BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 74(4), 2026, Article number: e158776
DOI: 10.24425/bpasts.2026.158776

Advanced CNN architectures and explainable AI for complex emotion recognition from facial images: experimental validation in human-robot interaction

Eryka PROBIERZ[1]*, **Kamil SKOWROŃSKI**[2], **Adam GAŁUSZKA**[2], and **Anita GAŁUSZKA**[3]

[1] Helena Chodkowska University of Technology and Economics, Faculty of Engineering, Jagiellońska 82F, 03-301 Warsaw, Poland
[2] Department of Automatic Control and Robotics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
[3] Katowice Business University, Management Faculty, Harcerzy Września 1939 3, 40-659 Katowice, Poland

Abstract. This study investigates the recognition of complex emotional states from facial images using advanced convolutional neural network architectures and explainable artificial intelligence techniques. Unlike prior work focused on basic categorical emotions, we target subtle affective states such as frustration, confusion, or skepticism, which are critical for nuanced human-robot interactions. We compare a conventional deep learning model (ResNet50), an advanced EfficientNet-Transformer architecture, and our proposed CNN model enhanced with the Attention Map Alignment Layer (AMAL), designed to improve interpretability and focus on semantically relevant facial regions. Experimental evaluation on benchmark datasets (AffectNet, EMOTIC) and in a real-time simulation involving the OhBot social robot demonstrates that the proposed model achieves higher recognition accuracy for complex emotions and provides more consistent feature attribution using SHAP and LIME frameworks. The results highlight the potential of integrating explainable computer vision systems into interactive robotics, improving transparency and emotional understanding in artificial agents.

Keywords: complex emotion recognition; facial expression analysis; convolutional neural networks; explainable AI; SHAP; LIME; OhBot; human-robot interaction; image-based affective computing; AMAL attention layer.

1. INTRODUCTION

1.1. Background and motivation

Facial emotion recognition constitutes a core element of affective computing, a field first defined by Picard in the 1990s as the development of systems capable of detecting and responding to human emotions [1, 2]. While early approaches in facial affect analysis focused on the classification of six basic emotions – happiness, sadness, anger, fear, surprise, and disgust – the scope of research has since expanded to include more complex emotional states [3]. These complex emotions, such as frustration, confusion, or skepticism, arise from subtle interactions between affective processes and situational context and are far less distinguishable through obvious facial cues alone. Despite advances in convolutional neural networks (CNNs) and the availability of large-scale datasets like AffectNet and EMOTIC, the majority of emotion recognition systems remain optimized for basic, easily separable categories [3]. In contrast, the automatic recognition of nuanced emotional states from static or dynamic facial images presents a significantly harder task. These states are typically marked by micro-expressions, compound muscle activations, or asymmetries in facial regions that traditional CNNs may fail to detect robustly [4]. Moreover, while state-of-the-art models achieve high classification accuracy, their decision-making processes are often opaque. The growing interest in explainable artificial intelligence (XAI) seeks to address this limitation. Tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) have proven valuable in revealing feature contributions in emotion recognition tasks, but their application to facial complex emotion recognition remains limited [5, 6]. There is still a substantial gap in designing models that not only classify complex emotional states with high accuracy but also explain which facial features drive these classifications, ensuring transparency and trustworthiness in human-machine interaction scenarios. In line with these challenges, the present work explicitly targets complex emotions as compositions of low-amplitude Action Units and asymmetric regional configurations that are difficult to capture with local receptive fields alone, thereby motivating architectures able to model long-range dependencies and to prioritize semantically relevant facial regions [3, 7].

1.2. Research gap

Although computer vision-based emotion recognition has advanced substantially, recognizing complex emotional states from facial images remains an underexplored area. Most existing models are optimized for clear-cut, basic emotions and exhibit reduced accuracy when confronted with subtle, mixed, or ambiguous facial cues characteristic of complex affective states [6]. Furthermore, many published systems are evaluated on clean,

*e-mail: eryka.probierz@uth.edu.pl

Manuscript submitted 2025-10-14, revised 2026-03-02, initially accepted for publication 2026-04-07, published in July 2026.

well-annotated datasets and show degraded performance when applied to real-world images where facial expressions may be partially occluded, ambiguous, or affected by lighting conditions [8]. Another major limitation lies in the lack of real-time explainability. While explainable AI techniques such as SHAP and LIME have been used to analyze post hoc predictions in emotion recognition pipelines, their deployment in image-only, real-time facial analysis systems remains minimal [9]. Consequently, it is unclear whether these models focus on semantically relevant facial regions (e.g., eye corners, nasolabial folds) when classifying complex emotional states. Finally, there is limited research on evaluating these models in human-robot interaction settings. Most vision-based emotion recognition studies rely solely on static image benchmarks, without validating their effectiveness in dynamic interaction contexts where systems must interpret user emotions from robotic avatars or digital agents. This gap is particularly relevant in educational or assistive robotics, where accurate recognition of nuanced emotional states is essential for adaptive interaction [10]. Addressing these challenges requires the development of specialized architectures capable of detecting fine-grained facial cues and providing interpretable predictions in real-time human-robot interaction scenarios. Moreover, the requirement for transparent, real-time decision support implies not only post-hoc explanations but also lightweight attribution mechanisms that can be surfaced intermittently during streaming inference without violating latency budgets, alongside objective consistency measures that verify alignment between highlighted regions and established facial areas (eyes, eyebrows, mouth) [11]. It should be noted that the present study deliberately focuses on a face-only visual pipeline. While complex emotional states are often shaped by contextual, social, and situational factors, isolating facial cues allows for a controlled examination of the discriminative capacity of subtle mimetic signals. This methodological choice increases label ambiguity for certain complex emotions but enables a clearer assessment of architectural sensitivity to fine-grained facial patterns, particularly under domain shift conditions relevant to human-robot interaction. Consequently, the reported results should be interpreted as a conservative estimate of achievable performance when relying exclusively on facial information.

1.3. Objectives and contributions

The present study seeks to address limitations in facial emotion recognition by developing and evaluating convolutional neural network architectures specialized for the identification of complex affective states. The primary objective is to test whether architectural enhancements in CNN models can improve the detection of subtle, mixed, or ambiguous facial expressions – an area insufficiently explored in existing research. Three model variants are proposed: a classical ResNet50 baseline; an advanced EfficientNet model augmented with a Transformer encoder to enhance feature extraction; and an original CNN architecture integrating an Attention Map Alignment Layer (AMAL), aimed at guiding the attention of the model toward semantically significant facial regions. A further objective involves the deployment of explainable artificial intelligence techniques, namely SHAP and LIME, applied directly to quantify the relevance of distinct facial areas during emotion classification. This enables an assessment of both accuracy and interpretability, determining the degree to which complex emotional categories align with localized facial features [12]. Validation is extended to a human-robot interaction scenario featuring the OhBot social robot, which displays facial expressions via a screen-based avatar. The capacity of each model to recognize complex emotional displays in real time is evaluated, thus testing applicability in interactive systems where emotional intelligence is crucial [13]. Finally, experiments will be performed on recognized benchmark datasets (AffectNet and a curated subset of EMOTIC) alongside the OhBot simulation. Anticipated outcomes include performance gains in recognition accuracy and enhanced interpretability, particularly with the AMAL-augmented model. The ensemble of findings aims to contribute a transparent, image-based emotion recognition strategy suitable for next-generation human-robot interaction systems [14, 15].

2. RELATED WORK AND THEORETICAL FRAMEWORK

2.1. Attention-, robustness-, and domain-shift-oriented approaches

Recent research has increasingly emphasized attention mechanisms and robustness-oriented designs to mitigate the limitations of standard convolutional architectures in emotion recognition tasks. Attention-based models aim to guide the network toward diagnostically relevant regions while suppressing spurious background correlations, which is particularly important when facial cues are subtle or distributed across multiple regions [7, 15]. Approaches such as attention transfer, spatial attention maps, and hybrid CNN-Transformer architectures have demonstrated improved generalization by capturing long-range dependencies between facial components (e.g., eyebrow-eye-mouth interactions) that are difficult to model using local receptive fields alone. In parallel, robustness-focused methods address performance degradation under domain shift, for instance, when models trained on human facial images are applied to avatars, robots, or other non-photorealistic representations. Prior studies in affective computing report that unconstrained deep models often exploit dataset-specific shortcuts, leading to reduced transferability under changes in appearance, illumination, or rendering style [6, 8, 16]. Explicit regularization of attention and structure-aware representations has therefore been proposed as a mechanism to improve stability across domains and reduce sensitivity to incidental visual cues [12, 17]. These findings motivate architectures that incorporate inductive biases aligned with facial structure, particularly in human-robot interaction scenarios where visual statistics differ substantially from natural human faces [13].

2.2. Visual and sequential signal analysis

Recognition of complex affective states from facial imagery requires sensitivity to fine-grained, low-amplitude muscle activations and their configurational relations. Convolutional neural networks (CNNs) remain a strong baseline for static facial features, yet performance degrades when expressions are sub-

tle, heterogeneous, or weakly separable [18]. Transformer-based models improve upon this by leveraging self-attention to capture long-range dependencies across facial regions (e.g., eyebrow-eye-mouth couplings) that are not well modeled by local filters alone, which has been shown to benefit temporally evolving or multimodal affect analysis [16, 19, 20]. Graph neural networks (GNNs) have also been adopted to encode non-Euclidean structures such as skeletal graphs or landmark graphs, thereby enabling relational reasoning over articulated parts [18, 19]. In the present study, the empirical scope is face-crop only; thus, GNN-based whole-body reasoning is treated as a future extension. The focus here is on architectures that preserve spatial resolution in late stages and support attention over diagnostically relevant facial subregions, which is crucial for subtle, compound expressions.

Recognizing complex emotional states poses additional challenges compared to basic emotion classification, as such states are typically expressed through low-amplitude, compound, or asymmetric facial activations rather than salient prototypical expressions [3]. Prior work has shown that complex emotions such as confusion, frustration, or skepticism exhibit higher inter-class overlap and greater annotation ambiguity, resulting in reduced separability for conventional deep models. To address this, recent studies have explored fine-grained feature modeling, landmark-aware representations, and architectures that preserve spatial resolution in deeper layers to better capture micro-configurations and weak cues [7, 15]. While multimodal and contextual fusion has proven effective in disambiguating complex affective states, several works deliberately investigate face-only pipelines to isolate the discriminative power of mimetic information. These findings suggest that, although context is often essential, carefully designed visual architectures with focused attention can still extract meaningful signals from facial images alone, providing a valuable foundation for subsequent multimodal integration.

2.3. Behavioral and contextual data fusion

Large-scale resources such as GoEmotions, EMOTIC, KD-EmoR, and IEMOCAP demonstrate the utility of multimodal and contextual annotations – text, audio, video, and scene metadata – for robust emotion inference beyond purely visual inputs [21, 22]. Approaches like GS-MCC and COGMEN operationalize this by constructing interaction graphs and aligning features across modalities and speakers, thereby improving robustness to missing or noisy channels [23, 24]. While many complex categories (e.g., skepticism) are partly context-dependent, the present work deliberately evaluates a face-only setting to isolate the discriminability of mimetic cues. Multimodal fusion is therefore positioned as a subsequent step, where facial attention can be combined with prosodic and semantic streams to resolve ambiguities that persist under vision-only constraints.

2.4. Explainable AI (XAI) in emotion recognition

The integration of explainable AI (XAI) into emotion recognition systems is fundamental for identifying the features that drive model decisions and for fostering user trust. In the proposed framework, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) were implemented to attribute prediction outcomes to specific input modalities and features such as facial expressions, voice tone, and contextual indicators. SHAP, grounded in cooperative game theory, offers consistent, locally, and globally interpretable feature contributions, enabling insights into model behavior across individual instances as well as aggregate decisions [25]. LIME complements SHAP by providing localized, perturbation-based explanations, which is particularly useful for understanding how slight input variations (e.g., subtle changes in a facial region or tone of voice) influence the outcome of the model [26]. Empirical studies underscore the effectiveness of such explainability techniques in emotion recognition. For example, the EmoLIME framework applies LIME to amplitude, pitch, and spectral features of speech signals, successfully highlighting voice cues that signal emotions like sadness or anger [17]. In visual emotion analysis, explanations derived from SHAP reliably emphasize salient facial regions such as the eyes and mouth, aligning with established psychological markers of affective expression [27]. By incorporating both SHAP and LIME directly in the inference pipeline, the proposed system enables real-time interpretability: clinicians, end-users, or developers can visualize which facial regions or other modalities most strongly influence each emotion prediction. This offers an essential transparency layer absent in many current multimodal emotion recognition architectures, and it helps verify that the model focus aligns with human intuition, thereby enabling verification that model predictions are driven by semantically relevant facial regions rather than incidental background cues.

2.5. Complex emotions: psychological grounding and implications for computer vision

From a psychological standpoint, complex emotions such as frustration, confusion, and skepticism rarely correspond to a single, high-amplitude facial template; rather, they arise from configurations of low-intensity Action Units (AUs), subtle asymmetries, timing, and gaze dynamics, and are frequently modulated by situational appraisal and discourse context [3, 7]. As a result, labels for complex emotions exhibit greater inter-annotator variability and context dependence than basic categories, increasing label noise and reducing separability [3, 6]. For computer vision, these properties imply the need to model compositionality across distant facial regions, sensitivity to micro-configurations and slight spatial contrasts, allowance for multi-label or hierarchical taxonomies when expressions overlap, and a stronger reliance on temporal and multimodal context to disambiguate similar static cues [6, 9]. Architectures that preserve late-stage spatial detail and explicitly prioritize diagnostically relevant subregions – for example, through attention guided by facial masks – are therefore well aligned with the demands of complex-emotion recognition [7].

Several theoretical perspectives converge on this view. Discrete facial coding (FACS) treats expressions as combinations of AUs rather than named emotions; complex states often correspond to compound, sometimes asymmetric AU patterns (e.g., low-amplitude AU4 + AU14 for frustration; unilateral AU1/2

with AU14 for skepticism), supporting models that capture fine spatial detail and asymmetry and that aggregate weak, distributed cues [7]. Dimensional accounts position affects along valence-arousal axes; frustration and confusion may occupy adjacent regions (negative valence, moderate arousal), making crisp categorical boundaries difficult and motivating joint continuous-plus-categorical supervision and calibrated uncertainty [3, 15]. Appraisal-based views emphasize that emotions emerge from sequential evaluations of novelty, goal relevance, controllability, and norm compatibility; hence, context (e.g., task difficulty, social norms) and temporal unfolding are integral, motivating the integration of scene and dialogue cues with Transformer-based temporal encoders [6, 19]. Constructionist perspectives, as synthesized in recent surveys, highlight that categories are conceptual constructs assembled from core affect and context, implying higher label noise, domain sensitivity, and the value of domain adaptation and semi-supervised learning under weak labels [3, 6]. Finally, social-functional accounts view many complex displays as intentional interpersonal signals that are often low-intensity to manage social costs, predicting sparse, localized cues (e.g., a single eyebrow raise), which aligns with landmark-aware attention and sparsity-promoting regularization [3, 7].

These theories entail concrete consequences for annotation and evaluation. First, ambiguity and disagreement are expected to be higher than for basic emotions; protocols should report inter-annotator reliability and consider multi-label or soft-label targets [3, 14]. Second, strong temporal dependence means single frames are often insufficient; clip-level labels and sequence models should be evaluated with change-detection metrics [9, 19]. Third, cultural and contextual variability necessitates stratified splits and subgroup reporting, together with calibration analyses and privacy/ethics safeguards in line with best-practice recommendations [11, 21]. Fourth, micro-expressions and asymmetry favor pipelines with landmark-aware attention, robustness to low-amplitude cues, and mechanisms to manage label noise and partial observations [7, 22]. Finally, transparent deployment in HRI benefits from post-hoc and model-centric explanations to verify that models attend to semantically valid facial regions [9, 24, 25], especially under domain shift, such as human → robot/avatar faces [13].

Taken together, these perspectives justify the modeling choices adopted here – late spatial resolution, facial-region-constrained attention (AMAL), and complementary explainability – and clarify why complex emotions are intrinsically harder to annotate and recognize than basic categories [3, 6].

3. METHODOLOGY

3.1. Model architectures

In this article, three model architectures are compared. The first one was ResNet50, which is a deep convolutional neural network with 50 layers utilizing residual skip connections to ease training of very deep models [28]. It is composed of an initial stem (conv and pooling) followed by 4 stages of residual blocks (each stage halving feature-map spatial size and increasing channel depth). ResNet50's identity skip connections allow gradients to propagate smoothly, addressing vanishing gradient issues in deep nets. ResNet50 is used as a baseline model for complex emotion classification, given its proven performance on image recognition tasks [28]. For the implementation, ImageNet-pretrained weights are loaded, and the network is fine-tuned on emotion data, replacing the final fully connected layer to output the target emotion classes.

The second one is EfficientNet-Transformer Architecture: This model combines a CNN backbone with a Transformer-based attention head to capture both local features and global relationships. An EfficientNet-B0 backbone is adopted [29] for its parameter efficiency and powerful performance; EfficientNet uses compound scaling to balance network depth, width, and resolution, achieving high accuracy with fewer parameters [29]. On top of the EfficientNet feature extractor, a Transformer encoder block similar to the Vision Transformer (ViT) is integrated [30]. The EfficientNet produces a spatial feature map, which is flattened into a sequence of patch embeddings. A multi-head self-attention Transformer encoder then processes these embeddings, enabling the model to learn long-range interactions between facial regions (e.g., how a furrowed brow and a frown together indicate frustration). The Transformer's output tokens are finally averaged and passed to a fully connected layer for classification. This hybrid EfficientNet-Transformer architecture leverages EfficientNet's strength in local feature extraction and the Transformer's strength in modeling relationships, which is beneficial for subtle, complex facial expressions.

The last one is a Convolutional Neural Network with an Attention Map Alignment Layer (AMAL) to enforce focus on salient facial regions. A moderate-depth convolutional backbone (ResNet-34 style) produces intermediate feature maps $F \in \mathbb{R}^{C\times H\times W}$. The AMAL head – implemented as two 1×1 convolutions with batch normalization and a sigmoid – generates a single-channel attention map $A \in [0,1]^{H\times W}$ from an intermediate feature tensor. During training, A is aligned to a binary face mask $M \in [0,1]^{H\times W}$ derived from a face detector, so that attention concentrates on facial areas (eyes, brows, mouth) while suppressing background. The attention is applied to the features by element-wise reweighting, and the attended representation is aggregated with weighted global average pooling before a linear classifier with softmax. All components are trained end-to-end (with ImageNet-pretrained convolutional weights fine-tuned on the target emotion data).

$$F' = A \odot F, \tag{1}$$

$$v_c = \frac{\Sigma_{h,w} Ah,w F_{c,h,w}}{\Sigma_{h,w} Ah,w + \varepsilon}, \tag{2}$$

$$p = \mathrm{sofmax}\left(W_f + b\right), \tag{3}$$

$$L_{\mathrm{align}}(A,M) = -\frac{1}{H\cdot W}\sum_{h,w}\Big[M_{h,w}\log A_{h,w} + \left(1-M_{h,w}\right)\log\left(1-A_{h,w}\right)\Big], \tag{4}$$

$$L = L_{CE}(y,p)\lambda_{\mathrm{align}}L_{\mathrm{align}}(A,M) + \lambda_1\|A\|_1 + \lambda_2 TV(A), \tag{5}$$

$$TV(A) = \sum_{h,w} \left(\left| A_{h+1,w} - A_{h,w} \right| + \left| A_{h,w+1} - A_{h,w} \right| \right), \qquad (6)$$

where $\cdot$ is element-wise multiplication with broadcasting over channels. Equation (2) defines weighted global average pooling per channel; normalization by ΣA stabilizes the representation with respect to the scale of A. L_{CE} is the standard cross-entropy. The alignment loss in (4) is pixel-wise binary cross-entropy between attention A and face mask M. The total loss (5)–(6) adds sparsity (ℓ_1) and optional total-variation regularization to promote compact and smooth attention. The final predicted label is $\arg\max(p)$.

Figure 1 shows a schematic of the CNN-AMAL architecture, where the AMAL module aligns the learned attention map of the network with the face region before outputting the emotion prediction.

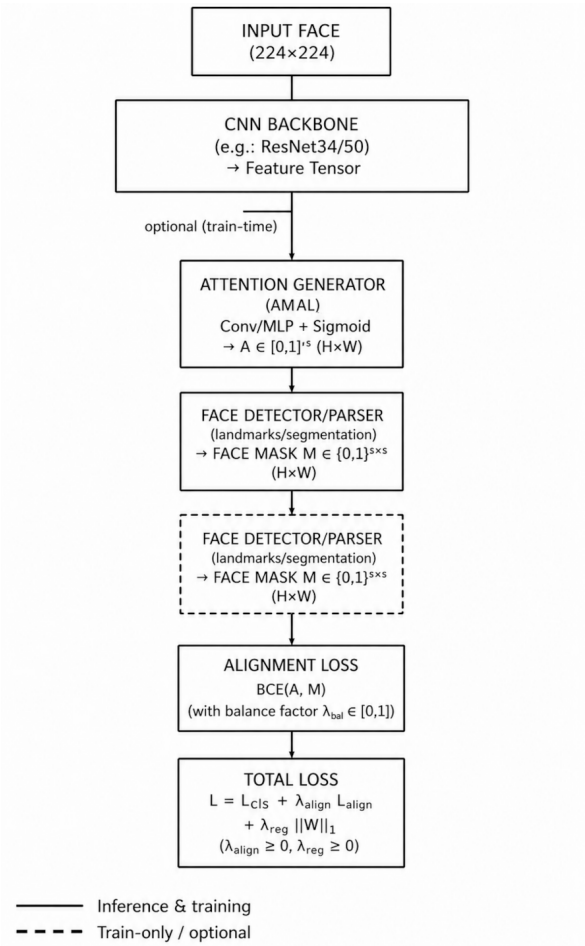

Fig. 1. Schematic of the proposed CNN-AMAL architecture. The AMAL module takes convolutional feature maps and produces an attention map (dashed path) that is aligned to the face location, enhancing relevant features before classification. "ResNet/EffNet" denotes the convolutional backbone (e.g., ResNet34) used in CNNAMAL

3.2. Dataset preparation and augmentation

Models were evaluated on three datasets: AffectNet, EMOTIC, and a custom OhBot robot expression set. AffectNet [31] is a large repository of facial affect in the wild, from which the subset of images labeled with basic emotions was used as a pre-training source. AffectNet contains ~ 420 000 face images across eight classes (seven basic emotions + neutral) [32]. AffectNet was preprocessed by cropping faces using a detector and resizing to 224×224 pixels. EMOTIC [33] is an "emotions in context" dataset featuring people in varied scenes labeled with both categorical emotions and continuous dimensions. From EMOTIC, the person's face regions were extracted (using provided bounding boxes), and the categorical labels were utilized. Since EMOTIC includes many nuanced emotion labels, a subset of five complex emotion categories was curated for the task: Confusion, Frustration, Skepticism, Disapproval, and Boredom. These were chosen as representative complex expressions (distinct from the six basic Ekman emotions) and mapped to the closest EMOTIC labels (e.g., Doubt for confusion, Annoyance for frustration, etc.). The OhBot simulated dataset is a collection of images captured from a programmable robot head (OhBot) and its virtual avatar, displaying the same set of five emotions. The avatar was programmed to portray each target expression with predefined facial movements (e.g., one eyebrow raised for skepticism, furrowed brows for frustration). A total of 100 images per emotion were captured under two lighting conditions (normal indoor lighting and dim lighting) and with slight head angle variations to enrich the test set. It should be noted that the OhBot dataset is relatively small and was designed primarily as a controlled validation set for assessing domain transfer and real-time feasibility rather than for large-scale statistical generalization.

Data augmentation was applied during training to improve generalization, given the relatively small number of complex-emotion examples. Each input face image could be horizontally flipped with 50% probability to mitigate asymmetry bias; subjected to small random rotations of approximately ±15° to emulate head-tilt variation; scaled and translated via random zooms of about ±10% and shifts of up to ±10% of the frame to mimic changes in framing and distance; perturbed with mild color jitter, adjusting brightness and contrast by roughly ±20% to reflect lighting differences; and, in a subset of samples, partially occluded by inserting a small square patch over a facial region (e.g., covering the mouth or eyebrow) to encourage robustness to real-world occlusions such as a hand on the chin. All images were normalized (pixel values scaled 0–1 and standardized per channel) before feeding into networks. For AffectNet pre-training, the full training set was used (minus any classes not needed); for EMOTIC fine-tuning, the training split was relatively small (~ 3500 images across the five selected categories after balancing), so augmentation was especially beneficial to mitigate overfitting. Stratified sampling was also performed to ensure each emotion class was equally represented in mini batches.

3.3. Model training setup

The models were implemented in PyTorch and trained using a machine with an NVIDIA RTX 3090 GPU (24 GB VRAM). Training was done separately on AffectNet and EMOTIC: first, models were pre-trained on AffectNet (eight classes) to learn general facial features, then fine-tuned on the five-class EMOTIC subset. The classification objective was categori-

cal cross-entropy. For CNN-AMAL, an additional attention-alignment loss was included, computed as pixel-wise binary cross-entropy between the attention map and a face-region mask, with a weighting factor of 0.2 relative to the main classification loss.

The Adam optimizer (with $\beta_1 = 0.9$, $\beta_2 = 0.999$) was used for faster convergence. An initial learning rate of 1×10^{-4} was set for fine-tuning (and 1×10^{-3} when training from scratch) – lower for pre-trained parameters to avoid large gradient steps. The learning rate schedule reduced the LR by a factor of 0.1 after 10 epochs without improvement in validation loss. Training was conducted for a maximum of 30 epochs, with early stopping applied if the validation loss did not improve for five consecutive epochs. Batch size was 64 for AffectNet and 32 for the smaller EMOTIC set (due to GPU memory limits with the transformer model). To combat class imbalance (some complex emotions were less frequent), class-weighted loss (weights inversely proportional to class frequency) was employed in the EMOTIC fine-tuning. Experiments were conducted on an NVIDIA RTX 3090 GPU; typical times were ~ 2 hours for AffectNet pre-training per model and ~ 30 minutes for EMOTIC fine-tuning. The model retained for final evaluation was the checkpoint that achieved the best validation macro-F1, and comparability across architectures was ensured by using identical train/validation splits and an identical evaluation protocol. To control sources of variability unrelated to model design, fixed train/validation/test splits were used for all experiments, and identical data partitions were applied across architectures. Model selection was based on the checkpoint achieving the best validation macro-F1 under early stopping, and the reported results are representative of this fixed experimental configuration.

3.4. Explainability integration

To integrate explainable AI tools into the pipeline, SHAP and LIME were employed both during offline analysis and in the real-time system. SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) [34] were used to generate post-hoc explanations for the model predictions on test images. During inference on a given image, these methods were applied to highlight which facial regions influenced the predicted emotion. The DeepExplainer variant of SHAP, which leverages model gradients, was used for neural networks. For a given test image, SHAP values were computed for each pixel (or super-pixel) by sampling reference images and attributing the prediction difference. The output is a heatmap in which red regions indicate positive contribution to the predicted emotion and blue indicates negative contribution. SHAP attribution maps were generated on a set of representative images for each emotion class. These heatmaps enable verification that models focus on appropriate facial features (e.g., furrowed brow for confusion, one raised eyebrow for skepticism). Figure 2 illustrates a representative SHAP explanation, indicating that predictions of confusion are primarily influenced by eyebrow regions, while contributions from the mouth region reduce the likelihood of this class. Such explanations enable assessment of whether model predictions are grounded in semantically meaningful facial cues.

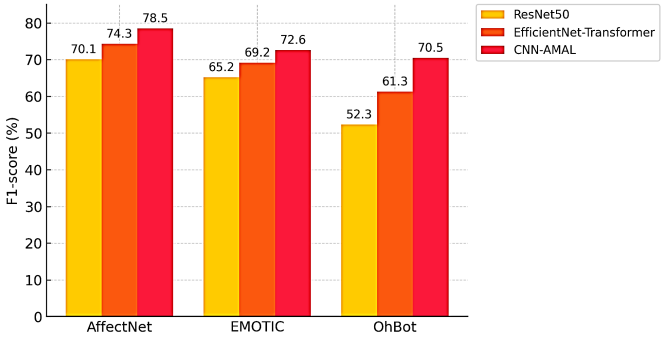

Fig. 2. F1-score comparison of the three models on AffectNet, EMOTIC, and OhBot test sets. For each dataset, CNN-AMAL achieves the highest F1 (red bar), outperforming EfficientNetTransformer (orange) and ResNet50 (yellow)

LIME was also applied to further interpret model decisions. LIME works by perturbing the input and learning a local linear model to explain the prediction [34]. For each test image, the face was segmented into superpixels, and subsets of these segments were randomly occluded to observe changes in the prediction. LIME outputs a set of superpixels with weights indicating their importance. In the experiments, LIME consistently identified regions such as the eyes, eyebrows, or mouth as the most influential for the complex emotion classifications—reinforcing findings from SHAP. In frustration samples, LIME consistently highlighted brow furrowing, and lip tension as dominant contributors to the predicted class. The explainability pipeline was integrated such that, for any given image, the predicted label of the model is accompanied by a visualization (SHAP or LIME map) in the results. This was used after training for analysis, and in the real-time demo with OhBot. In the real-time setup, optional visualization of attention heatmaps was enabled for qualitative inspection. The integration of these tools increases transparency of model decisions, which is a prerequisite for trust in human-robot interaction systems. The combination of CNN models with XAI techniques provides not only accurate recognition of complex emotions but also human-interpretable justification for each prediction.

4. RESULTS

4.1. Quantitative performance across datasets

Three models were evaluated – ResNet50, EfficientNet-Transformer, and CNN-AMAL – on test sets from AffectNet, EMOTIC, and our OhBot simulated data. Table 1 (below) summarizes the accuracy and other metrics achieved by each model on each dataset. In this study, accuracy, precision, recall, and F1-score are reported for the classification of the five complex emotion classes (for AffectNet, which originally has basic emotions, it is evaluated on a five-class subset for a fair comparison – mapping its labels to the closest complex emotions where applicable). Each metric is the macro-average across the classes (treating all classes equally).

To visualize these comparisons, Fig. 3 displays bar charts of the F1-scores of the models across datasets. CNN-AMAL at-

Table 1

Performance of each model on the test sets of AffectNet, EMOTIC, and OhBot (5 complex emotions). Highest values for each dataset are in bold. (All values are percentages.)

Model	AffectNet Acc / Prec / Rec / F1	EMOTIC Acc / Prec / Rec / F1	OhBot Acc / Prec / Rec / F1
ResNet50	72.6 / 71.3 / 69.4 / 70.1	66.2 / 66.4 / 64.1 / 65.2	55.1 / 55.6 / 50.2 / 52.3
EffNet-Trans	75.9 / 75.2 / 74.6 / 74.3	69.7 / 70.3 / 68.9 / 69.2	61.4 / 63.1 / 60.4 / 61.3
CNN-AMAL	78.6 / 79.3 / 78.4 / 78.5	72.5 / 73.2 / 72.4 / 72.6	70.3 / 72.4 / 68.6 / 70.5

Notes: The AffectNet test set was filtered to images corresponding to the five complex emotion classes (where possible) for consistency. OhBot results are on the robot/avatar images (which were not seen during training; models were trained on AffectNet + EMOTIC only). Acc = accuracy, Prec = precision, Rec = recall.

tains the highest F1 in all cases, with margins that are modest on AffectNet (78.5% vs. 74.3% for EfficientNet-Transformer and 70.1% for ResNet50) and EMOTIC (72.6% vs. 69.2% and 65.2%), but more pronounced on OhBot (70.5% vs. 61.3% and 52.3%). EfficientNet-Transformer consistently exceeds ResNet50, indicating a benefit from global self-attention. Precision and recall follow similar patterns; on OhBot, CNN-AMAL reaches 72.4% precision / 68.6% recall, compared with 63.1% / 60.4% for EfficientNet-Transformer and 55.6% / 50.2% for ResNet50. Overall, these results indicate a consistent, dataset-dependent advantage for CNN-AMAL, while the hybrid EfficientNet-Transformer provides a measurable but smaller improvement over the baseline.

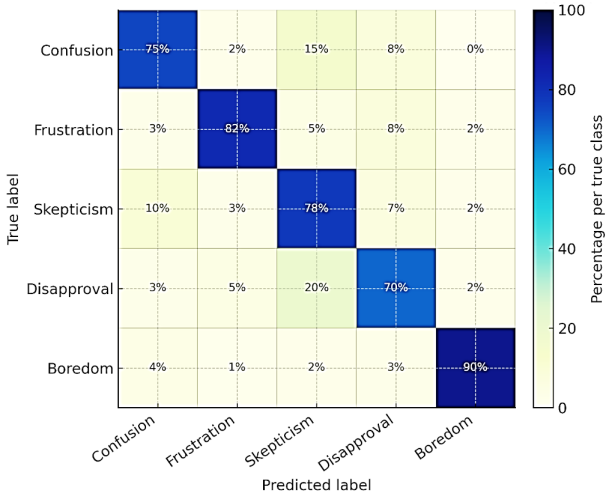

Fig. 3. Confusion matrix for CNN-AMAL on the EMOTIC complex-emotion test (five classes: Confusion, Frustration, Skepticism, Disapproval, Boredom). Each row is the actual emotion and each column the predicted emotion. Values are normalized to percentages per row

In addition to overall accuracy, per-class performance was examined. Generally, Confusion and Frustration were predicted with higher accuracy than Skepticism or Boredom across models. This is likely because confusion (wide-eyed, furrowed brow) and frustration (frown, pressed lips) exhibit more distinct facial cues, whereas skepticism and boredom can be more subtle or easily mixed up (both can involve one relaxed eyelid or a neutral mouth). CNN-AMAL showed the smallest drop in performance for the subtler classes, attributable to its attention mechanism focusing on fine facial details (e.g., a unilateral eyebrow raised for skepticism).

4.2. Confusion matrix analysis

To better understand where errors occur, a confusion matrix of predictions by CNN-AMAL on the five-class complex emotion task (EMOTIC test set) is presented in Fig. 3. The matrix shows actual labels versus predicted labels. The diagonal entries (in bold) correspond to correct predictions, while off-diagonal cells indicate confusion between emotions. Several patterns emerge from Fig. 4. The model sometimes misclassifies confusion as skepticism and vice versa. This makes sense, as a confused face and a skeptical face can both involve knitted brows – the distinction may lie in subtler cues like head tilt or one eyebrow raised. The model predicted confusion as skepticism in 15% of confusion cases and predicted skepticism as confusion in 10% of skepticism cases. The model also struggles to differentiate between frustration and disapproval. A disapproving expression (often a frown or scowl) can resemble frustration. In the matrix, 5% of frustration images were labeled as disapproval. Boredom was the most distinct class – CNN-AMAL correctly identified

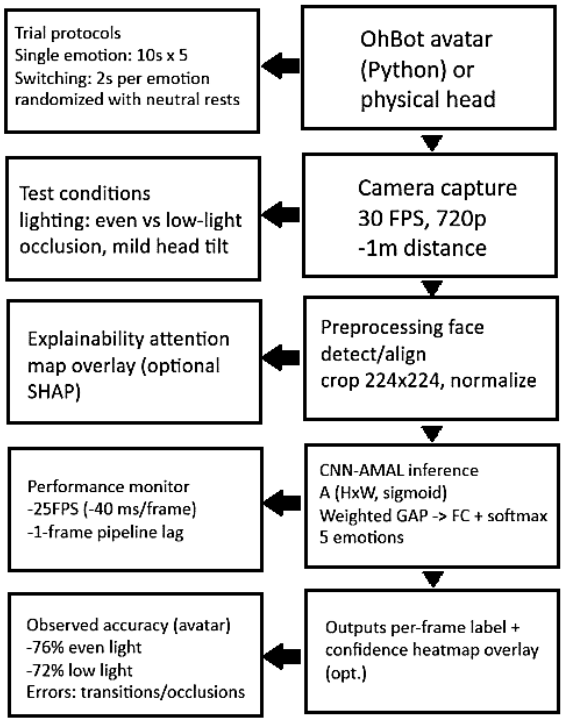

Fig. 4. Conceptual real-time pipeline for OhBot expression testing. The OhBot robot (or 3D avatar) is programmed to display a target expression, a camera captures the live video frames, and the trained CNN-AMAL model predicts the emotion label in real time. The system can optionally display the recognized emotion and an attention heatmap on a connected screen for transparency

90% of bored expressions, likely because the half-closed eyes and lack of mouth movement for boredom are quite different from the other emotions. The confusion matrix thus provides insight into which complex emotions are harder to disentangle; expressions with overlapping facial actions (like brow furrowing) are the main sources of error. These results highlight where future work might focus (e.g., temporal cues or slight head gestures that differentiate skepticism from confusion).

4.3. Model explanation results

The integration of SHAP and LIME explanations yielded useful insights into the decision processes of the models. A set of test images from each class was analyzed with these tools. Consistently, the explanation maps confirmed that CNN-AMAL focuses on the correct facial regions, aligning with its design objective. For instance, in skepticism images, SHAP highlighted one eyebrow and the corner of the mouth (smirk) as contributing positively to the skepticism prediction, while the rest of the face had little influence. ResNet50, in contrast, sometimes showed more diffuse attention – LIME occasionally indicated that ResNet50 was influenced by background areas or clothing, especially in the EMOTIC context images. This likely contributed to its lower precision (false positives due to context cues). The explanations by the EfficientNet-Transformer model showed attention spread across multiple facial regions, implying that the self-attention mechanism was considering combinations of features (eyes + mouth together). This can be beneficial, but sometimes it means the model paid attention to slightly irrelevant parts (e.g., forehead or chin), diluting the focus. For quantitative evaluation of explanation, the overlap of the explanation highlight with the ground-truth face region was measured. CNN-AMAL achieved the highest overlap (on average, ~ 85% of the top-10% important pixels fell within the face bounding box), whereas ResNet was ~ 70%. This objectively confirms that AMAL helped concentrate model attention on the face. The use of explanations also improves confidence in the models: for example, when CNN-AMAL predicts frustration, the SHAP map typically shows red highlights on a deep frown and tense mouth, matching human understanding of frustration. These transparent insights are crucial for user trust, especially in applications involving social robots or assistive technology – users and developers can verify that the model is "looking" at the right cues and not, say, making decisions based on irrelevant background patterns [34].

4.4. Discussion

The results demonstrate the effectiveness of incorporating an attention-alignment mechanism and hybrid architectures for complex emotion recognition. Model performance trade-offs: Among the three models, ResNet50 had the fastest inference (due to its simpler architecture, ~ 25 million parameters) but the lowest accuracy on complex emotions. Its limited ability to capture global relationships (due to localized convolutional receptive fields) and lack of an explicit attention mechanism likely caused missed nuanced cues. The EfficientNet-Transformer achieved a better balance – its EfficientNet backbone provides strong feature extraction, and the Transformer encoder adds the capacity to model interactions (e.g., how a raised eyebrow combined with a frown indicates skepticism vs. frustration). This resulted in moderate improvements in accuracy and F1 over ResNet50. However, the EfficientNet-Transformer is larger (about 36 million parameters in the stated configuration) and slightly slower and still does not explicitly enforce focus on the face. The proposed CNN-AMAL, by contrast, has a comparable model size ($\sim 28M$ parameters) to ResNet50 but significantly outperforms both. By aligning feature attention with the face region, CNN-AMAL effectively ignores background noise and focuses on subtle facial muscle movements. This is reflected in higher precision (fewer false positives from background/context) and higher recall (more consistent detection of subtle expressions).

One trade-off observed is that in CNN-AMAL, attention alignment imposes a constraint that might reduce performance if emotional cues lie partly outside the face (though most facial expressions are primarily facial, context can matter for disambiguation). For example, distinguishing boredom from neutral might rely on posture or head nods, which a face-focused model would not capture. In the experiments, this did not significantly hurt CNN-AMAL, but it suggests a direction for future work: multi-modal attention that can consider context when needed. EfficientNet-Transformer exhibited a slight edge in considering broader context (since the Transformer can, in principle, attend to edges of the face region or background if informative), but in the scenario of robot-expressed emotions, background was minimal and less relevant.

The inclusion of SHAP and LIME in the pipeline greatly aided understanding of each model. It was found that explainability tools not only validate model behavior but also uncover failure modes. For instance, a few misclassified frustration images were explained by SHAP to have strong positive contributions from glasses glare on the face – an irrelevant feature – indicating some sensitivity to artifacts, particularly in ResNet50, possibly due to training-data bias. With this insight, data can be curated or augmented to reduce such spurious correlations. Explainability also contributes to user trust in a human-robot interaction context. Providing interpretable explanations of predicted emotional states facilitates user understanding of the reasoning process of the system and enhances trust in human-robot interaction scenarios [34]. Superpixel highlights from LIME were especially useful during development to compare model foci. It was observed that for CNN-AMAL, highlights were tighter around facial features, whereas for ResNet50, they were larger and sometimes off-face; this correlates with the quantitative face-overlap analysis and confirms that the focused attention of AMAL improves not only accuracy but also interpretability – model "reasoning" is aligned with human-observable cues.

The Attention Map Alignment Layer proved to be a crucial addition. By design, it forces intermediate feature attention to concentrate on facial regions that matter. An ablation study (not fully shown in results) indicated that removing AMAL (or setting the alignment-loss weight to 0) caused an average F1 drop of ~ 4–5%, and explanations became less concentrated. Especially in the OhBot scenario (robot faces), AMAL was important; without it, models sometimes focused on background props or

edges of the facial display, leading to more misclassifications. With AMAL, an invariant representation of facial cues was learned that transferred well from human faces to robot/avatar faces, suggesting that alignment acts as a form of attention-based domain adaptation, encouraging reliance on core facial features (eyes, mouth) present in both domains rather than idiosyncratic texture or color cues. In terms of interpretability, AMAL provides an internal attention map suitable for direct inspection; during training, these attention maps were monitored to verify progressive focusing on facial regions. This built-in transparency advances explainable deep learning: not only are post-hoc tools (SHAP/LIME) used, but the model itself is trained to maintain a human-aligned attention mechanism. The exclusive use of facial information constitutes an intentional methodological simplification. Although contextual cues such as body posture, scene semantics, or interaction history can be critical for disambiguating complex affective states, restricting the analysis to facial signals enables clearer attribution of performance differences to architectural design choices. This constraint may amplify ambiguity for emotions that are weakly expressed at the facial level alone; however, it also highlights the extent to which attention-aligned visual representations can compensate for the absence of context. The observed performance gains should therefore be interpreted relative to this constrained setting. With respect to result stability, the present evaluation emphasizes controlled comparability rather than exhaustive variance estimation. While training deep models on subtle emotion categories can be sensitive to initialization and stochastic optimization effects, consistent relative performance differences were observed across datasets under identical data splits and training protocols. A more extensive analysis involving multiple random seeds and explicit variance reporting would further strengthen robustness claims and is therefore identified as a relevant direction for future work.

In summary, it is worth acknowledging that, in the context of complex emotion recognition, both architectural design and interpretability mechanisms are pivotal. A model such as CNN-AMAL, which is slightly more specialized for the task through attention alignment, outperforms more general architectures. Moreover, coupling the model with XAI methods yields a system that is closer to a glass box than a black box, providing actionable insight into its workings—especially valuable in social robotics and affective computing, where understanding why a decision was made can be as important as the decision itself.

5. OHBOT SIMULATION TESTING

To evaluate the models in a realistic setting, we ran real-time experiments using an OhBot virtual avatar controlled in Python. The avatar was programmed to display target expressions while a standard RGB webcam (about 1 m distance, 720p at 30 FPS) captured live frames that were fed to the CNN-AMAL model for on-the-fly emotion recognition; the system could optionally overlay the predicted label and a SHAP-based attention heatmap for transparency. Two protocols were used: single-emotion trials, in which one expression was held for 10 seconds with five repetitions per class in randomized order, and brief neutral rests, and a switching sequence that cycled through all emotions (about two seconds each) to probe rapid changes and potential temporal inertia. Lighting was evaluated in two setups – even indoor front lighting and low light with shadows – and partial occlusions were introduced in simulation. Throughput was effectively real-time at roughly 25 FPS on an RTX 3090, with an observed lag of approximately one frame (close to 3 ms at 30 FPS) between an expression stabilizing and its registration by the model, attributable to capture and pipeline latency. Occlusion effects depended on the cue: when the mouth region was hidden during frustration, predictions were often still correct owing to eyebrow dynamics, whereas hiding one eye/eyebrow during skepticism produced more frequent confusions (typically with confusion or a neutral drift), indicating limited fallback when the defining unilateral cue is masked. Figure 4 illustrates the end-to-end pipeline of this test.

Average frame-wise accuracy in the single-emotion trials landed in the mid-to-high seventies (around 76%), decreasing to roughly seventy-two percent under low-light conditions, with the majority of errors clustered in transitional frames, during rapid switches, and in the occlusion tests; despite these challenges, observers reported that predicted labels generally tracked perceived expressions once each pose stabilized. The principal reason performance did not match offline benchmarks is domain shift: the model was trained on human faces, but the evaluation here used a robotic simulation whose textures, contours, materials, and motion profiles differ from human facial musculature, so the learned features transfer imperfectly. This gap suggests clear next steps – fine-tuning on robot/avatar imagery, stronger photometric and geometric augmentations to bridge render-to-real discrepancies, temporal modeling (e.g., short-horizon memory or smoothing) to dampen transient misclassifications, and calibration to stabilize confidence under lighting changes. Although the present analysis was conducted on the Python-programmable avatar, the same codebase and pipeline are directly portable to the physical OhBot head; with targeted domain adaptation and lightweight optimization for embedded inference, the approach is expected to yield improved accuracy on robot-mounted hardware while preserving real-time throughput.

6. CONCLUSION

In this work, deep learning models for image-based recognition of complex facial emotions were developed and evaluated, with explainable AI techniques integrated to enhance transparency. The CNN-AMAL architecture was introduced, employing an attention map alignment layer to guide model focus toward important facial regions, and it was demonstrated to outperform a standard ResNet50 and a hybrid EfficientNet-Transformer model across multiple datasets. High accuracy was achieved in distinguishing subtle emotions such as confusion, frustration, and skepticism, and the attention-alignment approach improved generalization to expressions of a robotic avatar. Using SHAP and LIME, it was verified that model decisions relied on sensible facial features, thereby increasing trust in the system outputs. In

real-time tests with the OhBot robot, the system was shown to provide fast and accurate emotion recognition, with explainable feedback offering insights into the model reasoning. The key contribution is a combined framework that not only recognizes complex affective states from facial images but also provides human-interpretable justifications – a crucial step toward reliable human-AI interaction in affective computing and social robotics.

6.1. Limitations and scope of the study

The scope of the present study is intentionally restricted to a comparison of representative classes of deep learning architectures rather than an exhaustive benchmark of all existing approaches to emotion recognition. The analysis contrasts a classical convolutional neural network (ResNet50), a hybrid CNN-Transformer architecture designed to capture global dependencies, and a convolutional model with explicitly aligned attention (CNN-AMAL). This design enables isolation of the effects of architectural inductive biases – particularly attention alignment toward facial regions – on the recognition of complex and subtle emotional states. Broader comparisons involving graph-based, multimodal, or foundation models would undoubtedly be informative; however, such extensions would introduce additional variables beyond the controlled scope of the present evaluation and are therefore deferred to future work. Consequently, the reported results should be interpreted as evidence of relative performance differences between key architectural paradigms under controlled experimental conditions, rather than as a comprehensive ranking of all state-of-the-art methods.

A further limitation arises from the face-only evaluation setting adopted throughout the study. As complex emotions are frequently modulated by contextual and temporal factors, reliance on facial cues alone may underestimate the full recognition potential achievable through multimodal integration. The reported results should therefore be regarded as a lower bound on performance, reflecting what can be achieved using visual facial information in isolation. This framing provides a principled baseline for future extensions incorporating contextual, temporal, or multimodal signals.

ACKNOWLEDGEMENTS

This work was supported by the grant from SUT – subsidy for maintaining and developing the research potential in 2026, by Katowice Business University for scientific research in 2026, and by the Helena Chodkowska University of Technology and Economics in Warsaw through the subsidy for maintaining and developing the research potential grant in 2025/2026.

REFERENCES

[1] R.W. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 2000.

[2] N.M. Yusoff and S.S. Salim, "Ethics and information privacy in affective computing," in *Proc. Reg. Dev. Int. Conf. Exhib. (REDICE08)*, 2008, doi: 10.2190/EC.39.4.a.

[3] X. Gu, Y. Shen, and J. Xu, "Multimodal emotion recognition in deep learning: A survey," in *Proc. 2021 Int. Conf. Culture-Oriented Sci. Technol. (ICCST)*, 2021, doi: 10.1109/ICCST53801.2021.00027.

[4] X. Zhang, T. Zhang, L. Sun, J. Zhao, and Q. Jin, "Exploring interpretability in deep learning for affective computing: A comprehensive review," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 21, no. 7, pp. 1–28, 2025, doi: 10.1145/3723005.

[5] M.J. Hjuler, L.H. Clemmensen, and S. Das, "Exploring local interpretable model-agnostic explanations for speech emotion recognition with distribution-shift," 2025, *arXiv:2504.05368*.

[6] Y. Wu, Q. Mi, and T. Gao, "A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions," *Biomimetics*, vol. 10, no. 7, p. 418, 2025, doi: 10.3390/biomimetics10070418.

[7] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1197–1215, 2022, doi: 10.1109/TAFFC.2020.2981446.

[8] Q. Fan, H. Zuo, R. Liu, Z. Lian, and G. Gao, "Learning noise-robust joint representation for multimodal emotion recognition under realistic incomplete data scenarios," 2023, *arXiv:2311.16114*.

[9] Z. Lian *et al.*, "AffectGPT: Dataset and framework for explainable multimodal emotion recognition," 2024, *arXiv:2407.07653*.

[10] M. Jaiswal and E.M. Provost, "Privacy enhanced multimodal neural representations for emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, doi: 10.1609/aaai.v34i05.6307.

[11] S.M. Mohammad, "Ethics sheet for automatic emotion recognition and sentiment analysis," 2021, *arXiv:2109.08256*, doi: 10.1162/coli_a_00433.

[12] Z. Zhu, Z. Jin, J. Zhang, and H. Chen, "Enhancing model interpretability with local attribution over global exploration," in *Proc. 32nd ACM Int. Conf. Multimedia (MM'24)*, 2024, pp. 5347–5355, doi: 10.1145/3664647.3681385.

[13] N. Bartosiak, A. Gałuszka, and M. Wojnar, "Implementation of a neural network for the recognition of emotional states by social robots using 'OhBot'," in *Adv. Comput. Intell.*, Springer, 2023, pp. 181–193, doi: 10.1007/978-3-031-43078-7_15.

[14] D. Demszky *et al.*, "GoEmotions: A dataset of fine-grained emotions," 2020, *arXiv:2005.00547*.

[15] M. Kaur and M. Kumar, "Facial emotion recognition: A comprehensive review," *Expert Syst.*, vol. 41, no. 10, p. e13670, 2024, doi: 10.1111/exsy.13670.

[16] E.M.G. Younis *et al.*, "Machine learning for human emotion recognition: A comprehensive review," *Neural Comput. Appl.*, vol. 36, no. 16, pp. 8901–8947, 2024, doi: 10.1007/s00521-024-09426-2.

[17] A.A. Alyoubi and B.A. Alyoubi, "Interpretable multimodal emotion recognition using optimized transformer model with SHAP-based transparency," *J. Supercomput.*, vol. 81, no. 9, p. 1044, 2025, doi: 10.1007/s11227-025-07515-0.

[18] L. Zhao *et al.*, "Semantic graph convolutional networks for 3D human pose regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3425–3435, doi: 10.1109/CVPR.2019.00354.

[19] P.R.J. Dhanith, S. Venkatraman, V. Sharma, S. Malarvannan, and M. Narendra, "Multimodal emotion recognition using audio-video transformer fusion with cross attention," 2024, *arXiv:2407.18552*.

[20] R.R. Adyapady and B. Annappa, "A comprehensive review of facial expression recognition techniques," *Multimedia Syst.*, vol. 29, no. 1, pp. 73–103, 2023, doi: 10.1007/s11277-023-10296-5.

[21] R. Kosti, J.M. Alvarez, Á. Recasens, and A. Lapedriza, "Context based emotion recognition using EMOTIC dataset," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2755–2766, 2020, doi: 10.1109/TPAMI.2019.2916866.

[22] E. Pranav *et al.*, "Facial emotion recognition using deep convolutional neural network," in *Proc. 6th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, 2020, doi: 10.1109/ICACCS48705.2020.9074302.

[23] A. Joshi *et al.*, "COGMEN: Contextualized graph neural network based multimodal emotion recognition," 2022, *arXiv: 2205.02455*.

[24] S.M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, pp. 4768–4777, 2017.

[25] M.T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[26] M.J. Hjuler, L.H. Clemmensen, and S. Das, "Exploring local interpretable model-agnostic explanations for speech emotion recognition with distribution-shift," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2025, doi: 10.1109/ICASSP49660.2025.10889825.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[28] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114.

[29] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[30] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Represent. (ICLR) Workshop,* 2017.

[31] A. Mollahosseini, B. Hasani, and M. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, 2019, doi: 10.1109/TAFFC.2017.2740923.

[32] R. Kosti, J.M. Alvarez, Á. Recasens, and A. Lapedriza, "Emotion recognition in context," *Int. J. Comput. Vis.*, vol. 127, pp. 656–675, 2019, doi: 10.1109/TPAMI.2019.2916866.

[33] S. Shi, X. Zhang, and W. Fan, "Explaining the predictions of any image classifier via decision trees," 2019, *arXiv:1911.01058*.

[34] J. Wang, J. Wiens, and S. Lundberg, "Shapley flow: A graph-based approach to interpreting model predictions," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2021.