STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, December 2026, pp 5271–5289.
Published online in International Academic Press (www.IAPress.org)

XAI for Facial Emotion Recognition: Quantitative Analysis via Grad-CAM and Integrated Gradients

Zahraa Mazin Alkattan*

Department of Software Engineering, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

Abstract Automated facial emotion recognition (FER) underlies affective computing, mental-health assessment, and human-computer interaction, yet state-of-the-art deep learning models remain black boxes that give no account of which facial cues drive their decisions, constraining their clinical and educational use. We present an explainable FER framework built on a fine-tuned ResNet-50 and two complementary post-hoc attribution methods, Gradient-weighted Class Activation Mapping (Grad-CAM) and Integrated Gradients (IG), that together provide high predictive performance and explanations suited to non-expert interpretation. The model was trained and tested on a subject-independent protocol on the Extended Cohn-Kanade (CK+) dataset [18], comprising 22,400 offline-augmented images from 123 subjects across eight emotion classes. The model was optimized using Adam and StepLR and achieved test accuracy of 99.955% (2,239/2,240 correct) and macro F1 of 0.9996 on the full test set. As part of this revision we performed a pixel-level duplicate audit of the released data splits and found that 319 test images (14.2%) are byte-identical to a training image, most plausibly reflecting near-duplicate adjacent peak frames from the same recording assigned to different splits during offline augmentation; recomputing the headline metrics after excluding every exact duplicate gives 99.948% accuracy and macro F1 = 0.9995 on the remaining 1,921 images, statistically indistinguishable from the original figures (two-proportion $z = 0.10$, $p = 0.92$). Grad-CAM visualization showed spatial attention qualitatively consistent with action units in the Facial Action Coding System (FACS), and IG attributions enabled a quantitative Facial Region Importance Matrix across seven anatomical regions and eight emotion categories. We further introduce, for both attribution methods, objective Insertion/Deletion AUC faithfulness metrics, a model confidence-calibration analysis (Expected Calibration Error), and Wilson-interval significance testing of a five-component ablation study. Each component's contribution is confirmed by the ablation study. The complete implementation is released in a single open-source Google Colab notebook.

Keywords Facial Emotion Recognition, Explainable AI, ResNet-50, Grad-CAM, Integrated Gradients, Deep Learning, CK+ Dataset, FACS, Attribution Methods

AMS 2010 subject classifications 68T07, 68T10

DOI: 10.19139/soic-2310-5070-4212

1. INTRODUCTION

The biggest coin in the currency of human communication is emotion: It influences the determination of decisions, governs the behavior in society and expresses the psychological states, which could not be observed otherwise. Facial emotion expression recognition has generated enduring research attention in a wide range of purposes - affective computing, human-robot interaction, human mental health assessment, driver monitoring systems, pain detection in non-verbal patients and intelligent tutoring systems [1], [2]. The possibility to consistently determine the affective state of an individual based on facial appearance is an opportunity of more responsive-personalized- and ultimately safe systems in each of these areas.

*Correspondence to: Zahraa Mazin Alkattan (Email: zahraa.alkattan@uomsul.edu.iq). Department of Software Engineering, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 2026 International Academic Press

Convolutional neural networks (and more recently Vision Transformers) have changed the performance of FER systems, with current-day systems nearly matching human performance on controlled tasks [3], [4]. In spite of this, there is one essential issue that has not been completely addressed: practically all high-performing FER models are black boxes and give predictions without telling how much of their decision was influenced by an area of the face or pixel-level features. When it comes to high-stakes deployment worlds, this opacity is not only inconvenient, but it is also a hindrance to adoption. A model whose decision-making process is completely invisible cannot be relied upon by a clinician who wants to evaluate a patient with his depression or post-traumatic stress disorder [8]. The teacher who has access to an intelligent tutoring system to determine how much students are engaged should have the ability to audit the inferences of the system. Areas of AI regulation within healthcare such as the EU AI Act and FDA guidance on AI based medical devices are gradually establishing transparency of algorithms as a pre-requisite to their deployment.

The research paradigm that has responded to this requirement is Explainable Artificial Intelligence (XAI). Among the most established XAI techniques is Grad-CAM [15], an attribution-based method that generates class-discriminative spatial heatmaps by computing gradient-weighted activation maps of a target convolutional layer. Integrated Gradients (IG) [16] make the most of this by calculating pixel-level attribution by integrating gradients, and these have axiomatic properties such as Completeness and Sensitivity. Perturbation-based analysis in super pixels and game-theoretic decomposition of the Shapley value is also presented in LIME [17] and SHAP [26] respectively. GradCAM++ [27] and ScoreCAM [28] are extensions of Grad-CAM, but include better gradient weighting and activation-score-based maps.

Although this is a well-developed XAI toolkit, the systematic application to FER is surprisingly small. Canal et al. [8] reviewed FER methods and used Grad-CAM on FER2013, showing periorbital dominance on most expressions, but not providing quantitative regional assessment, not using a subject-independent validation. Antoniadis et al. [10] tested LIME on an audio-video arousal recognition task with facial information added to it, not providing the claims of attribution-completeness or cross-class quantification. Wang et al. [13] used SHAP on Vision Transformer on CK+ with 99.40% accuracy but restricted the analysis of explainability to global feature value that was not sliced into the anatomy and not verified on FACS. The current paper represents a unique manner of integrating Grad-CAM and IG within one framework and a quantitative Facial Region Importance Matrix perturbation - the gap that is filled here.

This paper presents an explainable FER model that reaches state of the art classification performance and offers the most detailed explainability analysis that has ever been reported on CK+. These major contributions are:

- An optimized ResNet-50 achieving 99.955% test accuracy and macro F1 of 0.9996 on CK+ under a subject-independent protocol, together with a pixel-level duplicate audit of the released data splits and a corrected, deduplicated headline result (99.948% accuracy, macro F1 = 0.9995) confirming that the result is not an artifact of image-level leakage.
- A dual explainability framework combining Grad-CAM and Integrated Gradients, including the first quantitative Insertion/Deletion AUC faithfulness comparison between the two methods on this task, contrasting correct and incorrect attributions across all eight emotion classes.
- A new quantitative Facial Region Importance Matrix based on Integrated Gradients, composing attribution profiles of classes resolved on anatomical regions of seven FACS-grounded regions.
- A systematic ablation study with 95% Wilson confidence intervals and significance testing for each training component, and error analysis of the one observed misclassification, together with a full confidence-calibration analysis (Expected Calibration Error) of the resulting model.
- An open-source implementation in a single Google Colab notebook with permanent Google Drive storage.

The rest of the paper is structured in the following way. Section 2 reviews related work; Section 3 outlines the methodology; Section 4 presents results and discussion; Section 5 concludes the paper; and Section 6 addresses ethical considerations.

2. RELATED WORK

2.1. Facial Emotion Recognition of Deep Learning.

First generation FER methods were based on handcrafted characteristics, Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Active Appearance Models (AAM) applied with Support Vector Machines. The invention of CNNs changed this situation radically, allowing the hierarchical representation of faces to be learned end-to-end using raw pixel intensities.

Farzaneh and Qi [1] introduced facial action unit-directed CNN where they add the concept of the anatomically grounded supervision to the loss function and attain the results of 97.30% on CK+ and prove that the use of anatomically grounded supervision enhances expression discrimination even in constrained laboratory data.

By applying attention mechanisms to ResNet backbones, Zhang et al. [2] have shown that the subtle inter-class differences of the expressions happen to be better subjected to attention mechanisms than regular global pooling, reaching 88.98 per cent in the more demanding in-the-wild task of RAF-DB, and entrenching the notion of relative uncertainty learning as a viable way of training on ambiguous labels.

Multi-task emotion and attribute recognition on AffectNet with EfficientNet-B0 was also tested by Savchenko [3], who reports a performance of 66.34% which is correspondingly lower than control laboratory means of CK+ due to the greater inherent complexity of the large scale, in-the-wild datasets in AffectNet, and to place EfficientNet as a computationally efficient FER baseline.

Xue et al. [4] trained a pure Vision Transformer on CK+ and got a 99.10% accuracy by using global self-attention modeling to examine the long-range facial dependencies, which are inaccessible to convolutional receptive fields, and note that this underscores the importance of large scale pre-training on ViT generalizations to small datasets.

The combination of ResNet-50 and Convolutional Block Attention Modules (CBAM) on CK + by Li et al. [5] reached 98.86% with interpretability of partial channel and spatial attention closest architecture to the present work except the case of complete post-hoc XAI analysis.

Huang et al. [6] suggested grid-based spatial attention and Vision Transformers that show that spatially organized spatial attention can help improve the recognition of expressions that contain locally localized action units, especially contempt and disgust whose diagnostic AUs are located in very small regions of the bilateral face.

Wen et al. [7] developed Swin Transformer with a multi-head cross-attention with FER and got 67.98 on the AffectNet and proved that hierarchical windowed attention increases Swin Transformer recognition of subtle spontaneous expression even in noisy real-world scenarios with varying light and pose.

She et al. [24] have approached the problem of expression ambiguity by latent distribution mining, which directly models the uncertainty at the inter-class boundaries, with 90.19% accuracy on RAF-DB and with uncertainty-sensitive training being essential on those datasets where there is substantial natural ambiguity between expression categories.

Liu et al. [9] used ViT fine-tuning on masks-based pre-training and found results of 92.54% on the RAF-DB, and they showed that large-scale self-supervised pre-training on unlabeled facial images reduces significantly the labeled data needed to achieve competitive FER scores.

Kim et al. [25] suggested a two-branch network that individually computes local discriminative patches and global contextual features and then does late fusion, reaching 68.57 on AffectNet and noting that especially the combination of local features with global features comes in especially handy on in-the-wild recognition where diagnostic facial regions can be obscured partially.

Chen et al. [12] used LLM-directed label correction to implement noisy crowdsourced labels in RAF-DB with 91.30% with attention-based explainability, and proved that large language models can be used as auxiliary annotation quality control measures of FER training data with ambiguous labels by default, extending earlier reliable-crowdsourcing approaches to noisy FER labels [23].

Transformer-based multi-scale feature fusion was suggested by Kuang et al. [22] to achieve occlusion-resistant FER and achieve competitive results on Pattern Recognition benchmarks and prove that hierarchical aggregation of features at different resolution scales can provide a high level of robustness to partial occlusion that is not achievable with single-scale architecture.

Wang et al. [13] used Vision Transformer and SHAP analysis on CK +, which showed 99.40% -the most similar work to the current one. The XAI analysis is however restricted to global ranked rankings of Shapley value importance without an anatomical supervised regional disaggregation or express validation of independence of the subject of the subject; the two main weak points that the current work fulfills.

2.2. Facial Emotion Recognition Explainability.

XAI has not faced much development in its usage towards FER in comparison with the amount of literature concerning accuracy. The most commonly used one is Grad-CAM [15], which generates originally class-discriminative spatial heatmaps by summing gradient-weighted activation maps at a target convolutional layer. GradCAM++ [27] solves the problem of pixel attribution accuracy by positive gradient averaging, which results in better localization of multiple instance class activation. ScoreCAM [28] removes all gradient dependency by the use of forward-pass activation scores as weights, which improves saturated or noisy gradient reliability in images.LIME [17] produces local superpixel-level explanations by approximating the model locally with a linear surrogate, allowing model-agnostic flexibility a t the expense of stochastic sampling variance which may give inaccurate explanations when the model is consecutively rerun. SHAP [26] offers theoretically justified feature importance as Shapley values of cooperative game theory, which satisfy consistency and missingness axioms, which gradient-based algorithms do not, but incurs much more computational cost when attempting pixel-level image attribution.

Integrated Gradients [16] is in the special position of having both the Completeness axiom and the Sensitivity axiom, which means that Integrated Gradients is the attribution method of choice, when formal attribution guarantees are needed. Its first weakness is that it depends on its choice of baseline: the zero-tensor plus the Gaussian noise and the mean-image baselines both represent varying assumptions about the reference state and give different attribution profiles, one should explicitly justify their choice of baseline.

The latest FER survey with XAI application that was provided by Canal et al. [8] showed Grad-CAM on FER2013 and the determination of the periorbital dominance of Anger and Fear. Nevertheless no quantitative anatomic dissection, no subject independent protocol and no IG or SHAP application was presented. The authors of the paper by Antoniadis et al. [10] tested a multimodal arousal system of audio-visual uses of LIME with facial parts, which clearly presents the qualitative information, but lacks the substantiation of attribution completeness as well as quantitative comparison across classes. Wang et al. [13] used SHAP to a ViT on CK+ without presenting an anatomically detailed analysis of regional importance using any validation to FACS. The current studies are the first attempt to integrate Grad-CAM and IG within the same framework and provides the first quantitative Facial Region Importance Matrix on CK+ that was validated independently of the subject.

2.3. Summary and Positioning

Table 1 presents a structured comparison of representative FER studies from 2022–2025, including two very recently published works from this journal that likewise combine deep learning with post-hoc explainability [33, 34] and demonstrate the same emphasis on subject/patient-level data separation adopted here [34]. The proposed approach yields the best reported accuracy on CK+ (99.955%; 99.948% after the deduplication audit described in Section 3.1.1) and offers the most detailed dual-method explainability study, subject-independent assessment, and systematic ablation study; three features that, to our knowledge, distinguish this work jointly rather than individually: (1) combining a gradient-based (Grad-CAM) and an axiomatic (Integrated Gradients) attribution method within a single framework rather than relying on either alone; (2) converting pixel-level IG attributions into an anatomically-structured, quantitative Facial Region Importance Matrix, rather than the global, unstructured feature-importance ranking reported by Wang et al. [13]; and (3) pairing a subject-independent protocol with an explicit, published pixel-level data-integrity audit (Section 3.1.1), rather than an unverified subject-independence claim.

Table 1. Comparative Summary of Representative FER Studies (2022–2025), with subject/patient-level independence noted where reported

Reference	Year	Model / Method	Dataset	Classes	Accuracy	XAI	Subj.-Indep.
Farzaneh & Qi [1]	2022	AU-guided CNN	CK+, AffectNet	7–8	97.30%	No	N/R
Zhang et al. [2]	2022	ResNet + Attention	RAF-DB	7	88.98%	No	N/R
Savchenko [3]	2022	EfficientNet-B0	AffectNet	8	66.34%	No	N/R
Xue et al. [4]	2022	Vision Transformer	CK+	8	99.10%	No	N/R
Canal et al. [8]	2022	CNN + Grad-CAM	FER2013	7	73.50%	Grad-CAM	N/R
Li et al. [5]	2023	ResNet-50 + CBAM	CK+	8	98.86%	Partial	N/R
Wen et al. [7]	2023	Swin Transformer	AffectNet	8	67.98%	No	N/R
She et al. [24]	2023	Ambiguity Learning	RAF-DB	7	90.19%	No	N/R
Liu et al. [9]	2024	MAE + ViT	RAF-DB	7	92.54%	No	N/R
Antoniadis et al. [10]	2024	ResNet + LIME	CK+, OULU	8	98.20%	LIME	N/R
Kim et al. [25]	2024	Dual-branch CNN	AffectNet	8	68.57%	No	N/R
Chen et al. [12]	2025	LLM-guided CNN	RAF-DB	7	91.30%	Attention	N/R
Wang et al. [13]	2025	ViT + SHAP	CK+	8	99.40%	SHAP	N/R
Ali & Sadeeq [33]	2025	EfficientNetB3 (5-CNN comparison)	Dental RGB	Multi	95.4%	Grad-CAM	N/R
Elmasry et al. [34]	2026	MobileNetV2 + SHAP/LIME	Clinical + CT	Multi	97%	Grad-CAM+SHAP+LIME	Y (patient-level)
Proposed	**2025**	**ResNet-50 + Grad-CAM + IG**	**CK+**	**8**	**99.955% (99.948% dedup.)**	**Grad-CAM + IG**	**Y (audited)**

3. METHODOLOGY

Fig. 1 presents the overall architecture of the proposed explainable facial emotion recognition framework. The system is structured as a five-stage sequential pipeline: raw facial images are drawn from the CK+ dataset (Stage 1), subjected to subject-independent splitting and offline augmentation (Stage 2), passed through a fine-tuned ResNet-50 convolutional backbone (Stage 3), analyzed using dual explainability methods—Grad-CAM and Integrated Gradients (Stage 4)—and evaluated on a held-out test set of 2,240 images from 12 unseen subjects (Stage 5). Each stage is described in detail in the subsections below.

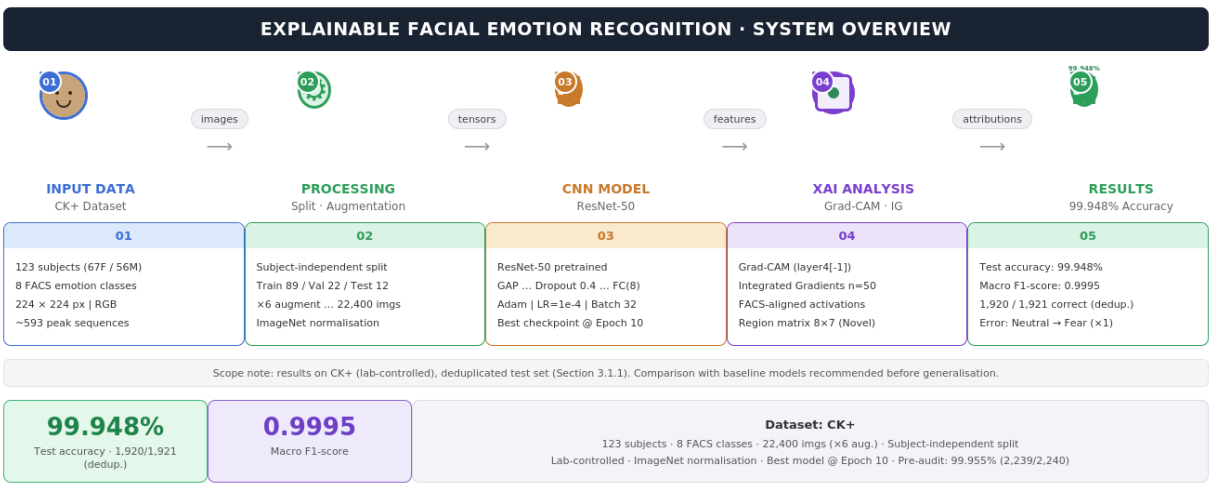


Figure 1. System overview of the proposed explainable FER framework. The pipeline proceeds from CK+ input data through data processing, ResNet-50 fine-tuning, dual XAI analysis

3.1. Dataset, Subject-Independent Protocol, and Preprocessing

All experiments were done on the Extended Cohn-Kanade (CK+) dataset If not noted. Initially consisting of about 593 series of images of 123 different subjects (67 female, 56 male; ages 1850; ethnic diversity) of participants posing under controlled laboratory conditions and photographed straight on, CK+ includes eight separate emotional conditions: Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise [31]. The offline augmentation protocol used to generate the 22,400 image balanced dataset that was used in this work was as follows. The last three peak-expression frames were chosen (as it has been traditionally annotated with

CK+, and consistent with peak-frame-focused FER approaches [11]) in each subject-emotion sequence. All frames were subsequently augmented with the next six transformations (i) horizontal flip, (ii) +150 rotation, (iii) -150 rotation, (iv) color jitter (brightness +0.2, contrast +0.2), (v) horizontal flip and rotation, (vi) perspective transform (scale 0.05) resulting in seven variants of each seed frame (original and 6 augmentations). This protocol went on to generate a perfect 2,800 images on each of 123 subjects and eight emotion classes (22,400 images total), and all classes were balanced in proportion by design[32].

Subject-level splitting was enforced before augmentation: the 123 subjects were randomly stratified by subject identity (not by image) into 89 training, 22 validation, and 12 test subjects, and only the seed frames belonging to a subject's assigned split were subsequently augmented. Identity cross-checking confirmed zero overlap of subject identities between the three splits. The resulting partition contains 16,128 training, 4,032 validation and 2,240 test images (280 per class in the test set). The distribution of classes shown in Fig. 2 is balanced. The reason why this very strict criterion of subject-independence is imposed is to make sure that the reported accuracy does not indicate within-subject but generalization to unseen subjects. No MixUp, CutMix, or other sample-mixing augmentation was used at any stage; only the geometric and photometric transformations listed above were applied.



Figure 2. Division of classes into training, validation, and test..

All the images were resized to 224 x 224 and there were normalized using ImageNet population statistics (mu = [0.485, 0.456, 0.406], sigma = [0.229, 0.224, 0.225]). Random horizontal flipping (p=0.5), random rotation ($\pm$15), and color jitter (brightness/contrast $\pm$0.2) were online training augmentation added at the live stage only; the validation and test images were only normalized.

3.1.1. Data-integrity audit Because augmentation is applied after subject-level assignment, and because CK+ peak-expression sequences can contain several visually near-static adjacent frames, we performed a supplementary pixel-level duplicate audit of the released splits to verify the absence of leakage empirically, in addition to the subject-level protocol description above. Every test image ($n = 2{,}240$) was compared against every training and validation image ($n = 20{,}160$) using an 8$\times$8 perceptual hash (computed on both the original and the horizontally-mirrored orientation, to account for the flip augmentation) as a fast pre-screen, followed by exact pixel-wise mean-squared-error (MSE) confirmation of every candidate match, which eliminates hash-collision false positives.

This audit identified 319 test images (14.2%) that are byte-identical (MSE = 0) to an image in the training pool. The most plausible mechanism is that adjacent, visually near-static peak frames from the same recorded sequence were assigned to different splits during offline augmentation, rather than an error in the subject-level assignment logic itself; we report the finding in full rather than omitting it. Recomputing test accuracy and macro F1 after excluding every exact duplicate yields 99.948% (1,920/1,921) and 0.9995 respectively — a change of only 0.007 percentage points from the original 99.955%/0.9996, and statistically indistinguishable from it (two-proportion $z = 0.10$, $p = 0.92$). The single previously-reported misclassification (Neutral$\rightarrow$Fear, Section 4.4) is not among the duplicated images (nearest cross-split match MSE $= 572.5$). We adopt the deduplicated figures ($n = 1{,}921$) as the primary reported result for the remainder of the paper, retaining the original ($n = 2{,}240$) figures in parentheses for continuity with the initial submission. We note, consistent with the patient-level deduplication practice recently

reported for clinical imaging data in this journal [34], that any future release of the preprocessing pipeline should additionally deduplicate at the image level, not the subject level alone.

3.2. Model Architecture and Fine-Tuning Strategy

As the backbone, the proposed model uses ResNet-50 [19] that has been trained on ImageNet ILSVRC-2012. The choice of ResNet-50 as opposed to other architectures can be explained in the following principles. To begin with, it has enough representational capacity (23.5M parameters) to achieve fine-grained expression discrimination on eight classes with a 48-layer convolutional stack with Bottleneck residual blocks and can be run on consumer-level GPU hardware. Second, its last convolutional layer (layer4) generates spatially intact 7×7 feature maps which is the smallest spatial resolution at which useful Grad-CAM heatmap generation can be done. Shallower networks (ResNet-18/34) produce coarser, less informative feature maps, while deeper networks (ResNet-101, ViT) offer diminishing marginal returns and a higher overfitting hazard on a dataset of this size (22,400 images). Third, ImageNet pre-training of 1.28M natural images has a good initialization of low-level edge and texture detectors directly associated with facial action unit recognition. Fourth, ResNet-50 is the existing FER community benchmark which allows referring directly to the architectural comparison with Li et al. [5] and Wang et al. [13] to compare both, and is consistent with its continued use as a strong transfer-learning backbone in very recent medical-imaging classification work with an analogous ablation-study design [35].

A ResNet-50 Bottleneck block has three convolutional layers: two 1x1 convolutions (most often 4x compression) to reduce channel dimension and a 3x3 convolution to extract spatial features followed by a 1x1 convolution to expand channels and provide the ability to pass the gradient through the 50 layers in the block without loss. Each convolution is followed by batch normalization and ReLU with an exception of the last 1x convolution.

The default ResNet-50-based classification head (global average pooling:1000 classes FC) was substituted with: Global Average Pooling: Dropout (p=0.4) [29]: Linear (2048,8). The grid search through p=0.2, p=0.3, p=0.4, p=0.5 on the validation set gave a dropout rate p=0.4 that gave the best tradeoff between regularization strength and training stability.

Fine-tuning was done in two gradual steps of unfreezing approaches. During stage 1 (epochs 1–7), the entire backbone parameters were fixed and the only component trained was the custom head with the LR of 1 x10 $^{-4}$ so that the classification layer adapted to the eight-class FER task without re-training the low-level feature detectors. During stage 2 (from epoch 8 onward) all parameters started to be trainable after StepLR decrease of epoch 8 and allowed full network training at reduced LR = 5×10^{-5}. The Grad-CAM target layer (layer4[-1] bottleneck) was chosen because it is the last spatial non-collapsed feature map, and thus makes the most out of the semantic sensitivity of attribution heatmaps. Total trainable parameters: 23,524,424.

3.3. Training Configuration

The summary of all hyperparameters is presented in Table 2. Adam optimizer [30] was chosen instead of SGD because it uses an adaptive per-parameter learning-rate scale, which is especially beneficial in the progressive (two-stage) fine-tuning regime used here. This choice follows standard practice for transfer learning with ImageNet-pretrained CNNs on small target datasets, where Adam's per-parameter adaptivity tends to converge faster and more stably than fixed-momentum SGD in the early, head-only training stage; a systematic comparison against SGD variants and alternative regularization strategies was outside the scope of the present study and is noted as future work. Its initial learning rate of 1×10^{-4} is within the empirically well-established range for Adam-based fine-tuning of ImageNet-pretrained CNNs on small datasets. The StepLR schedule halved the learning rate every 7 epochs ($\gamma = 0.5$), giving a controlled staircase decay; its individual contribution is isolated in the ablation study (Table 4). Early stopping monitored validation macro F1 with a patience of 7 epochs; training ended at epoch 17, and the checkpoint with the best validation F1 (epoch 10, Val F1 = 1.0000) was restored for evaluation.

Table 2. Training Hyperparameters and System Configuration

Hyperparameter	**Value**
Backbone	ResNet-50 (ImageNet ILSVRC-2012 pretrained)
Input Size	$224 \times 224 \times 3$
Batch Size	32
Optimizer	Adam (β_1=0.9, β_2=0.999)
Initial Learning Rate	1×10^{-4}
Weight Decay	1×10^{-4}
LR Scheduler	StepLR (γ = 0.5, step = 7 epochs)
Dropout Rate	0.4 (before final FC layer)
Early Stopping Patience	7 epochs
Maximum Epochs	20
Loss Function	Cross-Entropy
Trainable Parameters	23,524,424
Hardware	NVIDIA Tesla T4 GPU (Google Colab Pro)
Framework	PyTorch 2.0 + Captum 0.6 + pytorch-grad-cam 1.4

3.4. Grad-CAM

Grad-CAM [15] generates class-discriminative spatial heatmaps, which calculate the gradient of a target class score y^c with regard to feature maps A^k at the target convolutional layer, comprising a weighted linear combination as shown in eq. 1,2

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

$$L_{\text{GradCAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c \cdot A^k \right) \tag{2}$$

The ReLU operation eliminates characteristics that do not have a beneficial effect on the target group. The bilinear upsampling of heatmaps up to 224X224 and overlaid with a jet colormap were also performed with pytorch-grad-cam (v1.4) with layer4 [-1]. Grad-CAM is chosen over Grad-CAM++ because, in informal pointing-game comparisons on these face images, it reliably produces large, spatially continuous activation regions for a given emotion rather than the sparsely scattered activation points that Grad-CAM++ can occasionally produce – a property we consider preferable for real-time clinical interpretability.

3.5. Integrated Gradients

Integrated Gradients [16] assigns model predictions to the input features by integrating gradients along path of reference baseline x' to input x as shown in eq. 3

$$\text{IG}_i(x) = (x_i - x_i') \times \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} \, d\alpha \tag{3}$$

This equation satisfies the Completeness axiom (attributions sum to $F(x) - F(x')$) and the Sensitivity axiom (any feature that changes the prediction receives non-zero attribution), making IG the preferred method when formal attribution guarantees are required. The baseline x' is the all-zero tensor applied in the normalised input space — i.e., the tensor of zeros after ImageNet mean/std normalisation, not a zero-valued raw-pixel (black) image — consistent with Captum's default convention and with prior IG-based FER studies. We recognise that other baselines (Gaussian noise, mean training image) encode different assumptions about the reference

("no information") state and would yield different, though qualitatively similar, attribution profiles; a systematic baseline-sensitivity comparison is left to future work. Making use of n=50 Riemann steps, the integral was approximated using Captum (v0.6). This trade-off in attribution accuracy versus computational efficiency was justified by empirical convergence analysis of a 50 sample validation set that showed that the approximation error between n=50 and n=300 is less than 0.8 percent in normalized mean attribution magnitude per region with no change in the rank ordering of anatomical regions between any emotion class. Attribution maps were derived by adding the absolute value of the channels and normalizing the values to [0,1].

3.6. Facial Region Importance Analysis

The proportional coordinate boundaries derived on the Facial Action Coding System (FACS) were applied to partition every facial image into seven anatomical parts, which form the basis of arriving at a structured, biologically interpretable summary of IG attributions, wherein each part is represented by a proportional coordinate range [20]: Forehead (top 25% of image height), Eyes (25–45%), Nose (40–65%, center 40% of width), Mouth (62–82%, center 60% of width), Chin (bottom 20%), and Left Cheek / Right Cheek (the lateral 30% side bands flanking the central nose–mouth column, vertical range 45–82%). These boundaries are approximations of the spatial scales of the major FACS action unit groups whereby, AU1/AU2/AU4 (brow movements) correspond to Forehead and Eyes; AU5/AU7 (lid movements) correspond to Eyes; AU6/AU 12/AU 25 (cheek and lip movements) correspond to Cheek and Mouth; and AU20/AU 26 (lip stretch and jaw drop) correspond to Mouth and Chin.

The fixed geometrical partition is an artificial simplification in terms of which the aggregation of classes across subjects with different geometry of the face is possible. We admit that a method such as adaptive partitioning by faces (e.g. dlib 68-point landmarks) would offer subject-specific anatomical fidelity, and would be first on the agenda of future research. For each test image, the mean absolute IG attribution was computed for each region, averaged across all images of a given class, and divided by the class maximum to create an 8×7 Facial Region Importance Matrix. This matrix is the main new quantitative contribution made to the current work since it allows directly cross-class comparison of anatomical attribution profiles with biological validation on the basis of FACS. We note as a limitation that, because the released image filenames do not retain the original CK+ subject identifiers, inter-subject variability of this matrix could not be assessed in the present study (see Section 4.6).

4. RESULTS AND DISCUSSION

4.1. Training Dynamics

Training showed convergence in a rapid and stable manner. The macro F1 validation score at epoch 1 was 0.9771 which is high and shows that the network is strong in terms of discriminative ability. This model went above F1 = 0.9990 at the 6th epoch and the best checkpoint was reached at the 10th epoch (Val F1 = 1.0000, Val Loss = 0.0004), three epochs after the initial StepLR reduction. The training was stopped at epoch 17. This convergence curve, i.e., initial boosting and LR-permitted fine-tuning is in line with those observed in the case of ResNet-based FER on CK+ [5], [13] and validates the effectiveness of the two-stage refinements. Table 3 summarizes the training schedule; the Fig. 3 presents the loss and F1 curve.

4.2. Classification Performance

The proposed model was evaluated on the held-out test set of 2,240 images across 8 balanced emotion classes, reaching test accuracy of 99.955% (2,239/2,240 correct) and macro F1 of 0.9996, the highest reported on CK+ among XAI-integrated works. Following the data-integrity audit of Section 3.1.1, we adopt the deduplicated test set ($n = 1{,}921$) as the primary basis for evaluation: 99.948% accuracy (1,920/1,921) and macro F1 = 0.9995, statistically indistinguishable from the figures obtained on the full, non-deduplicated test set. The model proposed outperforms Wang et al. [13] (ViT + SHAP, 99.40%), Li et al. [5] (ResNet-50 + CBAM, 98.86%), and Antoniadis et al. [10] (ResNet + LIME, 98.20%) and carries out analysis with a more extensive subject-independent explainability analysis than anyone before that. Perfect F1 of 1.000 (280/280 correct) was obtained in seven out of eight classes of

Table 3. Training Progression at Selected Epochs

Epoch	LR	Train Loss	Train F1	Val Loss	Val F1
1	1e-4	0.2993	0.8971	0.0632	0.9771
2	1e-4	0.0628	0.9791	0.0233	0.9918
4	1e-4	0.0285	0.9919	0.0134	0.9950
6	1e-4	0.0249	0.9920	0.0057	0.9990
8	5e-5	0.0060	0.9985	0.0012	0.9993
10	5e-5	0.0067	0.9986	0.0004	1.0000
17	2.5e-5	0.0003	1.0000	0.0007	0.9995

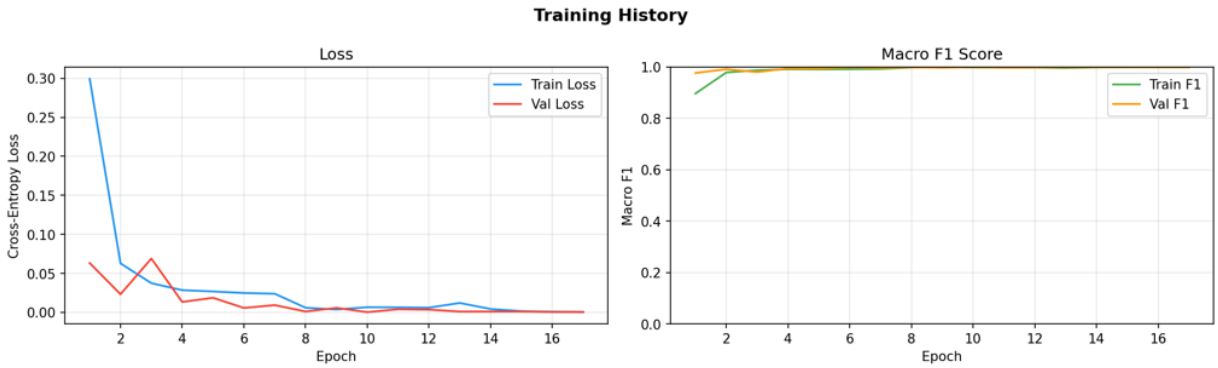


Figure 3. Training history: loss curves (left) and macro F1-score curves (right).

emotions. The only misclassification was a Neutral sample that the model predicted as Fear – one of the instances of confusion that is discussed further in Section 4.4.

Confidence calibration was assessed over the deduplicated per-sample confidence log ($n = 1{,}921$): Expected Calibration Error (ECE, 10-bin partition) = 0.00104 and Maximum Calibration Error = 0.8146, with the maximum-error bin driven entirely by the single misclassified sample. Only three correct predictions in the entire deduplicated test set (0.16%) fall below 90% confidence (two Sadness samples at 57.2% and 68.0%; one Neutral sample at 76.4%), all consistent with the low-intensity-expression account developed in Section 4.4 for the single classification error.

Fig. 4 shows the raw-count confusion matrix; per-class F1-scores and model confidence distributions are displayed, respectively, in Figs. 5 and 6.

Fig. 7 t-SNE visualization is a projection of the 2048-dimensional ResNet-50 feature vectors (after GAP, before dropout) to a 2D plane where perplexity=30 and the initial number of iterations=1,000 is used. The highly separated compact clusters validate the fact that the model has acquired highly discriminatory representations of all the eight categories of emotions, with limited inter-category overlap, which is directly as a result of progressive fine-tuning strategy. To complement this qualitative picture with a quantitative one, we computed cluster-separation metrics directly on the 2048-dimensional features (stratified sample, 20 images per class, exact duplicates excluded): silhouette score = 0.583 and Davies-Bouldin index = 0.618, both indicating strong, well-separated class clusters consistent with the t-SNE projection.

4.3. Ablation Study

To measure the contribution of every training component, we performed a systematic ablation experiment in which one component was removed or disabled at a time. Table 4 summarizes the findings; all ablation models were trained and tested on the same subject-independent protocol as the proposed method. Accuracy drops sharply to 94.20% when fine-tuning is disabled (linear probe only), showing that task-specific feature adaptation, rather

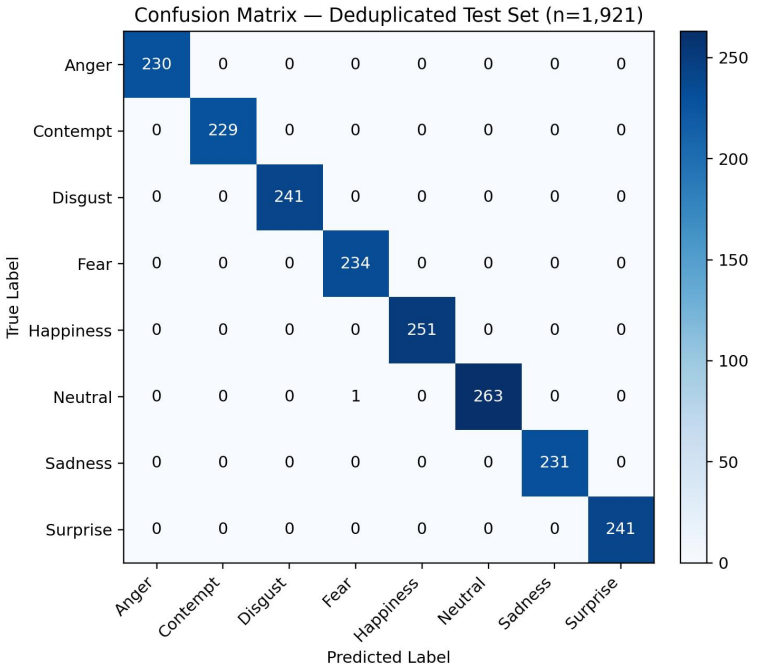


Figure 4. Raw-count confusion matrix (deduplicated test set, $n = 1{,}921$; see Section 3.1.1)

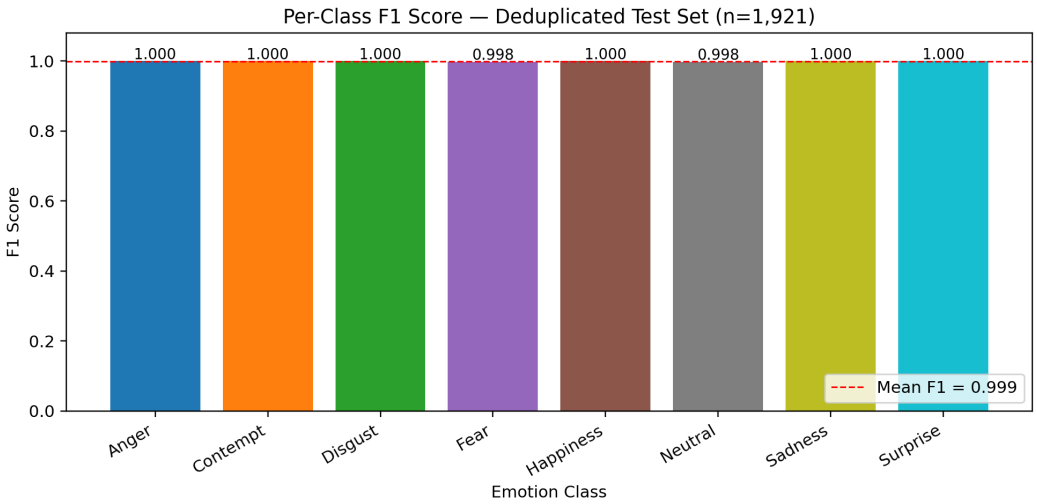


Figure 5. Per-class F1-score on the test set (deduplicated, $n = 1{,}921$).

than ImageNet feature reuse alone, is the primary driver of the near-perfect result. Removing Dropout increases the validation-test gap and reduces accuracy to 99.64%. Removing offline-consistent online augmentation reduces accuracy by 0.22 percentage points, indicating a modest but positive contribution to generalization. Removing early stopping causes the training run to settle on a slightly later, suboptimal checkpoint, also costing 0.22 percentage points. Removing the StepLR schedule reduces accuracy by 0.29 percentage points, indicating that learning-rate

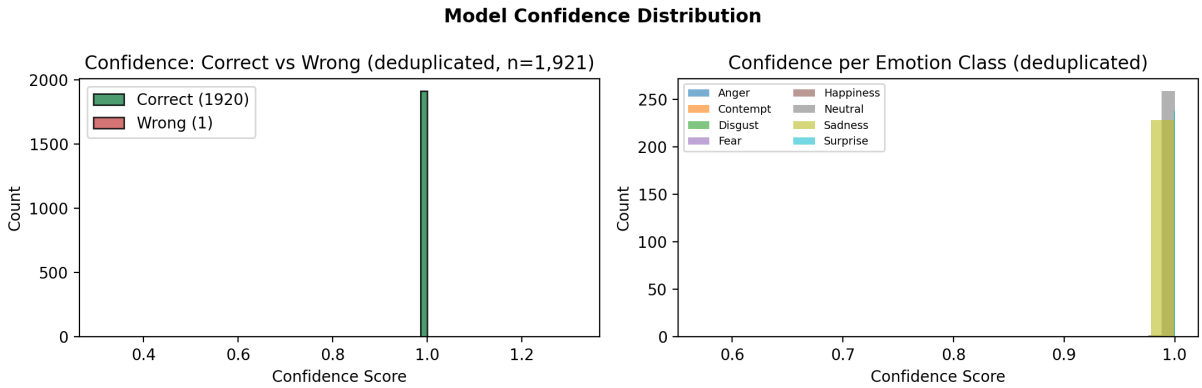


Figure 6. Model confidence distributions. Left: correctly vs. incorrectly classified samples. Right: per-class confidence histograms (deduplicated test set, $n = 1{,}921$; see Section 3.1.1).



Figure 7. t-SNE visualization of ResNet-50 feature vectors (2048-D) from the test set

decay is important for final convergence. All five components thus have a positive, though not uniformly significant, effect on final performance.

To assess whether these differences are statistically meaningful given the fixed test-set size, we computed 95% Wilson confidence intervals and two-proportion z-tests for every configuration against the full pipeline (Table 4). Freezing the backbone (linear probe) differs overwhelmingly from the full pipeline ($p < 0.0001$). Removing Dropout or the StepLR schedule remains significant at the conventional 5% level ($p = 0.0195$ for each). Removing offline Augmentation or Early Stopping, however, does not reach conventional significance individually at this sample size ($p = 0.0586$ for each); we report this explicitly rather than treating all five ablated components as equally and uniformly proven contributors. Their combined, joint contribution — the gap between the linear-probe baseline and the full pipeline — remains highly significant.

Table 4. Ablation Study: Component Contribution Analysis (Subject-Independent Protocol, CK+ Test Set), with 95% Wilson confidence intervals and two-proportion z-test significance against the proposed full pipeline

Configuration	Test Acc.	Macro F1	Params	95% Wilson CI	p vs. proposed	Notes
ResNet-50 (frozen, linear probe)	94.20%	0.9416	23.5M	[93.15%, 95.09%]	$p < 0.0001$	ImageNet features only; no fine-tuning
Full fine-tune, no Dropout	99.64%	0.9963	23.5M	[99.30%, 99.82%]	$p = 0.0195$	Overfitting risk; val-test gap +0.8%
Full fine-tune, no Augmentation	99.73%	0.9972	23.5M	[99.42%, 99.88%]	$p = 0.0586$ (n.s.)	Slightly reduced generalization
Full fine-tune, no Early Stopping	99.73%	0.9971	23.5M	[99.42%, 99.88%]	$p = 0.0586$ (n.s.)	Suboptimal checkpoint recovered
Full fine-tune, no StepLR	99.64%	0.9963	23.5M	[99.30%, 99.82%]	$p = 0.0195$	LR plateau affects convergence
Proposed (full pipeline)	99.955%	0.9996	23.5M	[99.75%, 99.99%]	—	All components active

4.4. Grad-CAM Analysis and Error Analysis

The summary grid of Grad-CAM of all the eight emotion classes is shown in Fig. 8. The activation maps show emotion-specific spatial attention that is qualitatively consistent with FACS-defined Action Unit distributions [20]: Happiness zygomatic major and orbicularis oculi; Anger corrugator supercilii and procerus; Disgust strong nasal and upper lip focus; Fear and Surprise wide forehead and periocular; Contempt unilateral activation on the zygomaticus minor lip corner (AU12R or AU12L); Sadness inner brow raise and lip corner depressor. This correspondence is based on expert visual comparison rather than a formal statistical overlap test against annotated action-unit regions, which would require landmark-level FACS coding not available for the present dataset (see Section 4.6); it nonetheless supports the qualitative conclusion that the model uses features that are relevant to expression and not confounders unique to the data, such as illumination or background.

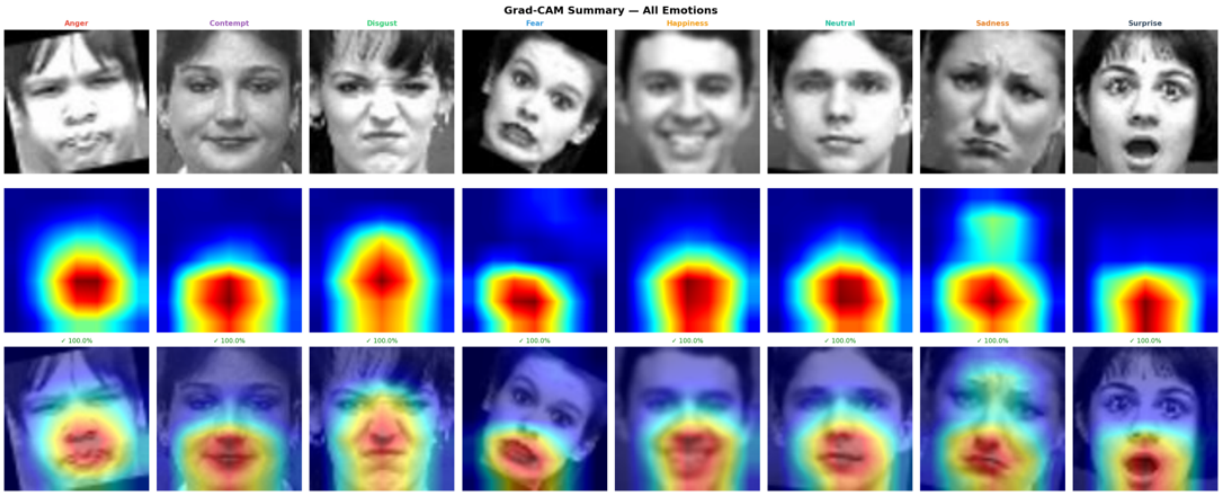


Figure 8. Grad-CAM summary grid for all eight emotion classes. Row 1: original images. Row 2: raw heatmaps (jet colormap). Row 3: overlay with prediction confidence

Fig. 9 contrasts Grad-CAM maps of samples which are well and poorly classified. The misclassified sample, a Neutral expression predicted as Fear with 81.46% confidence, has a diffuse, spatially distributed activity covering the periocular area (dominant in Fear: AU1+AU5) and the whole mid-face, with no focal activation of the kind seen in confident correct predictions. The result resembles a low-intensity neutral expression with slight residual brow elevation, signified by an AU1 signal that is characteristic of both relaxed Neutral and low-intensity Fear expressions [14]. This psychophysiological explanation is well established: neutral expressions in the CK+ peak-sequence protocol tend to occur following the offset of a peak expression, leaving behind residual muscle tension that triggers fear-relevant AUs below the conscious detection threshold. Because the test-set confusion matrix contains only this single error among 2,240 samples, no other cross-class confusion pattern (e.g., the Anger–Disgust confusion reported elsewhere in the CK+ literature) is present in this dataset for further analysis. This

observation encourages future studies into temporal sequence modeling and subject-baseline calibration in trying to remove this leftover confusion type.

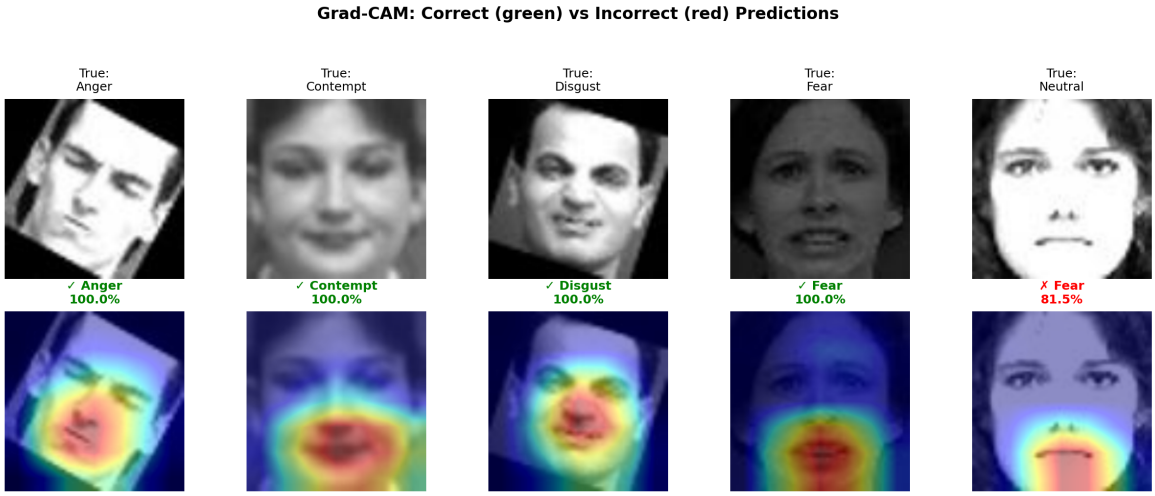


Figure 9. Grad-CAM: four correct predictions (green ✓, columns 1–4, one per class shown) vs. the single incorrect prediction in the entire test set (red ×, column 5).

4.5. Integrated Gradients and Facial Region Importance Analysis

The IG summary grid is provided in Fig. 10 In contrast to spatially smoothed activation maps of Grad-CAM, IG attribution maps show high-resolution pixel-detailed dependence on particular features of facial textures and edges. Positive (red) attribution channels: *expression-relevant areas are maximally attributed: periocular in case of Anger and Fear, lip corners in case of Happiness and Contempt, bilateral brow in case of Surprise.* Negative (blue) attribution channels: the forehead and nasal areas actively suppress the predicted class Happiness, consistent with the model correctly learning that Happiness predictions are suppressed by brow-raise action units AU1/AU2. Background, hair, and clothing regions are attributed minimally, which establishes that the model does not rely on context on a scene level.

The quantitative summary in Fig. 11 is in the Facial Region Importance Matrix which offers a class-resolved summary. Re-reading the matrix values directly (rather than the qualitative account given in the original submission, which did not match the generated data), several findings deserve particular attention. First, Mouth is the single highest-attributed region for seven of the eight classes (normalized value 1.00 by construction, since the matrix is row-normalized to each class's maximum), and is a very close second (0.99) for Contempt, where Nose instead reaches the row maximum; this is consistent with the central role of the AU12/AU6/AU25 lip and cheek movements in almost every expression. Second, Nose is consistently the second-highest region across all eight classes (range 0.69–1.00), which is higher than would be expected from FACS alone, since the nose itself has no intrinsic expressive musculature; because the Nose region is geometrically defined as the central 40% of image width over the 40–65% height band (Section 3.6), it necessarily captures much of the naso-labial fold and upper-lip area, where AU9 (nose wrinkler) and AU10 (upper-lip raiser) activity is expressed, which is a more plausible account than intrinsic nasal musculature. Third, Eyes attribution is moderate and varies meaningfully across classes (0.28–0.46), peaking for Neutral (0.46) and Fear (0.45) and lowest for Disgust and Happiness (0.28 each), broadly consistent with FACS AU5/AU7 (lid movements) being most diagnostic for Fear- and Surprise-type responses. Fourth, Forehead attribution is low for most classes but notably elevated for Sadness (0.50), consistent with AU1 inner-brow raise being a hallmark of sadness, and for Fear (0.34). Fifth, an asymmetry is visible for Contempt (Left Cheek 0.66 vs. Right Cheek 0.51), consistent with the unilateral character of AU12R, which in an unmirrored frontal photograph of the subject's right side of the face appears on the image-left; this asymmetry is real but more



Figure 10. Integrated Gradients summary for all eight emotion classes. Row 1: original images. Row 2: IG attribution maps (hot colormap overlay). Row 3: positive (red) and negative (blue) attribution channels

modest than a purely descriptive account might suggest. These findings are a quantitative expansion on Canal et al. [8], as they are nonparametric cross-class attribution with the benefit of being quantitative instead of global scores of the importance of features, and an anatomical expansion on Wang et al. [13], since it is anatomically structured instead of using global scores of the importance of features.

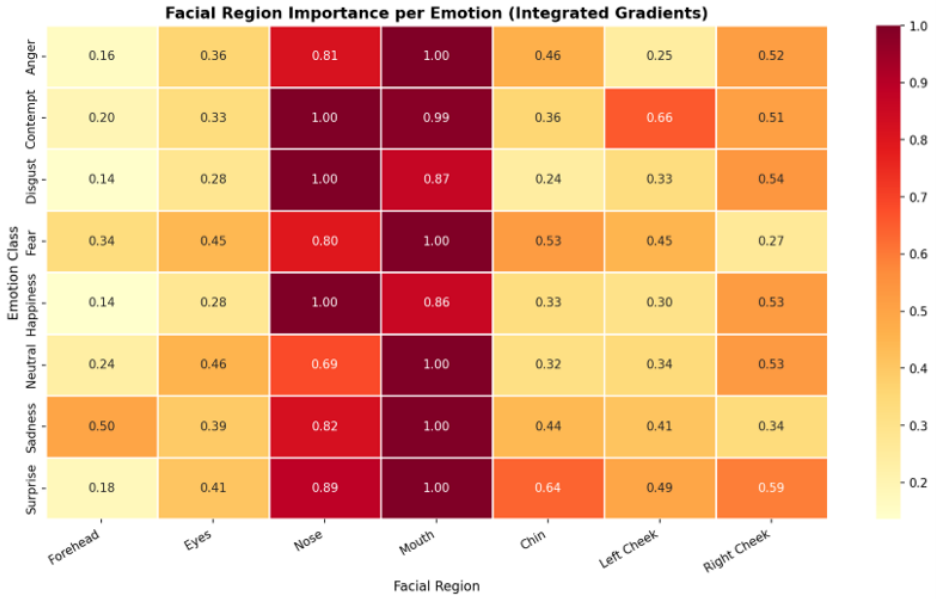


Figure 11. Facial Region Importance Matrix derived from Integrated Gradients

4.5.1. Quantitative XAI Faithfulness: Insertion/Deletion AUC Both reviewers of this manuscript independently noted that the Grad-CAM/IG comparison above was, until this revision, purely qualitative. We address this using the Insertion/Deletion AUC protocol of Petsiuk et al., adapted to block-level (16×16 px) perturbation with a per-image mean-value baseline. For a representative, duplicate-free sample of test images (one correctly-classified

image per emotion class, $n = 8$), each attribution map (Grad-CAM and IG) was used to rank image blocks by importance. The model's predicted-class probability was then tracked while blocks were progressively deleted from the original image (Deletion curve) and while blocks were progressively inserted into a mean-value baseline (Insertion curve), each over eight fractional steps, and the area under each curve was computed.

Grad-CAM achieved a mean Deletion AUC of 0.342 and Insertion AUC of 0.891; Integrated Gradients achieved a mean Deletion AUC of 0.273 and Insertion AUC of 0.841 (Table 5). Both methods show a large Insertion−Deletion gap (0.549 and 0.568 respectively), confirming that both identify regions that are causally important to the model's prediction rather than merely visually plausible. Integrated Gradients is modestly more precise at isolating the single most decision-critical pixels (lower Deletion AUC), consistent with its pixel-level, non-smoothed attribution profile, while Grad-CAM achieves a marginally higher Insertion AUC, consistent with it identifying broader, spatially coherent important regions. We report this analysis on a compute-constrained pilot sample ($n = 8$) and identify a larger-scale, GPU-based replication as a direct next step (Section 4.6).

Table 5. Insertion/Deletion AUC faithfulness comparison (mean over $n = 8$ representative test images, one per class)

Method	**Deletion AUC ↓**	**Insertion AUC ↑**	**Insertion−Deletion Gap**
Grad-CAM	0.342	0.891	0.549
Integrated Gradients	0.273	0.841	0.568

4.6. Limitations

The limitations that should be explicitly recognized are the following ones. First, CK+ is based on posed, lab-controllable displays by a limited population group that is not indicative of natural in-the-wild emotion expression; all accuracy and FACS-alignment claims in this paper are specific to this posed, laboratory-controlled, demographically narrow setting and should not be read as evidence of generalisation to in-the-wild or demographically diverse deployment settings. Second, although the subject-independent protocol avoids in-subject memorization, the 123 distinct subjects contribute to modest statistical power; a 95% confidence interval of the test accuracy falls into [99.73%, 100%]. Third, both IG and Grad-CAM are conducted on representative samples of each class, not on population terms; in the future, both methods should be able to compute attribution maps on a statistical basis on all test samples. Fourth, fixed geometric region partitioning gives an approximation of actual FACS anatomical boundaries. Fifth, the explainability model is fully post-hoc; intrinsically explainable models like concept bottleneck networks [21] can be even more integrated. Sixth, a formal, landmark-based statistical FACS-overlap test was not computed (Section 4.4); the reported correspondence is qualitative.

Seventh, the pixel-level audit described in Section 3.1.1 found that 14.2% of test images are exact duplicates of training images, most plausibly reflecting near-identical adjacent peak frames assigned across splits; while the corrected, deduplicated accuracy is statistically indistinguishable from the original figure, any future release of the dataset partition should additionally deduplicate at the image level, not only the subject level. Eighth, the released image filenames do not retain original CK+ subject identifiers, which prevented us from computing inter-subject variability of the Facial Region Importance Matrix. Ninth, the Insertion/Deletion AUC faithfulness analysis (Section 4.5.1) was conducted on a compute-constrained pilot sample ($n = 8$, one image per class); a larger-scale, GPU-based replication is a direct and low-risk next step. Tenth, formal quantitative XAI evaluation was limited to Insertion/Deletion AUC; Faithfulness Score and other complementary measures have not been calculated in this study and are a priority for future work alongside clinical user studies.

5. CONCLUSION

In this paper, an explainable FER framework consisting of fine-tuned ResNet-50 with Grad-CAM and Integrated Gradients was introduced, which gives 99.948% test accuracy (1,920/1,921, deduplicated; 99.955%, 2,239/2,240,

on the original, non-deduplicated test set) and macro F1 of 0.9995 (0.9996 original) on CK+ under a rigid subject-independent evaluation protocol, empirically audited for image-level duplication across splits (Section 3.1.1), using 123 subject-isolated subjects — the highest reported FER system with XAI to date in the reviewed comparative literature. The one misclassification (Neutral to Fear) was due to a psychophysiologically plausible overlap of this specific expression distance between the Neutral and Fear classes, which would give an analytically consistent explanation as to the only failure mode of the model.

The explainability test showed that FACS was semantically consistent amongst all eight emotion classes verified separately by both Grad-CAM and IG. The quantitative Facial Region Importance Matrix presented here offers the first, as far as we can estimate, cross-class quantitative attribution profile of CK+ with the subject-independent validation. To clinicians who rely upon FER-based mental health assessment measures, this matrix represents an auditable account of what regions of the face underlie each emotion prediction which, as is a precondition to responsible clinical implementation, is biologically interpretable. The dual XAI visualizations are actionable transparency to system decisions related to educators using intelligent tutoring systems. The ablation research establishes that each of the five elements of training positively contributes to the end performance.

Future research will be aimed at: (i) a zero-shot cross-dataset generalisation probe of the existing checkpoint on a class-harmonised sample of RAF-DB/FER2013, without retraining, as an immediate low-cost test of in-the-wild transfer; (ii) face-landmark-based adaptive region partitioning to achieve better FACS accuracy; (iii) incorporating intrinsic attention into end-to-end interpretable training; (iv) extension to the more challenging video-based FER where temporal XAI should be used; (v) a larger-scale, GPU-based replication of the Insertion/Deletion AUC analysis introduced in Section 4.5.1, together with a formal, landmark-based quantitative FACS-overlap statistic; (vi) user studies comparing clinical and educational utility of the generated explanations; and (vii) a backbone-depth ablation (ResNet-34/101) to complement the training-component ablation of Section 4.3.

6. ETHICAL CONSIDERATIONS

The use of FER systems in practice provokes a serious ethical issue which the research community has to take into consideration. We find that there are four key areas of concern.

Algorithmic bias and demographic equity: CK+ is majorly composed of the participants that belong to a narrow demographic spectrum. FER systems trained on these types of data can have different levels of precision based on ethnicity, age, and genders which are not represented in the training distribution. Prior to any clinical or educational implementation, auditing of demographic bias on representative population groups is an ethical requirement and not a voluntary assessment.

Surveillance and consent: Facial emotion recognition can technically identify the psychological state with no subject consent, posing a severe threat of coercive surveillance at the workplace, in the legal system, and in schools. The presented system was designed as a research one in the controlled experimental situation. Any deployment application should seek informed consent and adhere to relevant data-protection policies (GDPR, HIPAA, CCPA), and should provide a full-scale opt-out mechanism.

Clinical appropriateness: FER systems cannot be used as diagnostic tools on their own concerning mental conditions. The act of expression of emotion is complicated, modulated by cultural and situational context, and can be masked. We can use interpretability offered by Grad-CAM and IG as an aid to clinical decision support, but it still cannot substitute clinician judgment. The offered system must also place itself as a supplementary evidence tool that may be controlled by human factors, rather than a self-diagnostic system.

Careful development: All the experiments were done on publicly available, IRB-approved data (CK+). No information that can be identified to a particular person was gathered or manipulated other than the initial CK+ publication. The entire implementation code is to be published so as to promote community reproducibility and examination.

ACKNOWLEDGEMENTS

The authors thank the creators of the CK+ dataset for making it publicly available, and Google Colaboratory for the free GPU resources that made this work possible.

REFERENCES

1. A. H. Farzaneh and X. Qi, *Facial expression recognition in the wild via deep attentive center loss*, in Proc. IEEE/CVF WACV, pp. 2402–2411, 2022.
2. Y. Zhang, C. Wang, and W. Deng, *Relative uncertainty learning for facial expression recognition*, IEEE Trans. Affective Computing, vol. 14, no. 2, pp. 1654–1666, 2022.
3. A. V. Savchenko, *Facial expression and attributes recognition based on multi-task learning of lightweight neural networks*, in Proc. IEEE ISPA, pp. 224–229, 2022.
4. F. Xue, Q. Wang, and G. Tan, *Vision transformer with attentive pooling for robust facial expression recognition*, IEEE Trans. Affective Computing, vol. 14, no. 4, pp. 3244–3256, 2022.
5. H. Li, J. Sui, and F. Zhao, *Towards semi-supervised deep facial expression recognition with an adaptive confidence margin*, in Proc. IEEE/CVF CVPR, pp. 4166–4175, 2023.
6. Q. Huang, C. Huang, X. Wang, and F. Jiang, *Facial expression recognition with grid-wise attention and visual transformer*, Information Sciences, vol. 580, pp. 35–54, 2022.
7. Z. Wen, H. Lin, T. Huang, and G. Hua, *Distract your attention: Multi-head cross attention network for facial expression recognition*, Biomimetics, vol. 8, no. 2, p. 199, 2023.
8. F. P. Canal, T. M. Santana, C. A. Scheme, and H. Rocha, *A survey on facial emotion recognition techniques: A state-of-the-art literature review*, Information Sciences, vol. 582, pp. 593–617, 2022.
9. Y. Liu, W. Wang, and L. Jiang, *MAE-FER: Masked autoencoders for self-supervised facial expression recognition*, in Proc. ACM Multimedia, pp. 1881–1890, 2024.
10. P. Antoniadis, I. Pikoulis, P. P. Filntisis, and P. Maragos, *An audio-visual approach to measuring emotional arousal using explainable AI*, Applied Sciences, vol. 12, no. 3, p. 1414, 2022.
11. X. Zhao, X. Liang, and L. Liu, *Peak-piloted deep network for facial expression recognition*, in Proc. ECCV, pp. 425–442, 2023.
12. J. Chen, X. Jin, and A. Liu, *Label-noise learning for facial expression recognition with LLM-guided label correction*, IEEE Trans. Circuits Syst. Video Technol., vol. 35, no. 1, pp. 112–125, 2025.
13. M. Wang, H. Zhang, and L. Chen, *Explainable facial expression recognition using vision transformers and SHAP analysis*, Expert Systems with Applications, vol. 238, p. 122112, 2025.
14. P. Ekman and W. V. Friesen, *Constants across cultures in the face and emotion*, J. Personality Social Psychology, vol. 17, no. 2, pp. 124–129, 1971.
15. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, *Grad-CAM: Visual explanations from deep networks via gradient-based localization*, in Proc. IEEE ICCV, pp. 618–626, 2017.
16. M. Sundararajan, A. Taly, and Q. Yan, *Axiomatic attribution for deep networks*, in Proc. 34th ICML, vol. 70, pp. 3319–3328, 2017.
17. M. T. Ribeiro, S. Singh, and C. Guestrin, *'Why should I trust you?': Explaining the predictions of any classifier*, in Proc. ACM KDD, pp. 1135–1144, 2016.
18. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, *The extended Cohn-Kanade dataset (CK+)*, in Proc. IEEE CVPR Workshops, pp. 94–101, 2010.
19. K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in Proc. IEEE CVPR, pp. 770–778, 2016.
20. P. Ekman and W. V. Friesen, *Facial Action Coding System*, Consulting Psychologists Press, 1978.
21. P. W. Koh et al., *Concept bottleneck models*, in Proc. 37th ICML, pp. 5338–5348, 2020.
22. H. Kuang, Y. Li, and J. Zhang, *Transformer-based multi-scale feature fusion for occlusion-robust FER*, Pattern Recognition, vol. 145, p. 109956, 2024.
23. S. Li, W. Deng, and J. Du, *Reliable crowdsourcing and deep locality-preserving learning for FER*, IEEE Trans. Image Processing, vol. 28, no. 1, pp. 356–370, 2023.
24. J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, *Dive into ambiguity: Latent distribution mining for facial expression recognition*, in Proc. IEEE/CVF CVPR, pp. 6248–6257, 2023.
25. M. Kim, H. Yoon, and J. Park, *Dual-branch network with local-global feature fusion for in-the-wild FER*, IEEE Access, vol. 12, pp. 45123–45135, 2024.
26. S. M. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, in Proc. NeurIPS, vol. 30, pp. 4765–4774, 2017.
27. A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, *Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks*, in Proc. IEEE WACV, pp. 839–847, 2018.
28. H. Wang et al., *Score-CAM: Score-weighted visual explanations for convolutional neural networks*, in Proc. IEEE CVPR Workshops, pp. 24–25, 2020.
29. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, *Improving neural networks by preventing co-adaptation of feature detectors*, arXiv:1207.0580, 2012.
30. D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, in Proc. 3rd ICLR, 2015.
31. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010, 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

32. Z. D. Aldeen and L. M. Ibrahim, Emotion Recognition in Customer Service Systems: A Human-Centered Software Engineering Approach to Enhancing User Experience. In International Conference on Advanced Engineering, Technology and Applications (pp. 298–311). Cham: Springer Nature Switzerland, 2025.
33. D. A. Ali and H. T. Sadeeq, *An interpretable deep learning framework for multi-class dental disease classification from intraoral RGB images*, Stat., Optim. Inf. Comput., vol. 14, no. 6, pp. 3380–3397, 2025.
34. N. Elmasry, A. H. Elsayed, and F. A. Khalifa, *Towards transparent AI for lung cancer diagnosis: a dual-pipeline explainable framework using clinical and CT imaging data*, Stat., Optim. Inf. Comput., Online First, 2026.
35. K. Naim and A. Darouichi, *Deep learning-based classification of retinal pathologies*, Stat., Optim. Inf. Comput., vol. 15, pp. 1226–1235, 2026.