and Alpha bands, IG focuses on frontal and frontocentral areas, while GC includes parietal or occipital sites. The QT helps us evaluate IG's numerical performance compared to GC's, showing that IG provides more globally representative and model-aligned explanations, as highlighted by QT scores (6% vs. 22.2%).

Figure 2 illustrates the selected global class-wise explanations for the categorical emotion classification task.

In the 'sadness' classification, we can see a symmetric lateralization of the importance extracted in the alpha and theta bands. A similar phenomenon can be observed in the categorical emotions 'amusement' and 'neutral'. With lower intensity, this result is visible in all the other categorical emotions, except for 'anger'.

Overall, the theta and alpha frequency bands share higher importance compared to the beta and gamma bands.

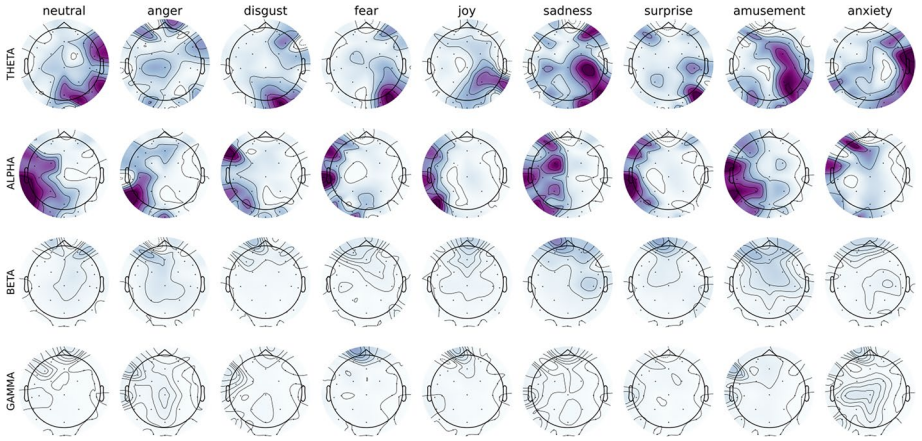

Fig. 2 Optimal global explanations obtained via the classification of nine different categorical emotions using BHI (HF-to-brain) features as input

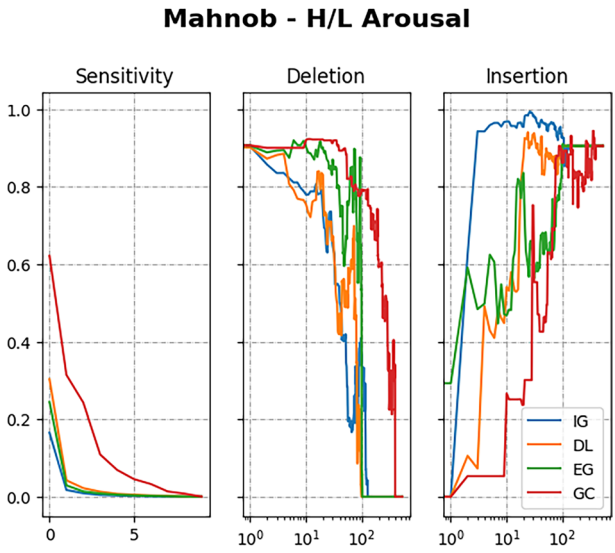

Fig. 3 Quality test results comparing the performance of different explanation methods, integrated gradients (IG), expected gradients (EG), DeepLIFT (DL), and GradCAM (GC), across insertion, deletion, and sensitivity tests on the Mahnob dataset in binary high versus low arousal classification

6 Discussion

The results obtained in our study include: (i) the evaluation of the BHI features performances in comparison with EEG-only features in recognizing emotional states; (ii) the extraction of the global optimal explanations from a set of local explainers from the best-performing architecture obtained in (i).

While this study does not aim to develop the highest-performing emotion recognition model, our primary focus is on generating optimal explanations that provide valuable physiological insights into how BHI features determine emotional states.

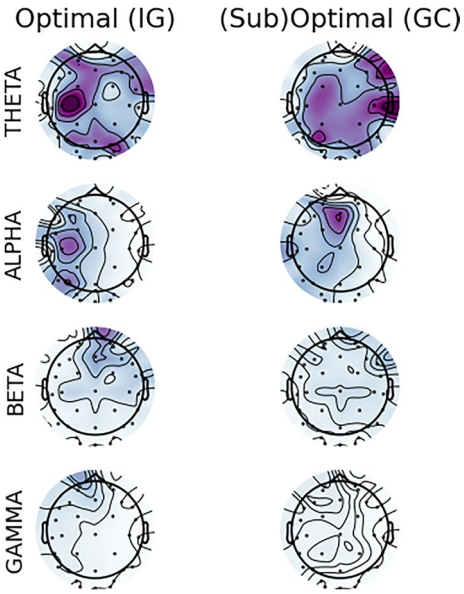

Fig. 4 Global explanations extracted by two methods, integrated gradient (IG) and GradCAM (GC) on the Mahnob dataset in binary high versus low arousal classification. Following Fig. 3 IG explanation has been selected as optimal by the QT assessment while GC is the least optimal

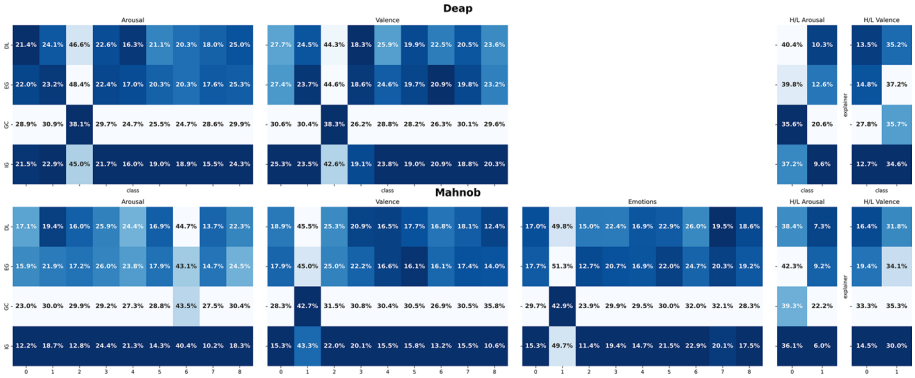

Fig. 5 QT results (i.e. the lower the better) on each output class obtained comparing different explanation methods, integrated gradients (IG), expected gradients (EG), DeepLIFT (DL), and GradCAM (GC), across two different datasets and classification setup

6.1 Classification performance

The proposed methodology was evaluated on the publicly available MAHNOB-HCI dataset, comparing EEG PDS and BHI feature sets rearranged in an image-like fashion. Both feature sets demonstrated satisfactory performance metrics, with the BHI outperforming the EEG-ones. Indeed, the results in Table 2 show a significant accuracy improvement: 20% for multiclass and 10% for binary classification on the MAHNOB dataset, and 50% for multiclass and 30% for binary classification on the DEAP dataset.

Thus, our study highlights the limitations of relying solely on EEG features (PSD) for emotion classification. Although the PSD-based analysis achieved reasonable accuracies,

Table 3 Comparison between the proposed approach and the state of the art on two benchmark datasets for emotion recognition, MAHNOB and DEAP, and different validation strategies, k-fold, LOTO, and LOSO. $LOSO^{ft}$ stands for the models tested with LOSO validation and fine-tuned with 20% of the testing data

Dataset	Approach	Validation	H/L arousal	H/L valence	Cat. emotions
DEAP	Yin et al. (2021)	5-fold	85.27%	84.81%	
	Wang et al. (2021)	5-fold	78.22%	83.85%	
	Gu et al. (2023)	5-fold	94.42%	94.87%	
	Dhara et al. (2024)	5-fold	91.72%	90.84	
	Zheng et al. (2017)	5-fold			69.67% (4)
	Issa et al. (2020)	10-fold			93.1% (4)
	Our Appr	10-fold	98.27%	96.56%	
	Li et al. (2020)	LOTO	74.51%	71.63%	
	Our Appr	LOTO	85.88%	86.6%	
	Zhong et al. (2024)	LOSO	**70.37%**	**71.11%**	
	Zhang et al. (2021)	LOSO	67.12%	68.18%	
	Li et al. (2023)	LOSO	58.04%	60.7%	
	Li et al. (2023)	LOSO	58.35%	57.73%	
	Ding et al. (2022)	LOSO	63.75%	62.27%	
	Our Appr	LOSO	61.51%	59.63%	
	Our Appr	$LOSO^{ft}$	72.04%	71.28%	
MAHNOB-HCI	Issa et al. (2020)	10-fold			94.4% (4)
	Our Appr	10-fold	98.27%	98.31%	96.98% (9)
	Li et al. (2020)	LOTO	82.18%	80.09%	
	Our Appr	LOTO	85.96%	87.34%	
	Zhong et al. (2024)	LOSO	**68.88%**	**71.15%**	
	Our Appr	LOSO	62.04%	67.31%	
	Our Appr	$LOSO^{ft}$	80.61%	83.61%	

Best mean values per classification task andvare marked in bold

similar to the one reported in the literature (Huang et al., 2019), the BHI features consistently outperformed them across all classification tasks.

Table 3 showcases a comprehensive comparison between our proposed approach and state-of-the-art methods in EEG-based emotion recognition, under various validation strategies: k-fold, leave-one-trial-out (LOTO), and leave-one-subject-out (LOSO).

The results on the k-fold validation strategy highlight the superior performances of the proposed methodology in binary arousal and valence classification tasks and categorical emotion recognition tasks compared to EEG-based methodologies (Yin et al., 2021; Wang et al., 2021; Gu et al., 2023; Zheng et al., 2017; Issa et al., 2020). Also, other state-of-the-art approaches in categorical emotion recognition identify the categorical emotions partitioning the arousal-valence space [i.e., circumplex space of affect model (Posner et al., 2005; Lang, 1995)] considering then only 4 emotions. The proposed methodology has instead been tested on the 9 self-assessed categorical emotions, expressed by subjects, in the MAHNOB-HCI dataset.

To test the generalization capabilities of the proposed approach both LOTO and LOSO validation strategies have been implemented resulting in superior recognition capabilities in both arousal and valence classification tasks with respect to the EEG-based methodologies (Li et al., 2020; Zhang et al., 2021; Li et al., 2023, 2023). While the approaches proposed in

Zhang et al. (2021), Zhong et al. (2024) and Ding et al. (2022) provided better performances than our approach with the LOSO validation schema ad EEG features, a simple fine-tuning on a small subset of test data allowed our approch to outperfom all of those..

6.2 Model-driven validation of class-wise explanations

Our XAI framework systematically evaluates and selects the optimal global explanations, thereby ensuring reliable and unbiased interpretability. Our proposed quality test evaluates both the *fidelity* and *sensibility* of each global class-wise explanation and uses them to compare different explanations rankings.

This allows to quantitatively measure the quality of the explanations. In contrast to other studies (Gagliardi et al., 2023b) that use different methods for extracting global importance, allowing only for a qualitative knowledge-driven evaluation of the explanations.

By ranking class-wise explanation candidates and guiding the human experts toward the best ones, the process becomes more efficient and trustworthy

From the considered XAI approaches, i.e., Integrated gradients, expected gradients, DeepLIFT, and GradCAM, integrated gradients outperformed on average the others in terms of fidelity and sensitivity.

Also, the XAI approach employed in this study enables a visual assessment of the importance of different scalp locations and frequency bands in the classification process, facilitating meaningful comparisons with previous studies (Sarma et al., 2020). Consistently, we observed that the theta and alpha bands of the EEG signals played a prominent role across all experimental tasks, which aligns with existing literature on the involvement of these frequency bands in emotion perception and processing (Sarma et al., 2020). However, it is important to emphasize that these results represent the initial attempt to map the importance of scalp locations in multiclass emotion perception, particularly by incorporating BHI features. Further research is warranted to validate and extend these findings, considering a broader range of emotional stimuli and exploring potential variations across individuals (Sarma et al., 2020).

By colour-coding the areas of the brain that are most influential for emotion recognition according to the CNN, we present intuitive and easily interpretable topographic representations of the AI-based explanations (see Fig. 2) (Li et al., 2020). To the best of our knowledge, this is the first study to provide different BHI (and EEG) information on such a detailed level of the emotional spectrum, allowing us to identify selectively informative scalp regions in their interplay with the autonomic system regarding specific emotions. This approach complements similar research in the field that arranges features extracted from physiological signals into image format (Gagliardi et al., 2023b; Jung & Sejnowski, 2019; Li et al., 2020), but stands out as the first to apply this strategy to provide explanations by analyzing inherently multi-modal and highly informative physiological features, particularly those related to the brain-heart relationship.

By considering the interplay between the brain and heart, we harness the full potential of these multi-modal features, resulting in more comprehensive and insightful explanations (Gagliardi et al., 2023b; Jung & Sejnowski, 2019; Li et al., 2020).

6.3 Neurophysiological insights

The performances obtained in a large variety of emotion classification tasks with different validation strategies emphasize the potential of incorporating cardiovascular information, specifically heart rate variability, to enhance the understanding of emotional states. The combination of EEG and HRV through the functional brain-heart interplay approach proved to be a more effective strategy for discerning emotions, as evidenced by the higher accuracies achieved. Several reasons contribute to this observation. Firstly, BHI features captures the interplay between cerebral and cardiovascular dynamics, incorporating information from both systems' activities. In contrast, EEG features solely reflect cerebral dynamics. By incorporating BHI features, we gain a more comprehensive understanding of the physiological processes underlying emotional processing. Secondly, human emotional processing and perception are known physiological phenomena arising from the interaction between the central and autonomic nervous system (Candia-Rivera et al., 2022; Thayer et al., 2012; Lang, 1995). Therefore, it is reasonable to deduce that features associated with this interaction are more informative to the task of emotional processing compared to features derived from a single system alone.

The incorporation of BHI features not only outperformed other approaches incorporating multimodal features (Zhang et al., 2022; Jung & Sejnowski, 2019), but also provided valuable neurophysiological insights. The attribution maps depicted the brain regions where the BHI HF-to-Brain features exerted the most influence, highlighting the importance of specific frequency bands and scalp locations. Notably, the alpha and theta frequency bands emerged as particularly influential in capturing BHI-related emotional states, while the beta and gamma ranges exhibited relatively lower importance. A reverse analysis can be conducted by comparing the attribution maps obtained from the CNN model used for classifying the EEG-PSD features. In this case, the model exhibits a distinct attentional focus on beta and gamma frequency ranges. Notably, within the gamma range, a discernible pattern of attention-lateralization emerges, as the network directs its attention toward the left hemisphere for each emotional group.

Delving deeper into the phenomenon of BHI, our experimental findings suggest that the HF-to-Brain feature, in particular, is associated with the best-performing schema. This observation aligns with previous studies that have explored how the perception of emotions influences cardiovascular dynamics and the directional functional BHI (Candia-Rivera et al., 2022). However, our results do not indicate clear distinctions regarding the directionality of the interaction (i.e., brain-to-heart versus heart-to-brain) or the specific heartbeat frequency bands considered (i.e., LF and HF). It appears that BHI-related features, regardless of their directionality or frequency bands, contribute significantly to emotion classification, achieving an accuracy higher than 93 These findings highlight the complex and dynamic nature of the brain-heart relationship in the context of emotion recognition. Indeed, the data used in this study were collected during dynamical emotional elicitation through image presentation, and the time-varying evolution of BHI changes was collapsed through temporal averaging. This temporal averaging may have obscured finer-grained dynamics and potential variations across specific temporal segments or emotional stimuli.

7 Conclusions

In this study, we introduced a novel approach to identify the best class-wise explanation. The method has been tested in emotion recognition tasks that integrates functional brain-heart interaction (BHI) estimates derived from EEG and HRV.

The integration of BHI features significantly enhanced emotion recognition accuracy over models using only EEG features and enabled the fine-grained distinction up to nine emotional classes. Beyond performance improvements, our methodology offers a more comprehensive understanding of emotional states by providing valuable neurophysiological insights via XAI-based global optimal attribution maps.

Due to the complexity and novelty of BHI features, that underscores challenges in fully validating these explanations against existing neurophysiological knowledge, the key contribution of our work is the development of a dynamic method for selecting the optimal global explainer.

This approach enriches the qualitative evaluation via visualization of explanation, with a quantitative assessment of the fidelity and sensitivity. This approach enables the selection of the most reliable explanation for each emotional class. While our method autonomously identifies the best optimal explanation, we acknowledge the importance of human validation—especially in critical applications such as medical scenarios.

Despite its promising results, the main limitation of the proposed framework side in its application context. Primarily, the method identifies optimal model-driven explanations in scenarios where human validation is not feasible due to the complexity or novelty of the input features, such as those derived from brain-heart interactions (BHI). In such contexts, these explanation techniques can serve as knowledge discovery tools, offering physiological insights into complex mechanisms like the brain-heart interplay. However, to establish a robust physiologically grounded knowledge discovery pipeline, the reproducibility and reliability of the extracted insights must be validated across multiple datasets and cross-dataset studies. In the domain of emotion recognition, this remains a challenge due to the limited availability of data specifically designed for BHI computation.

Nonetheless, such validation may be more feasible in alternative case studies, such as sleep stage classification, where EEG-ECG-related data are more abundant. Future work will explore the integration and assessment of the proposed framework in such domains.

Appendix

Appendix A: Arrangement of spatial features into an image

At this point, the BHI and EEG extracted features have been rearranged into images following the approach proposed in Gagliardi et al. (2023a, 2023b). The resulting input image is shown in Fig. 1.

For each heart frequency (LF, HF) and direction of BHI feature estimation (brain-to-heart, heart-to-brain), we extracted 32×4 features (EEG channels $\times$ EEG frequency bands) from each time window. These 32×4 input features, representing each frequency band (α, β, θ, and γ) and EEG channel, are arranged in a sparse matrix that mimics the spatial loca-

tion of the electrodes in the 10–20 EEG system. The four bands are concatenated within the same 2D input image (Fig. 1). This channel ordering scheme enhances the CNN model's ability to utilize the spatial relationships between sensors, thereby improving its performance (Gagliardi et al., 2023a).

Appendix B: CNN model

The input matrix is processed by a convolutional neural network (CNN) comprising two convolutional layers with depths of 32 and 64, respectively.

These layers employ the ReLU activation function. To prevent overlap of electrode positions across different frequency blocks, the convolutional filter size is designated as 3, consistent with the padding applied to the frequency blocks. Subsequently, the data is processed through two fully connected dense layers comprising 512 neurons and either 9 or 2 neurons, contingent upon the number of classes. Before the dense layers, a flattening layer is utilised to transform the two-dimensional representation into a one-dimensional vector.

The output layer, equipped with a softmax activation function, yields class membership scores. To enhance generalisation and mitigate overfitting, a dropout layer with $p = 0.2$ is incorporated after the flattening layer. The model is trained utilising the Adam Optimiser paired with a learning rate of 10^{-4}.

Appendix C: Model evaluation

The proposed CNN-based emotion recognition is tested for different features (i.e., EEG-PSD power, and 4 BHI feature sets separately) both in a *multiclass* and a *binary* classification fashion.

In the *multiclass* emotion recognition task, the model was trained to recognize different types of emotions using a 10-fold cross-validation methodology. The evaluation tasks included categorical emotion classification, namely neutral, anger, disgust, fear, joy, sadness, surprise, amusement, and anxiety, classification of 9 levels of arousal, and classification of 9 levels of valence.

In the MAHNOB-HCI dataset, the identified 9 levels of arousal and valence and the 9 categorical emotions correspond to the ones directly annotated by the subjects in the dataset. In the DEAP dataset instead, the arousal and valence were identified by subjects on a continuous 9-point scale. The levels utilized in this study were then determined by approximating their respective continuous values to the nearest integer.

For the binary emotion recognition task, we divided the discrete values of arousal and valence into two classes: high and low. The threshold for distinguishing low and high arousal/valence was set as 4.5, whereas samples with arousal/valence values less than 4.5 were labeled as low, otherwise, they were labeled as high. Similarly to the multiclass emotion recognition model, the binary classification model was evaluated using a 10-fold cross-validation strategy. After splitting the input data into training and test sets, the data has been standardized to mitigate any bias related to the overall EEG spectrum, specifically addressing the issue of increased power in low-frequency bands.

Both the leave-one-trial-out (LOTO) and leave-one-subject-out (LOSO) validation methodologies were utilized. Within the LOTO framework, the model is developed by utilizing all trials within the dataset—with the exception of one—and subsequently evaluated on the

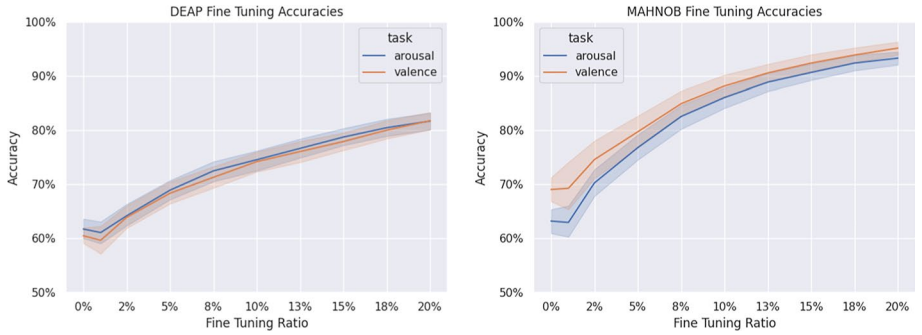

Fig. 6 Fine-tuning accuracy on the DEAP and MAHNOB datasets with varying percentages of training data

trial that was omitted. Each trial in the dataset is associated with a unique video-subject pairing, representing BHI or EEG samples gathered from the same session during which a subject viewed a video. The MAHNOB-HCI dataset comprises data from 27 subjects and 20 different videos, yielding a total of 540 trials. Correspondingly, the DEAP dataset resulted in 1280 trials. In the LOSO approach, the model is trained on data from all subjects except one, and it is tested on the excluded subject.

Appendix D: Real-world usability and generalization test

To assess the generalization capability of the proposed approach, we evaluated the best-performing CNN-based architecture, namely the CNN for binary classification trained on BHI HF-to-Brain features, using both leave-one-trial-out (LOTO) and leave-one-subject-out (LOSO) validation schemes.

The LOTO evaluation yielded accuracies of $85.96\% \pm 27.64\%$ for arousal classification and $87.34\% \pm 25.74\%$ for valence classification on the MAHNOB-HCI dataset. For the DEAP dataset, the approach achieved $85.88\% \pm 27.19\%$ for arousal and $86.6\% \pm 26.8\%$ for valence. In contrast, the LOSO validation results, summarized in Fig. 6, showed a performance drop: the model reached $61.51\% \pm 8.47$ for arousal and $59.63\% \pm 8.04$ for valence on the DEAP dataset, and $62.04\% \pm 8.18$ and $67.31\% \pm 10.16$ for arousal and valence, respectively, on the MAHNOB dataset.

To mitigate this decline and improve generalization, we introduced a fine-tuning strategy. Specifically, a portion of the testing subject's data was randomly sampled and incorporated as additional training instances. The CNN was then retrained for five epochs with the last layer unfrozen. As illustrated in Fig. 6, model performance consistently improved with larger proportions of fine-tuning data. Notably, using only 20% of the testing subject's data, the model achieved $72.04\% \pm 10.61$ (arousal) and $71.28\% \pm 10.9$ (valence) on the DEAP dataset, and $80.61\% \pm 13.35$ (arousal) and $83.61\% \pm 13.04$ (valence) on the MAHNOB dataset.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10994-025-06921-y.

Author contributions Guido Gagliardi: Conceptualization, formal analysis, investigation, software, visualization, writing—original draft. Antonio Luca Alfeo: Conceptualization, formal analysis, investigation,

software, writing—review and editing. Vincenzo Catrambone: Conceptualization, formal analysis, investigation, software, writing—review and editing. Mario G.C.A. Cimino: Formal analysis, visualization, funding acquisition, writing—review and editing Gaetano Valenza: Formal analysis, investigation, funding acquisition, writing—review and editing. Maarten De Vos: Formal analysis, investigation, funding acquisition, writing—review and editing.

Funding Open access funding provided by Università degli Studi di Firenze within the CRUI-CARE Agreement. Financial support was provided by the Research Foundation Flanders (FWO) via SBO mandate 1SH4Z24N. Work partially supported by Horizon 2020 Program under GA 101017727 of the project "EXPERIENCE"; FWO SBO Project: "Supporting the development of self-regulation in infants: a promising strategy in preventive mental health care", S003524N; FWO Research Project: 'Artificial Intelligence (AI) for data-driven personalized medicine', G0C9623N; FWO Research Project: 'Deep, personalized epileptic seizure detection', G0D8321N; Italian Ministry of University and Research (MUR) in the frameworks: "FAIR" PE00000013 Spoke1 "Human-centered AI"; National Center for Sustainable Mobility MOST/ Spoke10; "Reasoning" project, PRIN 2020 LS Programme, Project number 2493 04-11-2021; "THE" cod. ECS00000017 - CUP I53C22000780001; FoReLab project (Departments of Excellence); European Commission under the NextGeneration EU program, PNRR - M4 C2, Investment 1.5 "Creating and strengthening of "innovation ecosystems", building "territorial R&D leaders", project "THE - Tuscany Health Ecosystem"; Spoke 6 "Precision Medicine and Personalized Healthcare". 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 the 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.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

Ahmed, N., Aghbari, Z. A., & Girija, S. (2023). A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications, 17*, Article 200171.

Alfeo, A. L., Cimino, M. G. C. A. & Gagliardi, G. (2022). Concept-wise granular computing for explainable artificial intelligence. *Granular Computing*

Barra, S., Casanova, A., Fraschini, M., & Nappi, M. (2017). Fusion of physiological measures for multimodal biometric systems. *Multimedia Tools and Applications, 76*, 4835–4847.

Benarroch, E. E. (2008). The central autonomic network: Functional organization, dysfunction, and perspective. *Mayo Clinic Proceedings, 68*(10), 988–1001.

Calvo, R. A., & D'Mello, S. K. (2014). Affective computing and the impact of gender and age. *Pervasive and Mobile Computing, 17*, 55–63.

Candia-Rivera, D. et al. (2022) Cardiac sympathetic-vagal activity initiates a functional brain-body response to emotional arousal. *Proceeding of the National Academy of Science*

Candia-Rivera, D., Catrambone, V., & Valenza, G. (2021). The role of electroencephalography electrical reference in the assessment of functional brain-heart interplay: From methodology to user guidelines. *Journal of Neuroscience Methods, 360*, Article 109269.

Catrambone, V. & Valenza, G. (2021) *Functional Brain-Heart Interplay: From Physiology to Advanced Methodology of Signal Processing and Modeling*. Springer Nature.
Catrambone, V. & Valenza, G. (2023). Nervous–system–wise functional estimation of directed brain–heart interplay through microstate occurrences. *IEEE Transactions on Biomedical Engineering*
Catrambone, V. (2019). https://it.mathworks.com/matlabcentral/fileexchange/72704-brain-heart-interaction-indexes,
Catrambone, V., et al. (2019). Time-resolved directional brain-heart interplay measurement through synthetic data generation models. *Annals of Biomedical Engineering, 47*(6), 1479–1489.
Catrambone, V., et al. (2021). Intensification of functional neural control on heartbeat dynamics in subclinical depression. *Translational Psychiatry, 11*(1), 1–10.
Catrambone, V., Talebi, A., Barbieri, R., & Valenza, G. (2021). Time-resolved brain-to-heart probabilistic information transfer estimation using inhomogeneous point-process models. *IEEE Transactions on Biomedical Engineering, 68*(11), 3366–3374.
Choi, D. Y., Kim, D.-H., & Song, B. C. (2020). Multimodal attention network for continuous-time emotion recognition using video and EEG signals. *IEEE Access, 8*, 203814–203826.
Citi, L., Brown, E. N., & Barbieri, R. (2012). A real-time automated point-process method for the detection and correction of erroneous and ectopic heartbeats. *IEEE Transactions on Biomedical Engineering, 59*(10), 2828–2837.
Dhara, T., Singh, P. K., & Mahmud, M. (2024). A fuzzy ensemble-based deep learning model for EEG-based emotion recognition. *Cognitive Computation, 16*(3), 1364–1378.
Ding, Y., Neethu Robinson, S., Zhang, Q. Z., & Guan, C. (2022). Tsception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. *IEEE Transactions on Affective Computing, 14*(3), 2238–2250.
Di, W., Zhang, J., & Zhao, Q. (2020). Multimodal fused emotion recognition about expression-EEG interaction and collaboration using deep learning. *IEEE Access, 8*, 133180–133189.
Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., & Lee, S.-I. (2021). Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence, 3*(7), 620–631.
Faes, L., Kugiumtzis, D., Nollo, G., Jurysta, F., & Marinazzo, D. (2015). Estimating the decomposition of predictive information in multivariate systems. *Physical Review E, 91*(3), Article 032904.
Gagliardi, G., Alfeo, A. L., Catrambone, V., Candia-Rivera, D., Cimino, M. G. C. A. & Valenza, G. (2023) Improving emotion recognition systems by exploiting the spatial information of EEG sensors. *IEEE Access* (pp. 1–1).
Gagliardi, G., Alfeo, A. L., Catrambone, V., Candia-Rivera, D., Cimino, M. G. C. A., Valenza, G., & De Vos, M. (2023). Fine-grained emotion recognition using brain-heart interplay measurements and explainable convolutional neural networks. In *Proceedings of the 11th international IEEE EMBS conference on neural engineering* (pp. 1–1).
Gagliardi, G., Alfeo, A. L., Catrambone, V., Cimino, M. G. C. A., De Vos, M. & Valenza, G. (2023). Using contrastive learning to inject domain-knowledge into neural networks for recognizing emotions. In *2023 IEEE symposium series on computational intelligence (SSCI)* (pp. 1587–1592). IEEE
Gu, Y., Zhong, X., Qu, C., Liu, C., & Chen, B. (2023). A domain generative graph network for EEG-based emotion recognition. *IEEE Journal of Biomedical and Health Informatics*,
Guo, H., Jiang, N. & Shao, D. (2020). Research on multi-modal emotion recognition based on speech, eeg and ecg signals. In *Robotics and rehabilitation intelligence: First international conference, ICRRI 2020, Fushun, China, September 9–11, 2020, Proceedings, Part I 1* (pp. 272–288). Springer.
Hosseini, I., Hossain, M. Z., Zhang, Y., & Rahman, S. (2024). Deep learning model for simultaneous recognition of quantitative and qualitative emotion using visual and bio-sensing data. *Computer Vision and Image Understanding, 248*, Article 104121.
Huang, Y., Yang, J., Liu, S., & Pan, J. (2019). Combining facial expressions and electroencephalography to enhance emotion recognition. *Future Internet, 11*(5), 105.
Issa, S., Peng, Q., & You, X. (2020). Emotion classification using EEG brain signals and the broad learning system. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 51*(12), 7382–7391.
Jung, T.-P., Sejnowski, T. J., et al. (2019). Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing. *IEEE Transactions on Affective Computing, 13*(1), 96–107.
Khosrowabadi, R., Quek, H. C., Ang, K. K., & Tung, S. W. (2010). qeeg-based emotion recognition. *Neural information processing* (pp. 594–603).
Kim, S.-H. (2025). Mifu-er: Modality quality index-based incremental fusion for emotion recognition. *IEEE Access*
Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2011). DEAP: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing, 3*(1), 18–31.

Kumar, A., & Kumar, A. (2025). Human emotion recognition using machine learning techniques based on the physiological signal. *Biomedical Signal Processing and Control, 100*, Article 107039.

Lan, Y.-T., Liu, W., & Lu, B.-L. (2020) Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism. In *2020 international joint conference on neural networks (IJCNN)* (pp. 1–6). IEEE.

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist, 50*(5), 372.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience, 23*(1), 155–184.

Li, W., Fang, C., Zhu, Z., Chen, C., & Song, A. (2023). Fractal spiking neural network scheme for EEG-based emotion recognition. *IEEE Journal of Translational Engineering in Health and Medicine*,

Lian, Y., Zhu, M., Sun, Z., Liu, J., & Hou, Y. (2025). Emotion recognition based on EEG signals and face images. *Biomedical Signal Processing and Control, 103*, Article 107462.

Li, W., Wang, M., Zhu, J., & Song, A. (2023). EEG-based emotion recognition using trainable adjacency relation driven graph convolutional network. *IEEE Transactions on Cognitive and Developmental Systems, 15*(4), 1656–1672.

Li, Y., Yang, H., Li, J., Chen, D., & Min, D. (2020). EEG-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by grad-cam. *Neurocomputing, 415*, 225–233.

Lu, L., Yuan, L., & Chen, L. (2025). Deep learning based emotion recognition for analyzing students' psychological states during competitions. *Entertainment Computing*, 101005.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*.

Orini, M., Bailón, R., Mainardi, L. T., Laguna, P., & Flandrin, P. (2011). Characterization of dynamic interactions between cardiovascular signals by time-frequency coherence. *IEEE Transactions on Biomedical Engineering, 59*(3), 663–673.

Petsiuk, V., Das, A., & Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421,

Pillalamarri, R., & Shanmugam, U. (2025). A review on EEG-based multimodal learning for emotion recognition. *Artificial Intelligence Review, 58*(5), 131.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology, 17*(3), 715–734.

Saganowski, S., Perz, B., Polak, A., & Kazienko, P. (2022). Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review. *IEEE Transactions on Affective Computing, 14*(3), 1876–1897.

Said, Y., Saidani, T., Atri, M., Alsheikhy, A. A., & Shawly, T. (2026). Computational intelligence for emotion recognition in autism spectrum disorder: A systematic review of signal-based modeling, simulation, and clinical potential. *Biomedical Signal Processing and Control, 111*, Article 108367.

Sarma, G. P., Reinertsen, E., Aguirre, A., Anderson, C., Batra, P., Choi, S.-H., Achille, P. D., Diamant, N., Ellinor, P., Emdin, C., et al. (2020). Physiology as a lingua franca for clinical machine learning. *Patterns, 1*(2), Article 100017.

Sedehi, J. F., Dabanloo, N. J., Maghooli, K., & Sheikhani, A. (2025). Develop an emotion recognition system using jointly connectivity between electroencephalogram and electrocardiogram signals. *Heliyon, 11*(2).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision, 128*, 336–359.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *International conference on machine learning* (pp. 3145–3153). PMlR

Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2011). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing, 3*(1), 42–55.

Suhaimi, N. S., Mountstephens, J., Teo, J. et al. (2020). Eeg-based emotion recognition: A state-of-the-art review of current trends and opportunities. *Computational Intelligence and Neuroscience*.

Sundararajan, M., Taly, A. & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319–3328). PMLR.

Thayer, J. F., Åhs, F., Fredrikson, M., Sollers, J. J., & Wager, T. D. (2012). A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews, 36*(2), 747–756.

Torres-Valencia, C. A., Garcia-Arias, H. F., Lopez, M. A. A. & Orozco-Gutiérrez, A. A. (2014). Comparative analysis of physiological signals and electroencephalogram (EEG) for multimodal emotion recognition using generative models. In *2014 XIX symposium on image, signal processing and artificial vision* (pp. 1–5). IEEE.

Valenza, G., Citi, L., Saul, J. P., & Barbieri, R. (2018). Measures of sympathetic and parasympathetic autonomic outflow from heartbeat dynamics. *Journal of Applied Physiology, 125*(1), 19–39.

Valenza, G., Passamonti, L., Duggento, A., Toschi, N., & Barbieri, R. (2020). Uncovering complex central autonomic networks at rest: A functional magnetic resonance imaging study on complex cardiovascular oscillations. *Journal of the Royal Society Interface, 17*(164), 20190878.

Valenza, G., Sclocco, R., Duggento, A., Passamonti, L., Napadow, V., Barbieri, R., & Toschi, N. (2019). The central autonomic network at rest: Uncovering functional MRI correlates of time-varying autonomic outflow. *Neuroimage, 197*, 383–390.

Wang, Z., Tianhao, G., Zhu, Y., Li, D., Yang, H., & Wenli, D. (2021). FLDNet: Frame-level distilling neural network for EEG emotion recognition. *IEEE Journal of Biomedical and Health Informatics, 25*(7), 2533–2544.

Wang, Z., & Wang, Y. (2025). Emotion recognition based on multimodal physiological electrical signals. *Frontiers in Neuroscience, 19*, 1512799.

Wei, Y., Lil, Y., Xu, M., Hua, Y., Gong, Y., Osawa, K. & Tanaka, E. (2023) A real-time and two-dimensional emotion recognition system based on EEG and HRV using machine learning. In *2023 IEEE/SICE international symposium on system integration (SII)* (pp. 1–6). IEEE.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., & Ravikumar, P. K. (2019). On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32

Yin, Y., Zheng, X., Bin, H., Zhang, Y., & Cui, X. (2021). EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM. *Applied Soft Computing, 100*, Article 106954.

Zhang, Y., Cheng, C., Wang, S., & Xia, T. (2022). Emotion recognition using heterogeneous convolutional neural networks combined with multimodal factorized bilinear pooling. *Biomedical Signal Processing and Control, 77*, Article 103877.

Zhang, G., Minjing, Yu., Liu, Y.-J., Zhao, G., & Zhang, D. (2021). *and Wenming Zheng*. Sparsedgcnn: Recognizing emotion from multichannel EEG signals. *IEEE Transactions on Affective Computing*.

Zheng, W.-L., Zhu, J.-Y., & Bao-Liang, L. (2017). Identifying stable patterns over time for emotion recognition from EEG. *IEEE Transactions on Affective Computing, 10*(3), 417–429.

Zhong, X., Wu, F., Yin, Z., & Liu, G. (2024). An attention-enhanced retentive broad learning system for subject-generic emotion recognition from eeg signals. In *ICASSP 2024-2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2310–2314). IEEE.

Authors and Affiliations

Guido Gagliardi[1,2,3] · Antonio Luca Alfeo[4] · Vincenzo Catrambone[1,5] · Mario G. C. A. Cimino[1,5] · Maarten De Vos[2,6] · Gaetano Valenza[1,5]

✉ Guido Gagliardi
guido.gagliardi@phd.unipi.it

[1] Department of Information Engineering, University of Pisa, Pisa, Italy

[2] Department of Electrical Engineering, KU Leuven, Leuven, Belgium