Received: 11 April 2026 | Revised: 2 September 2026 | Accepted: 20 September 2026 | Published online: 20 October 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–13
DOI: 10.47852/bonviewAIA62029975

BON VIEW PUBLISHING

A Lightweight Hybrid CNN–Multi-Head Attention Framework for Multi-Dataset Facial Emotion Recognition

Ayman Odeh[1,*], Nada Odeh[2] and Munther Oder[3]

[1] *College of Engineering, Al Ain University, UAE*

[2] *College of Information System, United Arab Emirates University, UAE*

[3] *Mathematics and Natural Sciences, Carl von Ossietzky Universitaet Oldenburg, Germany*

Abstract: Facial emotion recognition requires a model to identify small local changes in facial appearance while also relating information from spatially separated regions. We developed a lightweight architecture that combines a convolutional neural network (CNN) with multi-head self-attention (MHSA). Three convolutional blocks extract local features from 48 × 48 grayscale images. The resulting 6 × 6 feature map is converted into 36 tokens and processed by a two-head MHSA module before classification. Attention is therefore used to refine intermediate CNN features rather than replace the convolutional backbone. The model was independently trained and evaluated on CK+, FER2013, RAF-DB, and AffectNet. Subject-independent five-fold cross-validation on CK+ produced a mean accuracy of 97.76% and a macro F1-score of 96.94%. Compared with the matched CNN baseline, the proposed configuration improved accuracy by 3.20 percentage points on FER2013, 3.50 points on RAF-DB, and 3.90 points on AffectNet. The ablation results showed that two attention heads performed better than one- and four-head configurations under the fixed embedding dimension. Grad-CAM maps indicated that predictions were influenced mainly by the mouth, eyes, eyebrows, and mid-face. These findings show that refining compact convolutional features with MHSA can improve FER performance with a measured increase from 1.2 to 1.5 million parameters.

Keywords: facial emotion recognition, convolutional neural networks, multi-head self-attention, fine-grained expression recognition

1. Introduction

Facial emotion recognition (FER) infers emotional categories from facial images or video. Because a face is biometric data and emotion labels are interpretive, deploying FER can create privacy, consent, and reliability concerns [1]. The field of FER plays an important part in the study of affective computing since it attempts to make machines understand the emotion of people. It is applied in many areas, including human-computer interaction, driver monitoring systems, healthcare evaluation, and socially intelligent robots [2, 3]. Through the capability of making computer systems determine emotions through facial expressions, FER contributes to adaptive technologies. Although there are advances in deep learning technologies, there is still a challenge in performing powerful deep learning in real-world situations. This is mainly because of the nature of facial expressions that are usually faint and span different parts of the face. As a result, recognition requires capturing both local information and global information about those areas. Moreover, structural variance of the face, different postures of the face, lighting conditions, and intensities of emotion can all cause intra-class variance [4, 5]. Such issues get further exacerbated in unconstrained real-world data scenarios. The feedback signal is generally noisy, and the class distribution is uneven. Such aspects negatively influence the generalization capability of the model. CNNs are usually applied in FER applications owing to their proficiency in learning the hierarchical representation of the input data. CNN architecture is ideal for identifying local spatial features. However, CNNs are not very adept at capturing long-range spatial dependencies. This may impair the system's ability to recognize complex or composite facial expressions that rely on the interaction of features in several facial locations. [3, 6].

Attention models, especially multi-head self-attention (MHSA), have been introduced to improve the capability of global feature modeling. The MHSA model allows for the interaction between distant features, which makes it possible to model long-range relationships. Models based on transformers with attention mechanisms have shown excellent results in computer vision applications. However, their success relies on extensive pre-training and significant computing power, which limits their usage in FER problems with small datasets and time-constrained environments [7, 8]. Hybrid CNN-attention models try to combine the strengths of both CNN and attention models. Although

***Corresponding author:** Ayman Odeh, College of Engineering, Al Ain University, UAE. Email: ayman.odeh@aau.ac.ae

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

these approaches have demonstrated good results, there is no architectural justification for attention usage in many current approaches that use attention [9, 10]. As such, attention can be employed in an all-purpose manner rather than as a task-oriented process tailored for the features of facial expressions. Table 1 provides a comparative analysis of traditional CNN-based FER models against the proposed CNN–MHSA architecture.

To overcome these problems, we present our solution with the following list of contributions:

1) We propose a lightweight CNN-MHSA architecture, where attention acts as a refinement method for intermediate features without removing the convolutional network from its role.
2) For a more efficient way to model global dependencies, we propose the attention-as-refinement mechanism.
3) We also present our reproducible architecture design, where tokenization and attention are well-defined.
4) The proposed method is experimented on several benchmark data sets, and the results show that our architecture outperforms CNN baselines in terms of accuracy with less computation cost.

2. Literature Review

Early approaches to FER based on CNN frameworks often relied on handcrafted features, including facial landmarks, Gabor filters, and local binary patterns (LBP). While these techniques offered meaningful representations, they depended heavily on manual design and tended to perform poorly under variations in lighting, head pose, and other real-world conditions. The implementation of CNNs was a significant step because these kinds of neural networks are able to learn hierarchical features straight from the images. This allows for extracting subtle facial features, such as textural changes and micro-movements. More progress has been made by using deep neural networks such as ResNet, which benefits from residual learning to ensure smooth training and gradient flow [3, 6, 11]. Lightweight channel attention blocks, like the ECA-Net, enhance CNN features without adding much computational burden [12]. Nevertheless, they do not explicitly account for the long-range interaction between different parts of the face. Attention models allow a model to understand the relationship between spatially distant parts of an image. Specifically, multi-head self-attention (MHSA) has emerged as a popular method for capturing global context. Vision Transformers (ViTs) can capture global interactions well and perform competitively on various FER tasks. However, the complete backbones of ViTs require extensive pre-training and computation resources, and the lack of locality bias may be detrimental when dealing with limited and noisy FER data [7, 8]. Thus, our proposed architecture still employs a CNN backbone and applies MHSA solely to fine-tune compact intermediate features.

Hybrid CNN–attention architectures aim to combine the strengths of convolutional feature extraction with the global modeling capability of attention mechanisms. This combination has shown promising results, particularly in fine-grained visual recognition tasks. For example, Bukhari et al. proposed an efficient CNN–attention model for FER, while detection-based approaches such as YOLO have demonstrated strong performance in challenging real-world conditions, including occlusions and pose variations [9, 13, 14]. Efficient FER Models: Recent studies in facial emotion recognition (FER) have increasingly focused on computationally efficient architectures, model compression, and optimization techniques to meet the constraints of resource-limited environments. Although these approaches improve computational efficiency, their primary emphasis is often on reducing model size and computational cost rather than enhancing feature representation [15, 16]. Multimodal Emotion Recognition: The use of multimodal techniques in FER enables the inclusion of faces along with spoken/text/pulse data to improve accuracy in difficult situations [17, 18, 19]. While they often work better by using multiple data sources, they need extra sensors and complex synchronization, making them impractical for many real-world uses.

Research Gap and Motivation: As evident from the analysis above, there are a few gaps in the currently available FER techniques. Specifically, CNN models excel in extracting local

Table 1
Technical comparison between CNN-based FER and the proposed CNN–MHSA (attention-as-refinement) framework

Aspect	CNN-Based FER	Proposed CNN–MHSA (Attention-as-Refinement)
Feature Representation	Learning hierarchical local features through convolutional layers	Combining local feature extraction with global relational modeling
Receptive Field Modeling	Capturing global context indirectly through deep stacking	Modeling global context explicitly using attention mechanisms
Long-Range Dependencies	Learned implicitly through network depth	Modeled explicitly via attention-based interactions
Attention Mechanism	Not utilized	Applied as a targeted refinement stage
Spatial Interactions	Limited interaction between distant facial regions	Enhanced interaction across spatial regions via attention
Data Efficiency	Effective without large-scale pre-training	Maintained through attention on compact feature maps
Computational Complexity	Low computational cost	Slightly increased but remains computationally efficient
Scalability (Small Datasets)	Performs well on limited data	Maintains performance without reliance on large-scale pre-training
Robustness	Moderate robustness to variations	Improved robustness through global contextual modeling
Interpretability	Limited interpretability	Enhanced interpretability via attention and visualization methods
Fine-Grained Recognition	Limited ability to capture subtle expressions	Improved recognition of subtle and complex expressions

spatial information but cannot account for the long-range dependencies between facial areas that are needed for the recognition of fine and complex expressions [3, 6]. Transformer architecture can compensate for this weakness through the use of global attention mechanisms but suffers from increased computational complexity and requires extensive pre-training, which makes them less applicable to FER tasks with small and noisy datasets [8]. A hybrid CNN-attention technique appears to be an effective solution in this context; yet existing solutions often lack any theoretical justification regarding the design architecture, and particularly the use of attention mechanisms in FER tasks [9]. Compared with CNN-only FER baselines [20, 21], integrating multi-head self-attention (MHSA) can strengthen the modeling of long-range dependencies among facial regions, which is important for recognizing subtle or compound expressions. Self-attention and its adaptation to Vision Transformer architectures enable models to capture global relationships among spatial features [22], and subsequent work has shown that CNN–attention hybrids can outperform purely convolutional models on fine-grained visual recognition tasks by combining local feature extraction with global relational modeling [23].

Recent studies have explored complementary strategies for improving FER. Meena et al. developed a CNN-based approach and evaluated it on the CK+ and FER2013 datasets [24]. Tang et al. proposed a dual-stream attention network designed to improve FER robustness under occlusion and head-pose variations [25]. An and Jia combined self-supervised learning with fine-grained feature selection to reduce dependence on large, high-quality labeled datasets [26]. Yao combined multiscale convolutional features with an attention module to emphasize salient facial information and suppress irrelevant features [27].

For clarity purposes, Table 2 presents a systematic overview of previously suggested approaches and their main strengths and weaknesses.

The above observations further highlight the importance of having a good first and striking integration. The problem is related to CNN feature extraction and, at the same time, does not require any increase in computing costs or amount of data. Our CNN–MHSA architecture consists of using attention in order to fine-tune features extracted on intermediate layers.

3. Methodology

3.1. Overview

In this paper, we develop a lightweight hybrid CNN-MHSA network for facial emotion recognition (FER), which can capture both the local fine-grained features and the global spatial dependencies simultaneously. This model uses an attention-as-refinement approach, and it uses multi-head self-attention (MHSA) on intermediate representations obtained using a CNN backbone network. There are four steps in our proposed network, including (1) feature extraction by CNN, (2) tokenization and embedding, (3) attention-based refinement, and (4) classification.

3.2. Datasets and preprocessing

We evaluate the model on four established FER benchmarks: CK + [28], FER2013, RAF-DB, and AffectNet. Their availability and original dataset sources are provided in the Data Availability Statement. CK+ contains posed facial expression sequences captured in controlled laboratory conditions; our processed subset contains 981 labeled images. The experiments use seven output categories. For consistency with the in-the-wild datasets, CK+ contempt samples are mapped to the neutral output category. Table 3 reports the class counts used in each dataset. This gives us a comparative description of the class distribution in all four datasets for the study of FER. We notice here that the CK+ dataset has the class called "contempt," while the "neutral" class is present in FER2013, RAF-DB, and AffectNet datasets. Datasets vary widely in terms of size and class balance; FER2013 and AffectNet have a highly imbalanced class distribution, with "happy" and "neutral" being the dominant classes. Such differences bring new complexities to the problem of model generalization.

All images are converted to grayscale, resized to 48 × 47 pixels, and normalized to [0,1]. The emotion categories are assigned integer values between 0 and 6. In the case of the seven-category CK+ data used in our experiment, the "contempt" examples are assigned to the neutral category. The complete preprocessing pipeline is summarized in Table 4.

CK+ Dataset Note: Since the CK+ dataset contains tightly cropped faces, any face detection and alignment procedures need

Table 2
Research gap analysis and positioning of the proposed CNN–MHSA framework

Research Direction	Key Strength	Primary Limitation	Implication for FER	Proposed Solution
CNN-based FER [6, 11]	Powerful local features Extraction.	Global dependency modeling limitation.	Difficulty capturing subtle cross-region interactions	MHSA Introduction to Explicit Global Modeling.
Transformer-based FER [8]	Direct modeling of long-term dependency.	High computational cost and data dependency.	Limited applicability to small/noisy datasets	Light attention enhancement on built-in features.
Hybrid CNN–Attention [9]	Local and global features Combination.	Lack of clear attention placement strategy.	Suboptimal efficiency and interpretability	Structured design for attention as an improvement.
Optimization-driven FER [15, 16]	Efficiency Improving across tuning.	Focus on parameters rather than representation.	Limited improvement in feature modeling capacity	Improving the architectural level.
Multimodal FER [17, 18]	Robustness across multi-source data Enhancement.	Additional modalities and synchronization	Reduced practicality in real-world deployment	Strong single-mode frame for FER.

Table 3
Dataset class distribution by emotion category

#	Emotion Class	CK+	FER2013	RAF-DB	AffectNet
1	Angry	135	3,995	705	25,000
2	Disgust	177	436	717	3,800
3	Fear	75	4,097	281	6,400
4	Happy	207	7,215	4,771	134,000
5	Sad	84	4,830	1,982	25,000
6	Surprise	249	3,171	1,290	14,000
7	Neutral	—	4,965	2,528	75,000
8	Contempt	45	—	—	—
	Total	981	28,709	12,274	283,200

Table 4
Preprocessing pipeline

Step	Description
Grayscale Conversion	RGB images are converted to grayscale images in order to reduce computation and keep facial information intact.
Face detection and Alignment	Facial alignment is done on RAF-DB and AffectNet images to adjust any pose variations that might exist.
Image resizing	All images are resized to 48 × 48 pixel dimensions, similar to the FER2013 dataset.
Normalization	Pixel values are normalized to the [0,1] interval to ensure stable optimization.
Data Augmentation	The process of data augmentation is done through random flipping, random rotation (±10°), and altering the brightness (+/- 0.2) during the training process.
Label Encoding	The emotional labels have been encoded as integers between 0 and 6. The emotion of contempt in the CK+ database has been combined with neutral emotions.

to be performed uniformly on both train and test data sets and clearly stated. This would enable efficient training on benchmark data sets like FER2013 and RAF-DB.

3.2. Model architecture

The proposed model combines a CNN-based feature extractor with a lightweight MHSA refinement module, followed by a classification stage. The CNN backbone captures local facial details, including textures and muscle movements, while MHSA models relationships between spatially distant regions, such as the eye and mouth. The refined representation is passed to the classification head to produce probabilities for the seven emotion classes. The overall architecture is illustrated in Figure 1.

CNN-Based Feature Extraction: A lightweight CNN backbone is employed to extract hierarchical local features from facial images [29]. The architecture is composed of three convolutional blocks, each consisting of a $\mathbf{3 \times 3}$ convolutional layer followed by a ReLU activation function and a $\mathbf{2 \times 2}$ max-pooling operation. The number of filters increases progressively across the network (32, 64, and 128), resulting in a sequence of feature maps with increasing representational capacity.

$$(48 \times 48 \times 1) \rightarrow (24 \times 24 \times 32) \rightarrow (12 \times 12 \times 64) \rightarrow (6 \times 6 \times 128)$$

Where $F \in R^{\mathrm{H}xWxC}$ refers to the final convolutional feature map, H = 6 and W = 6 are its spatial dimensions, and C = 128 is the number of channels. The feature map encodes local facial structures such as edges, textures, and muscle activations.

Tokenization and Linear Embedding: To apply the attention mechanism, the feature map F is first reshaped into a sequence of tokens, following the approach used in Vision Transformers [30]. Specifically, the spatial dimensions of the feature map are flattened to form a set of token vectors, as defined in Eq. (1).

$$F \in R^{6\ x\ 6\ x\ 128} \rightarrow\ T_0\ \in R^{\ \mathrm{N}\ x\ \mathrm{C}}, \qquad N = 36,\ \ C = 128 \tag{1}$$

Each token corresponds to a spatial location in the feature map. Next, a linear projection is then applied to map each token into a lower-dimensional embedding space as shown in (2):

$$T = T_0 W_e, \quad W_e \in R^{128\ \mathrm{x\ d}} \tag{2}$$

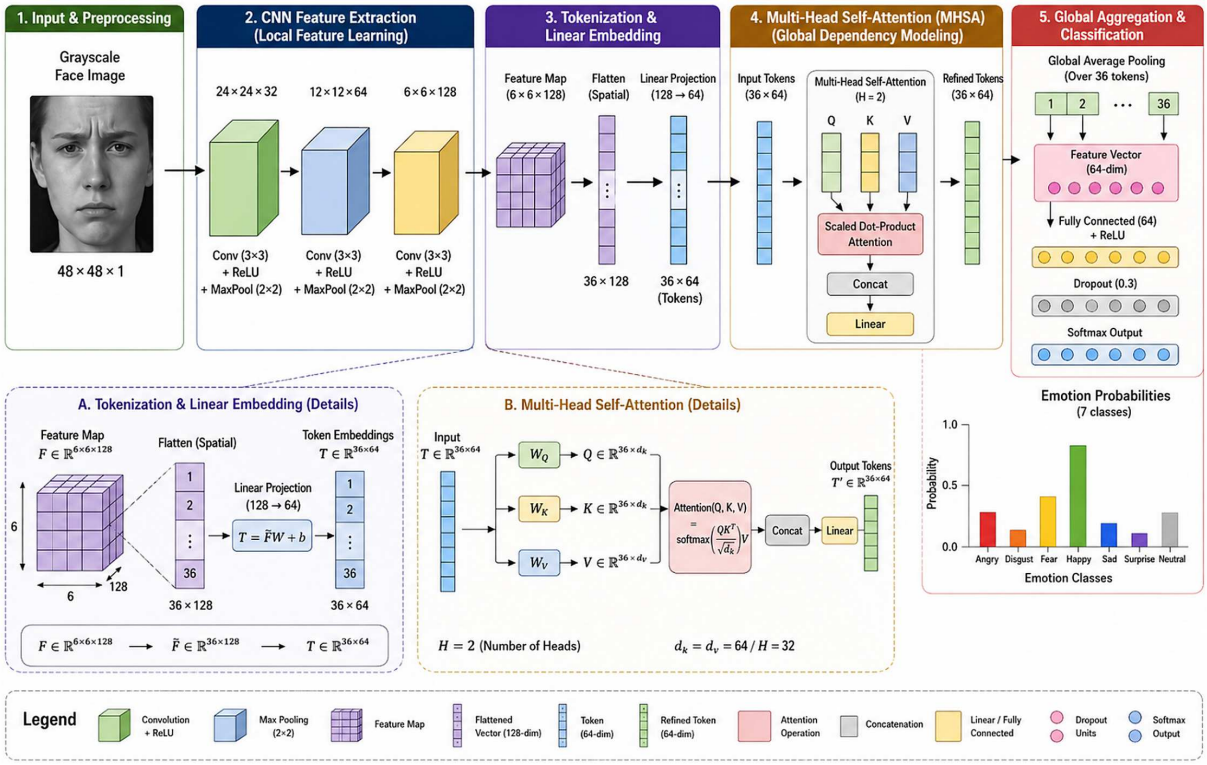
In Equations (1) and (2), N = H × W = 36 denotes the number of spatial tokens, C = 128 is the token input dimension, $T \in \mathbb{R}^{N \times C}$ is the flattened token sequence, $W_e \in \mathbb{R}^{c \times d}$ and $b_e \in \mathbb{R}^d$ are the learnable embedding parameters, $E \in \mathbb{R}^{N \times d}$ is the embedded token sequence, and d = 64 is the embedding dimension. This embedding process is consistent with standard transformer formulations [22].

Multi-Head Self-Attention (MHSA): The embedded token sequence $\boldsymbol{F} \in \boldsymbol{R}^{\mathbf{N} \times \mathbf{d}}$ is processed using a multi-head self-attention module to model global dependencies, following the standard multi-head self-attention formulation used in Vision Transformer architectures [22]. In this work, and for each head (h), the equations (3) are used to calculate query, key, and value matrices. The attention operation is defined in equation (4).

$$Q_h = TW_h^Q\ , \quad K_h = TW_h^K\ , \quad V_h = TW_h^V \tag{3}$$

Figure 1

The proposed framework's general architecture in which CNN captures the local features from the face image, the MHSA layer captures the dependency between the distant regions, and the classification layer generates probabilities for the emotions



$$Attention(Q_h,\ K_h, V_h) = softmax\left(\frac{Q_h K_h^T}{\sqrt{d_h}}\right) V_h \tag{4}$$

This formulation enables direct pairwise interactions between all tokens. This issue is facilitating explicit modeling of long-range spatial dependencies. The output of all H heads is concatenated and linearly projected as displayed in (5); in this work, H = 2 and d = 64 provide a balance between representational capacity and computational efficiency.

$$T' = Conact(head_1,\ \ldots\ldots,\ head_H)W_0 \tag{5}$$

Global Feature Aggregation: To obtain a compact global representation, the global average pooling is applied across the token dimension as demonstrated in (6). This calculation is widely used for spatial aggregation in deep networks due to its efficiency and regularization properties.

$$z = \frac{1}{N}\sum_{i=1}^{N} T'_i \tag{6}$$

Global average pooling Classification Head: The aggregated feature representation $\boldsymbol{z} \in \boldsymbol{R^d}$, which encodes both local and global facial information, is mapped to emotion class probabilities through a fully connected classification head. The final prediction is calculated using (7), and this component consists of:

1) a dense layer with 64 units and ReLU activation,
2) a dropout layer with rate 0.3 for regularization, and
3) a softmax output layer:

$$\hat{y} = softmax(W_c z + b_c) \tag{7}$$

where $\hat{\boldsymbol{y}}$ stands for the predicted distribution of probabilities to the categories of emotion, and Wc and bc are learnable parameters. The classification head transforms feature representations into probability distributions using a linear transform and softmax activation function for class separability. A dropout layer with a rate of 0.3 is applied to the classification head as a regularization measure to reduce overfitting [31].

3.3. Training configuration

Adam (learning rate = 0.001) and sparse categorical cross-entropy loss are used to train the model, the latter being the typical one used for multi-class FER where class labels are integers. The batch size is 32, up to 25 epochs with early stopping at 5 patience epochs. The main hyperparameters are presented in Table 5.

In the evaluation step, we used two protocols: (1) Protocol A: Image-level 80/20 split (supplementary). (2) Protocol B: Subject-independent 5-fold cross-validation (primary). In Protocol B, subjects are partitioned such that no identity appears in both training and testing sets. This protocol provides a more reliable assessment of generalization performance.

3.4. Computational complexity

The computational complexity of the MHSA module is calculated using $O(N^2d)$, where N = 36 is the number of tokens and d = 64 is the embedding dimension [22]. Because attention is applied to compact CNN feature maps rather than to image patches at the input resolution, its computational overhead remains limited relative to a full transformer backbone.

Table 5
Key Hyperparameters

Parameter	Setting	Notes
Input size	48 × 48 × 1	Grayscale FER format
Optimizer	Adam	Learning rate = 0.001
Loss	Sparse CE	Handles 0-6 emotion labels
Dropout	0.3	Classifier regularization
Batch size	32	Balances speed and gradient stability.
MHSA	2 heads, d = 64	Lightweight refinement
Augmentation	Flips, ±10° rot, brightness	Prevents overfitting on imbalanced data

4. Results

4.1. Evaluation overview

To measure the generalization capability and robustness, two evaluation protocols have been adopted. Firstly, subject-independent 5-fold cross validation (Protocol B) is performed on CK + to validate the model's capability in handling unseen identities. Secondly, independent training and testing of the model are performed on FER2013, RAF-DB, and AffectNet datasets using their respective splits. This is multi-dataset evaluation; it is not a train-on-one-dataset/test-on-another cross-dataset transfer protocol. Performance is measured using accuracy, precision, recall, and macro F1-score, supplemented by ablation experiments and Grad-CAM visualization.

4.2. Subject-independent performance on CK +

We primarily tested using subject-independent 5-fold cross-validation on CK+. As shown in Table 6, our model reached 97.76% mean accuracy and 0.9694 macro F1, staying very consistent across all folds.

Table 6
Fold-wise performance under Protocol B (subject-independent 5-fold cross-validation)

Fold	Accuracy	Precision	Recall	F1 Score
Fold 1	0.9695	0.9496	0.9662	0.9511
Fold 2	0.9592	0.9663	0.9392	0.9473
Fold 3	0.9898	0.9858	0.9858	0.9858
Fold 4	0.9847	0.9821	0.9781	0.9790
Fold 5	0.9847	0.9845	0.9841	0.9837
Mean	0.9776	0.9737	0.9707	0.9694
Std	0.0128	0.0156	0.0192	0.0186

Performance remains consistent across the five folds (accuracy standard deviation = 0.0128), indicating stable generalization to unseen identities. The results support the use of MHSA for modeling relationships among spatially separated expression cues, although they do not constitute cross-dataset transfer evidence.

4.3. Evaluation across multiple FER datasets

The model is trained and evaluated independently using the FER2013, RAF-DB, and AffectNet datasets, as shown in Table 7 and Figure 2. Compared with its matched baseline CNN, the model CNN–MHSA yields an improvement of 3.20 percent on FER2013, 3.50 percent on RAF-DB, and 3.90 percent on AffectNet. FER2013 and AffectNet have extensive variations in terms of illumination, pose, expression intensity, labeling accuracy, and class distribution. The consistent outperformance over the matched baselines can be taken as proof of the effectiveness of attention-based refinement for learning context-dependent features under such variations. The poor performance of AffectNet further verifies the importance of the dataset size and imbalance issue.

Such improvement shows that MHSA helps to improve the ability of the model in capturing the dependencies between distant facial regions.

4.4. Comparison with state-of-the-art methods

Table 8 provides a reference-indexed comparison with published methods that report results on matching datasets. ResEmoteNet [32] reports 79.79% on FER2013, 94.76% on RAF-DB, and 72.93% on AffectNet-7, whereas HLA-ViT [8] reports 90.45% on RAF-DB and 65.07% on AffectNet. Because preprocessing, class definitions, and training protocols may differ across studies, these literature values are provided for context rather than as a controlled head-to-head comparison.

The proposed lightweight CNN–MHSA model achieves 72.40% on FER2013, 87.60% on RAF-DB, and 61.20% on AffectNet-7. Although the cited transformer-based and ResEmoteNet results are higher on several datasets, they use different architectures and evaluation pipelines. Accordingly, Table 8 is presented as a reference-indexed literature comparison rather than a controlled claim of superiority. The motivation for the proposed CNN–MHSA design is to retain convolutional locality and apply attention only to compact intermediate features.

4.5. Ablation Study and Computational Efficiency

Ablation: The results of the ablation study are presented in Table 9 and Figure 3. The use of MHSA boosts performance compared to the CNN-only model for CK+, FER2013, and RAF-DB datasets. Among the attention-only architectures, the two heads yield the best results on all three datasets.

1) As compared to the CNN-only architecture, the two-head CNN-MHSA model increases accuracy by 1.86%, 3.20%, and 3.50% on CK+, FER2013, and RAF-DB, respectively.
2) Going from one to two heads increases the accuracy, while four heads do not bring any further improvement at the constant embedding size. Such an outcome confirms the decision to use two heads in the attention-as-refinement model.

Table 7
Comparison of performances on different FER datasets

Dataset	Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CK+	CNN (baseline)	95.90	95.20	95.00	94.80
CK+	Proposed	97.76	97.37	97.07	96.94
FER2013	CNN (baseline)	69.20	68.50	67.80	68.10
FER2013	Proposed	72.40	71.80	71.10	71.30
RAF-DB	CNN (baseline)	84.10	83.50	83.00	83.20
RAF-DB	Proposed	87.60	87.10	86.70	86.80
AffectNet	CNN (baseline)	57.30	56.40	55.90	56.10
AffectNet	Proposed	61.20	60.50	59.80	60.00

Figure 2
Comparison of performances between CNN baseline model and the proposed CNN-MHSA model on different datasets.

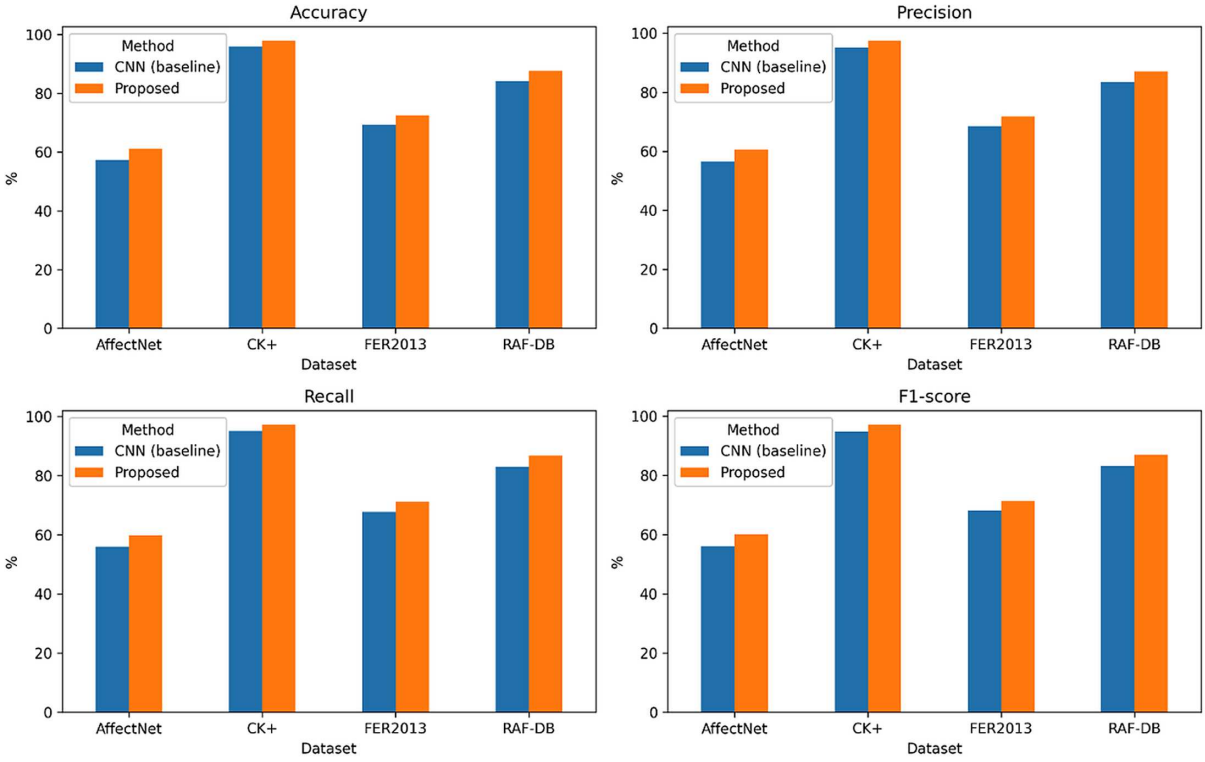


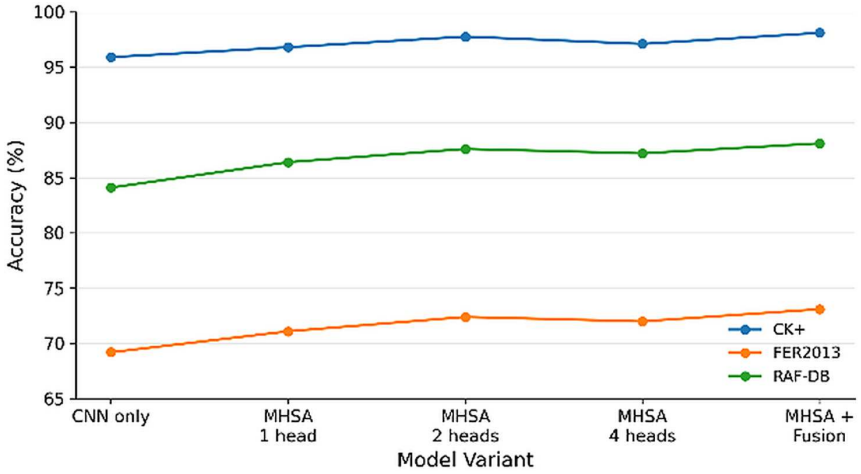
Table 8
Reference-indexed comparison with published FER methods

Dataset	Method	Year	Accuracy (%)
CK+	Ours (CNN-MHSA)	2026	97.76
FER2013	ResEmoteNet [32]	2025	79.79
FER2013	Ours (CNN-MHSA)	2026	72.40
RAF-DB	HLA-ViT [8]	2024	90.45
RAF-DB	ResEmoteNet [32]	2025	94.76
RAF-DB	Ours (CNN-MHSA)	2026	87.60
AffectNet-7	HLA-ViT [8]	2024	65.07
AffectNet-7	ResEmoteNet [32]	2025	72.93
AffectNet-7	Ours (CNN-MHSA)	2026	61.20

Table 9
Ablation study of proposed architecture

Model Variant	CK+ Acc (%)	FER2013 Acc (%)	RAF-DB Acc (%)
CNN only	95.90	69.20	84.10
CNN + MHSA (1 head)	96.80	71.10	86.40
CNN + MHSA (2 heads, proposed)	97.76	72.40	87.60
CNN + MHSA (4 heads)	97.10	72.00	87.20
CNN + MHSA + Feature Fusion	98.10	73.10	88.10

Figure 3
Ablation study of the proposed model



3) The feature-fusion model yields the highest accuracy in Table 9 (98.10% for CK+, 73.10% for FER2013, and 88.10% for RAF-DB). However, the two-head model was chosen as the one suggested by the model since it directly incorporates the idea of attention-as-refinement without any extra pathway. Since the parameters and the time latency for each variant were not evaluated, we cannot prove that this decision is the most efficient one empirically; only the accuracy-based upper variant is presented.

Computational Efficiency: Table 10 shows the results of computational performance measurement done in this study for the baseline CNN model and the CNN_MHSA architecture. Adding MHSA increases the parameter count from 1.2 M to 1.5 M, the estimated computation from 0.45 to 0.60 GFLOPs, and the measured inference time from 4.8 to 6.2 ms under the same experimental environment.

Table 10
Computational efficiency comparison measured in this study

Model	Parameters (M)	FLOPs (G)	Inference Time (ms)
CNN baseline	1.2	0.45	4.8
CNN + MHSA	1.5	0.60	6.2

The above findings suggest there is reasonable computational overhead compared to the CNN counterpart of the same architecture. The comparison with existing ViTs is done in a qualitative manner since the computation time and FLOP values are obtained using different system settings.

4.6. Confusion matrix

Confusion matrices are shown in Figure 4 and Figure 5 for both training and testing sets. While the training confusion matrix is perfectly diagonal, there are just a few off-diagonal errors in the testing confusion matrix. The confusion lies mainly in the recognition of minority or finer classes, such as sadness against anger and surprise vs contempt.

Figure 4
Training confusion matrix (Protocol A)

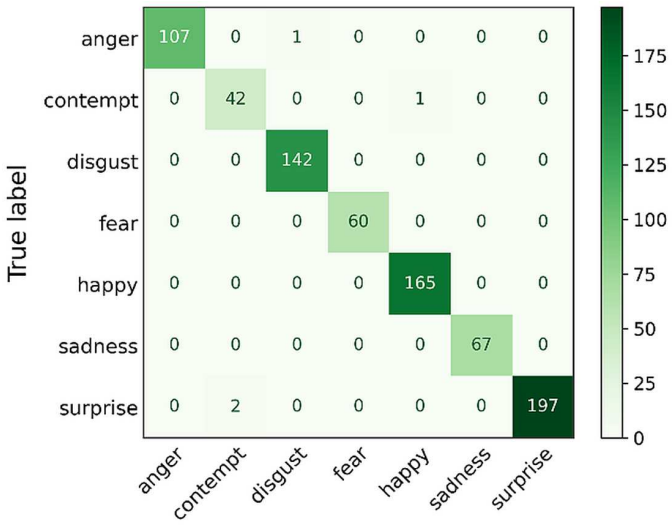
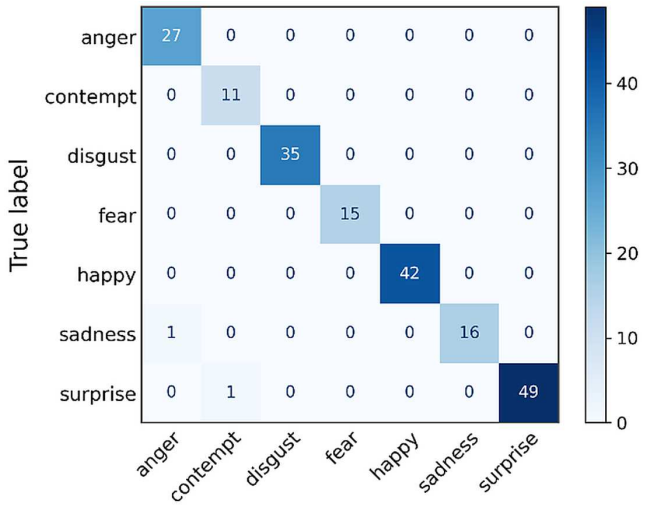


Figure 5
Test confusion matrix (Protocol A)



1) Sadness ↔ Anger: Confusion at low intensity.
2) Surprise ↔ Contempt: Class overlapping is rare.
3) Fear ↔ Surprise: Same eye opening.

4.7. Training dynamic curves

Figure 6 and Figure 7 illustrate loss and accuracy throughout the number of epochs. The curves of training and validation match closely, implying stability during optimization in the case of Protocol A. Figure 6 illustrates smooth and clean convergence—training and validation curves match closely, and there is no explosion or divergence. In the beginning, curves jump rapidly and then stabilize starting from epoch 12-15.

1) Training and validation curves follow each other and converge with losses of about 0.18 and 0.22, respectively. The accuracy reaches nearly 97% value, and early stopping happens at epoch 18.
2) Loss curves are close to each other (validation loss ~0.22 vs train ~0.18)
3) Accuracy converges at 97%
4) Early stopping is initiated at epoch 18 (patience=5)

The closeness of these two curves implies stability during optimization in the case of Protocol A. It can be supposed that dropout and data augmentation can have the role of regularization. This point is true only for this split since overfitting cannot be ruled out in other settings.

These mistakes reflect the class imbalance and the visual similarities of the low-intensity expressions. These should be treated as deficiencies in the model and not just as features of the data. The confusion matrices suggest excellent discrimination between the classes but poor discrimination for the minority classes and those with similar facial expressions.

4.8. Grad-CAM interpretation

In order to understand our predictions, we construct Grad-CAM overlays for various correctly classified test samples (see Figure 8). In general, the most salient activations are found in the regions that have anatomical meaning—eyes/eyebrows and mouth—and not in background pixels. For happiness, the activation is more intense around the mouth and cheeks; for surprise, the activation map focuses more on the open mouth and eye region; for disgust, the activation centers around the mid-face (nose/upper lip); and for fear, the map emphasizes the eye and brow regions. Our interpretation of these maps is qualitative evidence of influential regions and not a definite causation mechanism.

Figure 6
Training and validation loss curves of the proposed model

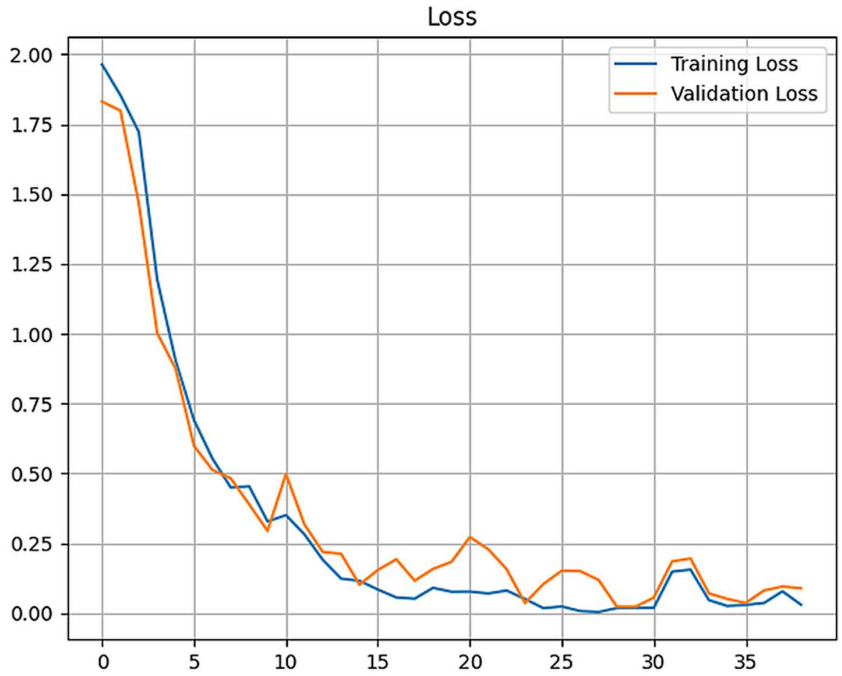


Figure 7
Training and validation accuracy curves of the proposed Model

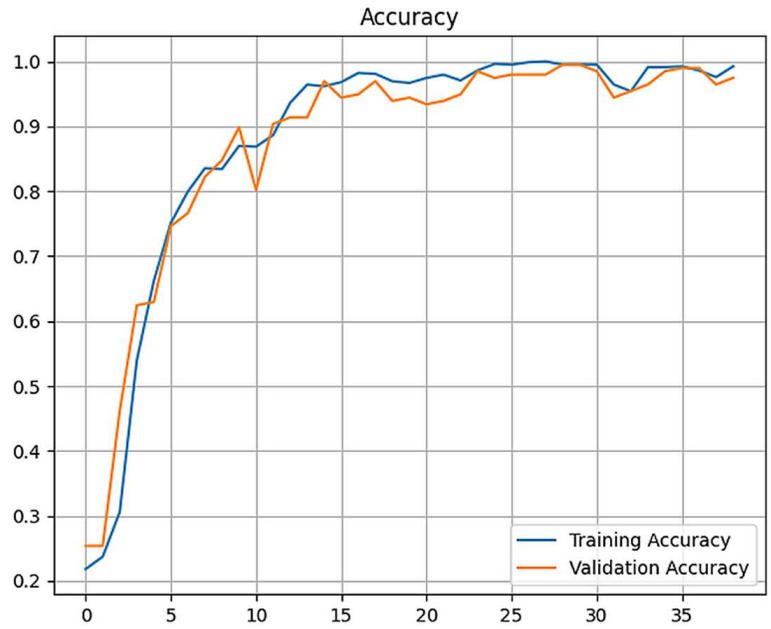
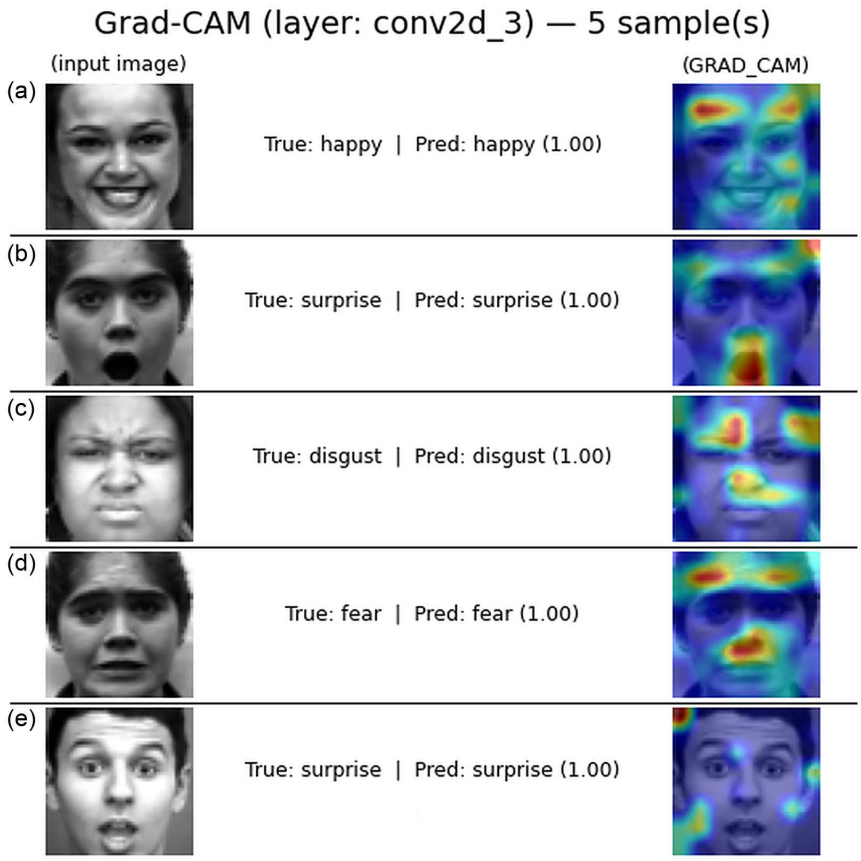


Figure 8
Grad-CAM visualizations for five correctly classified samples (example layer: conv2d_3)



The heat maps activate eyes, eyebrows, mouth, cheeks, and mid-face and not background regions. This observation suggests that the model activates in relation to relevant anatomical regions.

1) Emotion-specific activation:
2) Happiness → mouth, cheeks (smile)
3) Surprise → mouth wide open, eye region
4) Disgust → nose, upper lip (wrinkles)
5) Fear → eye, raised brows

The activation patterns are generally coherent with the facial regions related to expressions. Since Grad-CAM cannot establish causation, these maps are considered only as qualitative evidence.

4.9. Discussion

The results show that using MHSA as a refining phase enhances the performance of the baseline CNN with matching capabilities on the studied datasets. The reason for this is due to the nature of the self-attention process, which models the relationship between different spatially distributed parts of the face. Long-range relations are captured implicitly through the stacked receptive field in the convolution layers, while MHSA explicitly captures the relationship between tokens.

The advancements made are particularly clear in cases where it is used for challenging databases such as FER2013 and AffectNet, which have high variance concerning pose, illumination, and annotation. This indicates that incorporating attention in the model makes the network more robust to variance through context-based reasoning. In this aspect, our findings are similar to previous studies, which have highlighted that the use of attention mechanisms aids in handling long-range dependency problems [5, 10].

As demonstrated by the ablation study, it is better to have two heads in the examined attention-only architectures, whereas the best result in terms of accuracy is achieved with the help of the feature-fusion architecture. As for the proposed two-head architecture, it is kept for simplicity; however, further measurements are required to evaluate its complexity. Another important issue related to AffectNet is label noise.

As far as computational issues are concerned, the architecture proposed in this work represents a decent balance between speed and efficiency. By means of applying MHSA to compressed feature maps rather than directly applying it to the image, the issue of quadratic complexity in transformer-based models is solved. Such a characteristic of architecture makes it particularly appropriate for the cases where efficient performance and real-time operation are needed, such as, for example, in the case of embedded devices, smartphones, or driver monitoring systems. However, the current version of the model is limited to static facial expression recognition (FER) from images and does not have any temporal aspect that would help in the process of emotion recognition from videos.

5. Conclusion

Therefore, the current framework was based on designing a hybrid lightweight CNN-MHSA framework for facial emotion recognition through multiple datasets. Our current framework can integrate both local feature extraction and explicit global dependency modeling. Unlike transformer models, where attention is used as the entire framework, our current approach uses multi-head self-attention as a refining step in the intermediate convolutional features.

This architecture preserves the efficiency and inductive bias of the convolutional networks. At the same time, it provides an opportunity for the model to learn connections among different facial regions that are crucial for expression recognition. All the experiments performed on several benchmark datasets demonstrate good results from the model. The experiment on CK+ under the conditions of a subject-independent scenario proves the good ability of generalization on unseen subjects, while the experiments on FER2013, RAF-DB, and AffectNet demonstrate good robustness in case of noise and class imbalance. Moreover, the results of ablation prove the effectiveness of the attention-as-refinement strategy.

One of the major advantages of this paper is the proof of efficiency of the global context reasoning for FER. With the use of the attention on compact intermediate features, the proposed technique offers an efficient and accurate approach, which makes it suitable for real-time applications. Still, despite the advantages offered by the approach, there are certain limitations as well. Firstly, the proposed approach is limited to static images only and therefore does not consider any temporal data, which can be helpful in performing emotional reasoning. Secondly, the proposed approach is affected by label noise and class imbalance in large datasets.

Future directions in research would seek to expand the presented framework in order to include temporal modeling for FER based on sequences and to devise approaches to better deal with noisy and imbalanced data. Also, further experimentation on datasets that are diverse and population-balanced will be necessary in order to gauge fairness and generalization performance.

Overall, it can be stated that the introduced CNN-MHSA architecture is an efficient approach to the FER problem, considering the major issues arising from the nature of facial emotions.

Ethical Statement

This study did not involve a questionnaire, participant recruitment, direct interaction with human or animal participants, or the collection of new personal data. It involved secondary computational analysis of the publicly available CK+, FER2013, RAF-DB, and AffectNet facial-expression benchmark datasets. The datasets were accessed and used in accordance with their respective research-use conditions. No separate ethical approval or informed consent was obtained by the authors for this secondary analysis because no new human-participant research was conducted. Responsibility for the original data collection, including any required ethical approvals and participant consent, rests with the respective dataset creators.

Conflicts of Interest

The authors declare that they have no conflicts of interest in this work.

Data Availability Statement

The data that support the findings of this study are openly available in the CK+ dataset at https://www.jeffcohn.net/resources/, reference number [28]; in Kaggle at https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data; in the RAF-DB dataset at http://www.whdeng.cn/RAF/model1.html; and in the AffectNet dataset at https://mohammadmahoor.com/pages/databases/affectnet/.

Author Contribution Statement

Ayman Odeh: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Supervision, Project administration. **Munther Odeh:** Methodology, Software, Investigation, Data curation, Writing – original draft. **Nada Odeh:** Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization.

References

[1] Mattioli, M., & Cabitza, F. (2024). Not in my face: Challenges and ethical considerations in automatic face emotion recognition technology. *Machine Learning and Knowledge Extraction*, *6*(4), 2201–2231. https://doi.org/10.3390/make6040109

[2] Sambasivarao, K. V., & Bhima, A. S. R. D (Eds.). (2026). *Artificial intelligence, computational intelligence and inclusive technologies*. London:CRC Press. https://doi.org/10.1201/9781003740100

[3] Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, *13*(3), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446

[4] Xiong, K., Qing, L., Li, L., Guo, L., & Peng, Y. (2025). Facial expression recognition based on local–global information reasoning and spatial distribution of landmark features. *The Visual Computer*, *41*(1), 535–548. https://doi.org/10.1007/s00371-024-03345-y

[5] Zhang, Y., Zheng, X., Liang, C., Hu, J., & Deng, W. (2024). Generalizable facial expression recognition. In *European Conference on Computer Vision*, 231–248. https://doi.org/10.1007/978-3-031-72630-9_14

[6] Sălăgean, G. L., Leba, M., & Ionica, A. C. (2025). Leveraging symmetry and addressing asymmetry challenges for improved convolutional neural network-based facial emotion recognition. *Symmetry*, *17*(3), 397. https://doi.org/10.3390/sym17030397

[7] Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., . . . , & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, *8*(3), 331–368. https://doi.org/10.1007/s41095-022-0271-y

[8] Tian, Y., Zhu, J., Yao, H., & Chen, D. (2024). Facial expression recognition based on vision transformer with hybrid local attention. *Applied Sciences*, *14*(15), 6471. https://doi.org/10.3390/app14156471

[9] Bukhari, S. M. S., Zafar, M. H., Moosavi, S. K. R., & Sanfilippo, F. (2025). Emotion recognition with a randomized CNN-multihead-attention hybrid model optimized by evolutionary intelligence algorithm. *Array*, *26*, 100401. https://doi.org/10.1016/j.array.2025.100401

[10] Farzan, M., Rezaei, M., Mottaghi, S., & Mirsaleh, Y. R. (2025). Facial emotion recognition in offspring of parents with bipolar disorder. *Middle East Current Psychiatry*, *32*(1), 30. https://doi.org/10.1186/s43045-025-00523-z

[11] Talele, M., & Jain, R. (2025). A comparative analysis of CNNs and ResNet50 for facial emotion recognition. *Engineering, Technology & Applied Science Research*, *15*(2), 20693–20701. https://doi.org/10.48084/etasr.9849

[12] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155

[13] Utanğaç, M., & Lander, K. (2026). Facial emotion recognition, empathy & psychopathic traits: Understanding differential effects of measures. *BMC Psychology*, *14*(1), 272. https://doi.org/10.1186/s40359-026-04084-6

[14] Alshammari, A., & Alshammari, M. E. (2024). Emotional facial expression detection using YOLOv8. *Engineering, Technology & Applied Science Research*, *14*(5), 16619–16623. https://doi.org/10.48084/etasr.8433

[15] Joseph, C., & Maheswari, P. U. (2025). Facial emotion based smartphone addiction detection and prevention using deep learning and video based learning. *Scientific Reports*, *15*(1), 18025. https://doi.org/10.1038/s41598-025-99681-7

[16] Ajlouni, N., Özyavaş, A., Ajlouni, F., Takaoğlu, F., & Takaoğlu, M. (2025). Enhanced hybrid facial emotion detection & classification. *Franklin Open*, *10*, 100200. https://doi.org/10.1016/j.fraope.2024.100200

[17] Sanku, S. R., & Sandhya, B. (2024). Multi-modal emotion recognition feature extraction and data fusion methods evaluation. *International Journal of Innovative Technology and Exploring Engineering*, *13*(10), 18–27. https://doi.org/10.35940/ijitee.J9968.13100924

[18] Jinnuo, Z., Goyal, S. B., Rajawat, A. S., Nassar Waked, H., Ahmad, S., Randhawa, P., . . . , & Naik, N. (2025). Analysis of existing techniques in human emotion and behavioral analysis using deep learning and machine learning models. *Engineering Research Express*, *7*(1), 012201. https://doi.org/10.1088/2631-8695/ada68b

[19] Kocaleva, M., Stojanova, A., Koceska, N., & Koceski, S. (2025). Emotion detection from physiological markers using machine learning. *Informatica*, *49*(21), 29–44. https://doi.org/10.31449/inf.v49i21.7442

[20] Dixit, I., Prakash, C., Ramya, G., Dinesh, M., Giri, J., & Amer, A. (2024). Human emotion detection using deep CNN. In *2024 8th International Conference on Electronics, Communication and Aerospace Technology*, 1391–1395. https://doi.org/10.1109/ICECA63461.2024.10800985

[21] Goud, K. M., & Hussain, S. J. (2021). Estimation of emotion using CNN. In *2021 Second International Conference on Electronics and Sustainable Communication Systems*. https://doi.org/10.1109/ICESC51422.2021.9532763

[22] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., . . . , & Tao, D. (2023). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 87–110. https://doi.org/10.1109/TPAMI.2022.3152247

[23] Zhao, Y., Chen, J., Zhang, Z., & Zhang, R. (2022). BA-Net: Bridge attention for deep convolutional neural networks. In *European Conference on Computer Vision*, 297–312. https://doi.org/10.1007/978-3-031-19803-8_18

[24] Meena, G., Mohbey, K. K., Indian, A., Khan, M. Z., & Kumar, S. (2024). Identifying emotions from facial expressions using a deep convolutional neural network-based approach. *Multimedia Tools and Applications*, *83*(6), 15711–15732. https://doi.org/10.1007/s11042-023-16174-3

[25] Tang, H., Li, Y., & Jin, Z. (2024). A dual stream attention network for facial expression recognition in the wild. *International Journal of Machine Learning and Cybernetics*, *15*(12), 5863–5880. https://doi.org/10.1007/s13042-024-02287-0

[26] An, H. Y., & Jia, R. S. (2024). Self-supervised facial expression recognition with fine-grained feature selection. *The*

Visual Computer, *40*(10), 7001–7013. https://doi.org/10.1007/s00371-024-03322-5

[27] Yao, L. (2024). Facial expression recognition based on multiscale features and attention mechanism. *Automatic Control and Computer Sciences*, *58*(4), 429–440. https://doi.org/10.3103/S0146411624700548

[28] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

[29] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

[30] Xue, F., Wang, Q., Tan, Z., Ma, Z., & Guo, G. (2022). Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Transactions on Affective Computing*, *14*(4), 3244–3256. https://doi.org/10.1109/TAFFC.2022.3226473

[31] Salehin, I., & Kang, D. K. (2023). A review on dropout regularization approaches for deep neural networks within the scholarly domain. *Electronics*, *12*(14), 3106. https://doi.org/10.3390/electronics12143106

[32] Roy, A. K., Kathania, H. K., Sharma, A., Dey, A., & Ansari, M. S. A. (2025). ResEmoteNet: Bridging accuracy and loss reduction in facial emotion recognition. *IEEE Signal Processing Letters*, *32*, 491–495. https://doi.org/10.1109/LSP.2024.3521321

How to Cite: Odeh, A., Odeh, N., & Oder, M. (2026). A Lightweight Hybrid CNN–Multi-Head Attention Framework for Multi-Dataset Facial Emotion Recognition. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62029975