

Information Sheet

Title: Model-driven Validation of Visual Explanations for Multimodal Emotion Recognition

This submission presents a novel framework for validating visual explanations in multimodal emotion recognition models. The proposed approach autonomously selects the most reliable explanations for a black-box model trained on brain-heart interaction (BHI) features from EEG and HRV data. By leveraging a Convolutional Neural Network (CNN) and a model-agnostic methodology, we systematically evaluate and score multiple explainability techniques, identifying Integrated Gradients as the most effective for capturing global insights.

The study addresses a key challenge in AI-based emotion recognition, ensuring that model-generated explanations reflect the underlying decision-making mechanism of the network. Extensive experiments on the MAHNOB-HCI and DEAP datasets demonstrate consistently high classification accuracy (97–98%) across arousal, valence, and discrete emotion classification tasks.

This work builds on existing research in explainable AI and multimodal affective computing but uniquely introduces a dynamic evaluation framework for explanation selection.

This work is not under evaluation in a different journal, and no parts of the work have been previously published.