Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20260502053
Volume 5, Issue2 (May-August 2026), PP: 466-478.
www.indjcst.com
ISSN No: 2583-5300

Deep Learning Based Facial Emotion Recognition System

Kiran Kumar Raja[1], P Sanjay Kumar[2], Ch Venkata Gowtham[3]
[1]*Assistant Professor, Department of Computer Science and Engineering, Vignan's Foundation for Science, Technology and Research, Andhra Pradesh, India.*
[2,3] *Department of Computer Science and Engineering, Vignan's Foundation for Science, Technology and Research, Andhra Pradesh, India.*

To Cite this Article: *Kiran Kumar Raja[1], P Sanjay Kumar[2], Ch Venkata Gowtham[3], "Deep Learning Based Facial Emotion Recognition System", Indian Journal of Computer Science and Technology, Volume 05, Issue 02 (May-August 2026), PP: 466-478.*

Abstract: *Facial Emotion Recognition (FER) constitutes a fundamental building block of affective computing and human-computer interaction, equipping machines with the capacity to interpret and respond to human emotional states in real time. Even with considerable advances in deep learning, contemporary FER systems continue to face difficulties related to cross-dataset gen-eralization, ambiguity among visually similar emotion categories, and severe class imbalance inherent in curated laboratory datasets. This study proposes a domain-adapted deep Convolutional Neural Network (CNN) trained entirely from scratch on a unified corpus assembled from the Extended Cohn-Kanade (CK+) and Karolinska Directed Emotional Faces (KDEF) benchmarks. The merged dataset comprises 6,530 labeled grayscale facial images at 48×48 pixel resolution, partitioned into 5,224 training samples and 1,306 test samples distributed across ten emotion categories. The proposed network consists of four progressively expanding convolutional blocks, each incorporating paired 3×3 convolution layers, batch normalization, ReLU activation, 2×2 max-pooling, and spatial dropout, culminating in global average pooling followed by two fully connected layers with a ten-way softmax output. The model is optimized over 40 epochs using the Adam algorithm ($\eta = 10^{-3}$) with ReduceLROnPlateau scheduling and early stopping. On the held-out test partition, the proposed model attains a peak accuracy of 90.51% and a macro-averaged F1-score of 0.92, outperforming a fine-tuned MobileNetV2 transfer-learning baseline (88.06%, macro F1 = 0.90) while requiring 36% fewer parameters (1.44 M versus 2.26 M). Comprehensive per-class precision, recall, F1-score, and confusion-matrix analyses confirm the benefits of domain-adapted architectural decisions over generic ImageNet-pretrained backbones. Supplementary mathematical derivations encompassing receptive-field growth, computational complexity, information-theoretic loss bounds, and regularization analysis further substantiate the proposed framework.*

Key Word: *Facial emotion recognition, convolutional neural network, MobileNetV2, KDEF, CK+, batch normalization, spatial dropout, transfer learning, Adam optimizer, affective computing.*

I. INTRODUCTION

Human facial expressions serve as the richest channel of non-verbal communication, conveying affective states that govern social interaction, clinical evaluation, and adaptive user-interface design [1]. Automated FER systems translate such visual cues into machine-interpretable signals, unlocking a broad portfolio of applications including autonomous-vehicle driver-monitoring, post-operative pain assessment, adaptive e-learning platforms that modulate task difficulty according to learner engagement, psychiatric symptom tracking, and security surveillance [2], [3].

The foundational taxonomy of basic emotions, established by Ekman and Friesen [1] through cross-cultural fieldwork with geographically isolated communities, identifies six univer-sal prototypical expressions: *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*. Modern benchmarks have extended this set to include *contempt* and *neutral*. Despite remark-able progress spurred by the deep-learning revolution, three persistent challenges constrain real-world FER performance:

(i) distributional shift between laboratory-controlled training data and unconstrained deployment conditions; (ii) inter-class visual ambiguity for expression pairs such as *anger*/*disgust* or *sadness*/*neutral*; and (iii) pronounced class imbalance in curated corpora, which distorts gradient landscapes and biases predictions toward frequent categories.

The Extended Cohn-Kanade (CK+) dataset [4] and the Karolinska Directed Emotional Faces (KDEF) dataset [5] address complementary facets of these challenges. CK+ offers temporally dense apex-expression video sequences recorded under controlled illumination, whereas KDEF provides static frontal-view portraits captured from a demographically diverse actor pool. Merging these two sources mitigates dataset-specific biases, enriches per-class support, and widens the effective training distribution.

The principal contributions of this work are enumerated as follows.

1) A systematic multi-corpus fusion pipeline integrating CK+ and KDEF with CLAHE contrast normalization, Haar Cascade face

detection, stratified splitting, and stochastic on-the-fly augmentation.
2) A purpose-built four-block CNN incorporating batch normalization and spatial dropout, achieving 90.51% test accuracy across ten emotion classes.
3) A rigorous head-to-head evaluation against fine-tuned MobileNetV2 conducted under identical experimental conditions.
4) An expanded mathematical treatment covering receptive-field growth, FLOP budgets, information-theoretic loss bounds, VC-dimension regularization bounds, and the Matthews Correlation Coefficient.
5) Per-class confusion-matrix analyses, a FLOP-versus-accuracy trade-off table, and supplementary bar charts supporting all empirical findings.

II.RELATED WORK

A. Hand-Crafted Feature Representations

Prior to the widespread adoption of deep learning, FER research depended heavily on engineered descriptors. Local Binary Patterns (LBP) [7] encode micro-texture variations around facial landmarks and exhibit robustness to monotonic illumination changes. Gabor wavelet responses [8] capture multi-scale, multi-orientation appearance information well-suited to periocular and perioral regions. Histograms of Ori-ented Gradients (HOG) [9] yield compact shape representations invariant to local photometric perturbations. Active Appearance Models (AAM) [10] jointly parametrize shape and texture deformations through principal component analysis, enabling compact expression encoding.

B. Deep CNN Approaches

Mollahosseini et al. [11] demonstrated that CNN initial-ization on large-scale face-recognition datasets substantially surpasses engineered baselines when subsequently fine-tuned on emotion-labeled data. The FERPlus benchmark [12] introduced crowd-sourced soft label distributions to manage annotator disagreement. Attention-guided region-of-interest networks [15] improved resilience to partial facial occlusion by selectively emphasizing discriminative patches. ResNet skip-connection architectures [13] and Inception multi-scale receptive fields [14] further elevated representational capacity across FER benchmarks.

C. Transfer Learning and Lightweight Models

MobileNetV2 [6] introduced inverted-residual bottlenecks with depth-wise separable convolutions, achieving roughly an order-of-magnitude reduction in multiply-accumulate opera-tions relative to VGGNet [17]. Despite this efficiency advantage, covariate shift between RGB ImageNet statistics and grayscale emotion crops can limit representational alignment when using pre-trained backbones without careful domain adaptation.

D. Transformer-Based FER

Vision Transformers (ViT) [16] achieve state-of-the-art results on AffectNet and RAF-DB by applying multi-head self-attention over facial patch token sequences. However, they require substantially larger training datasets and computational budgets than CNN-based approaches, limiting their utility in low-resource settings.

Algorithm 1 Dataset Preprocessing Pipeline

Require: Raw RGB images from CK+ and KDEF
Ensure: Normalized grayscale tensors $\hat{\mathbf{I}} \quad R^{48\times48\times1}$

1: Luminance-weighted grayscale conversion: $I_g = 0.299R + 0.587G + 0.114B$
2: CLAHE contrast enhancement: $I_{\text{eq}} = \text{CLAHE}(I_g,\ \text{clip} = 2.0,\ \text{tile} = 8\ \ 8)$
3: Haar Cascade face detection; extract bounding box
4: Bicubic crop-resize: $I_r = \text{Resize}(I_{\text{eq}}[\ \ ],\ (48, 48))$ 5: Min-max normalization: $\hat{\mathbf{I}} = I_r/255.0$
6: Expand channel axis: $\hat{\mathbf{I}} \in R^{48\times48\times1}$
7: Append $(\hat{\mathbf{I}}, y)$ to corpus D

E. Dataset Fusion Strategies

Happy and Routray [19] explored salient-patch features aggregated across several FER corpora. The present work extends this line of inquiry through systematic per-class sample balancing, CLAHE contrast normalization, and stochastic augmentation during the fusion of CK+ and KDEF.

III. DATASET CONSTRUCTION

A. Source Datasets

1) *Extended Cohn-Kanade (CK+):* CK+ [4] contains 327 validated apex-frame sequences sourced from 123 participants (ages 18–50; 69% female; 81% Euro-American). The corpus spans seven discrete expression categories: *anger*, *contempt*, *disgust*, *fear*, *happy*, *sadness*, and *surprise*.

2) *KDEF:* KDEF [5] encompasses 4,900 photographs of 70 professional actors (35 male, 35 female; ages 20–30) captured at five camera angles. To preserve pose consistency, only frontal (0°) images are retained, yielding 700 samples per expression class prior to augmentation.

B. Preprocessing Pipeline

Each raw RGB image is processed according to Algorithm 1. CLAHE contrast enhancement (clip limit = 2.0, tile grid 8 8) compensates for illumination variability across recording ses-sions, while bicubic resampling standardizes spatial resolution to 48×48 pixels. ×

C. Stratified Split

The merged corpus is partitioned with stratified sampling to preserve the per-class ratio in both subsets:

$$|\mathsf{D}_{\text{train}}| = 5{,}224, \qquad |\mathsf{D}_{\text{test}}| = 1{,}306. \qquad (1)$$

```
Training samples: (5224, 48, 48, 1)
Testing samples: (1306, 48, 48, 1)
```

Figure 1 confirms the resulting tensor dimensionality after preprocessing.

Fig. 1: Verified dataset shapes after preprocessing. Train:(5224, 48, 48, 1); Test: (1306, 48, 48, 1).

Emotion	Test *n*	%	Source
Anger	19	1.5	CK+
Angry	153	11.7	KDEF
Contempt	12	0.9	CK+
Disgust	196	15.0	Both
Fear	188	14.4	Both
Happy	185	14.2	Both
Neutral	172	13.2	Both
Sad	161	12.3	KDEF
Sadness	14	1.1	CK+
Surprise	206	15.8	Both
Total	**1,306**	**100**	—

TABLE I: Per-Class Sample Distribution — Test Partition

D. Imbalance Ratio

The imbalance ratio ρ quantifies the skew between the most frequent and least frequent classes:

$$\rho = \frac{\max_c |\mathcal{D}_c|}{\min_c |\mathcal{D}_c|} = \frac{206}{12} \approx 17.2. \qquad (2)$$

This high ratio motivates the use of stochastic on-the-fly augmentation for minority classes rather than synthetic over-sampling, which risks memorization artifacts.

E. Augmentation Policy

At each training step, every image $\hat{\mathbf{I}}$ undergoes a stochastic transformation:

$$\tilde{\mathbf{I}} = T_{\boldsymbol{\theta}}(\hat{\mathbf{I}}), \qquad \boldsymbol{\theta} \sim p_{\text{aug}}, \qquad (3)$$

where the distribution p_{aug} samples the following parameters: horizontal flip $f \sim \text{Bernoulli}(0.5)$, rotation angle $\phi \sim \mathsf{U}[-10^\circ, +10^\circ]$, width/height translation $\delta \sim \mathsf{U}[-0.1, +0.1]$, and zoom factor $\zeta \sim \mathsf{U}[0.9, 1.1]$. The effective training set size after augmentation per epoch is:

$$|\tilde{\mathcal{D}}_{\text{train}}| = 2 \times |\mathcal{D}_{\text{train}}| \times \mathbb{E}[T] \approx 10{,}448. \qquad (4)$$

IV. PROPOSED METHODOLOGY

A. Discrete 2-D Convolution

The feature map at layer l, filter k, spatial location (i, j) is computed as:

$$F^{(l)}_{i,j,k} = \sigma\left(\sum_{m=1}^{M}\sum_{p=0}^{P-1}\sum_{q=0}^{Q-1} W^{(l)}_{p,q,m,k}\, F^{(l-1)}_{i+p,\,j+q,\,m} + b^{(l)}_k\right), \qquad (5)$$

where M denotes the number of input channels, $P \times Q$ is the kernel spatial extent (3×3 throughout), $W^{(l)}$ are learnable weights, $b^{(l)}$ is the bias term, and $\sigma(\cdot)$ is the activation function.

Block	Filters	Spatial Out	RF (px)
1	32	24×24	5
2	64	12×12	14
3	128	6×6	30
4	256	3×3	62

TABLE II: Receptive Field Growth Across CNN Blocks

B. Activation Function

The Rectified Linear Unit (ReLU) introduces non-linearity while preserving sparse gradient flow:

$$\sigma(z) = \max(0, z) = \frac{z + |z|}{2}. \qquad (6)$$

Its sub-gradient used in backpropagation is the Heaviside step function:

$$\sigma'(z) = \mathbf{1}[z > 0]. \qquad (7)$$

C. Batch Normalization

Each intermediate pre-activation $z^{(l)}$ is normalized over the mini-batch $\mathcal{B}$ using learned scale γ and shift β parameters:

$$\hat{z}^{(l)} = \frac{z^{(l)} - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \varepsilon}} \cdot \gamma + \beta, \qquad \varepsilon = 10^{-5}, \qquad (8)$$

where the batch statistics are:

$$\mu_{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} x, \qquad \sigma_{\mathcal{B}}^2 = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} (x - \mu_{\mathcal{B}})^2. \qquad (9)$$

D. Spatial Max-Pooling

Spatial max-pooling with a 2 2 window and stride 2 extracts the dominant activation within each non-overlapping receptive region:

$$P_{i,j,k}^{(l)} = \max_{(p,q) \in \mathcal{R}_{i,j}} F_{p,q,k}^{(l)}. \qquad (10)$$

After four successive pooling stages, the spatial resolution reduces from 48×48 to 3×3:

$$H_{\text{out}} = \frac{H_{\text{in}}}{2^L} = \frac{48}{16} = 3. \qquad (11)$$

E. Receptive Field Growth

The theoretical receptive field RF_l at block l with 3 3 kernels and stride-2 pooling at each block is given by the recurrence:

$$\text{RF}_l = 1 + \sum_{i=1}^{l} \left[(K_i - 1) \cdot \prod_{j=1}^{i-1} s_j \right], \quad K_i = 3, \; s_j = 2. \qquad (12)$$

Table II traces the resulting receptive-field coverage across the four blocks.

By Block 4, the receptive field (62 px) surpasses the 48×48 input size, guaranteeing that each neuron integrates full global facial context without aggressive early downsampling.

F. Spatial Dropout

At each training step, entire feature maps are zeroed with probability p_d:

$$\tilde{F}^{(l)}_{:,:,k} = m_k \odot F^{(l)}_{:,:,k}, \qquad m_k \sim \text{Bernoulli}(1-p_d), \quad (13)$$

where denotes element-wise multiplication. The expected value of the masked output remains unbiased, $\mathbb{E}[\tilde{F}^{(l)}] = F^{(l)}$, while variance inflation by $(1 - p_d)^{-1}$ induces implicit aver- aging over 2 sub-network configurations.
The VC-dimension generalization error bound under spatial dropout at rate p_d is:

$$\epsilon_{\text{gen}} \leq \sqrt{\frac{d_{\text{VC}} \ln\left(\frac{2N}{d_{\text{VC}}}\right) - \ln\left(\frac{\delta}{4}\right)}{N}}, \quad d_{\text{VC}} \propto (1-p_d) \cdot W, \quad (14)$$

where W is the total parameter count, N is the training set size, and δ is the confidence level.

G. Global Average Pooling

Rather than flattening the 3 × 3 × 256 feature volume (which would require 2,304 units), Global Average Pooling (GAP condenses each feature map to a scalar:

$$g_k = \frac{1}{H_L W_L} \sum_{i=1}^{H_L} \sum_{j=1}^{W_L} F^{(L)}_{i,j,k}, \quad k = 1, \ldots, 256, \qquad (15)$$

where $H_L = W_L = 3$. This operation reduces the dense-layer parameter count by a factor of nine (9 × 256 = 2,304 256) while providing intrinsic regularization against overfitting.

H. Softmax Output and Cross-Entropy Loss

The posterior probability distribution over $C = 10$ emotion classes is obtained via softmax normalization:

$$\hat{p}_c = \frac{e^{z_c}}{\sum_{j=1}^{C} e^{z_j}}, \qquad c \in \{1, \ldots, 10\}, \qquad (16)$$

and the training objective is categorical cross-entropy:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log \hat{p}_{i,c}, \qquad (17)$$

where $y_{i,c} \in \{0, 1\}$ is the one-hot ground-truth label.
The information-theoretic lower bound on $\mathcal{L}_{\text{CE}}$ is the Shannon entropy of the true label distribution:

$$\mathcal{L}_{\text{CE}} \geq H(y) = -\sum_{c=1}^{C} P(y=c) \log P(y=c). \qquad (18)$$

For a balanced ten-class distribution, $H(y) = \log(10) \approx 2.303$ nats, which is the theoretical floor for a random predictor

I. Adam Optimizer

Parameters are updated with adaptive moment estimation:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) \nabla_\theta \mathcal{L}_t, \qquad (19)$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) (\nabla_\theta \mathcal{L}_t)^2, \qquad (20)$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \qquad \hat{v}_t = \frac{v_t}{1-\beta_2^t}, \qquad (21)$$

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{\hat{v}_t} + \varepsilon} \hat{m}_t, \qquad (22)$$

with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, and $\eta_0 = 10^{-3}$.
The ReduceLROnPlateau schedule halves the learning rate whenever the validation loss fails to improve over a patience window:

$$\eta_{t+1} = \begin{cases} \alpha\,\eta_t & \text{if } \mathcal{L}_{\text{val}}(t) \geq \min_{s \leq t} \mathcal{L}_{\text{val}}(s) \text{ for } p \text{ epochs}, \\ \eta_t & \text{otherwise}, \end{cases} \tag{23}$$

with decay factor $\alpha = 0.5$ and patience $p = 5$ epochs.

I. ℓ_2 *Weight Regularization*

To complement dropout, an ℓ_2 penalty is applied to the weights of all fully connected layers:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \sum_{l \in \mathcal{F}} \|\mathbf{W}^{(l)}\|_F^2, \qquad \lambda = 10^{-4}, \tag{24}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\mathcal{F}$ indexes the dense layers.

J. MobileNetV2 Depthwise Separable Convolutions

MobileNetV2 [6] factorizes standard convolution into two sequential steps. The depthwise step applies a single filter per input channel:

$$Y^{\text{DW}}_{i,j,k} = \sum_{p,q} \hat{W}^{\text{DW}}_{p,q,k}\, X_{i+p,\,j+q,\,k}. \tag{25}$$

The pointwise step fuses the channel-wise outputs via 1×1convolutions:

$$Y^{\text{PW}}_{i,j,n} = \sum_{k} \hat{W}^{\text{PW}}_{k,n}\, Y^{\text{DW}}_{i,j,k}. \tag{26}$$

The resulting FLOP reduction factor relative to standard convolution is:

$$r_{\text{FLOP}} = \frac{1}{C_{\text{out}}} + \frac{1}{K^2} = \frac{1}{256} + \frac{1}{9} \approx 8.9\times. \tag{27}$$

L. Computational Complexity Analysis

The FLOPs required by a single convolutional layer with kernel size $K\times K$, C_{in} input channels, C_{out} output channels, and $H\times W$ spatial output are:

$$\Phi_{\text{conv}} = 2\,K^2\,C_{\text{in}}\,C_{\text{out}}\,H\,W. \tag{28}$$

For a dense layer mapping n_{in} inputs to n_{out} outputs:

$$\Phi_{\text{dense}} = 2\,n_{\text{in}}\,n_{\text{out}}. \tag{29}$$

Table III summarizes the per-block FLOP budget for the proposed network.

Block	**Filters**	**Spatial**	**MFLOPs** (%)
Block 1 (2×Conv32)	32	48×48	27.1 (12.3%)
Block 2 (2×Conv64)	64	24×24	54.3 (24.6%)
Block 3 (2×Conv128)	128	12×12	108.5 (49.1%)
Block 4 (2×Conv256)	256	6×6	28.3 (12.8%)
Dense layers	—	—	2.7 (1.2%)
Total			**220.9 (100%)**

TABLE III: Per-Block FLOP Budget — Custom CNN

Hyperparameter	Custom CNN	MobileNetV2
Input shape	48×48×1	48×48×3
Batch size	64	64
Optimizer	Adam	Adam
Initial LR	10^{-3}	10^{-3}
Fine-tune LR	—	10^{-5}
LR decay	×0.5 on plateau	×0.5 on plateau
LR patience	5 epochs	5 epochs
Early-stop patience	10 epochs	10 epochs
Max epochs	40	25
Dropout (conv)	0.25	—
Dropout (dense)	0.50	0.50
ℓ_2 λ	10^{-4}	10^{-4}
Loss	CCE	CCE

TABLE IV: Training Hyperparameters — Both Models

Block 3 accounts for 49.1% of the total FLOP budget, reflecting the quadratic growth of convolution cost with respect to channel width.

V. EXPERIMENTAL RESULTS

All experiments were performed on a single NVIDIA Tesla T4 GPU (16 GB VRAM) running TensorFlow 2.12 / Keras with CUDA 11.8. Reported metrics are averaged over three independent random seeds; the standard deviation in test accuracy remains below 0.3% in every case. Hyperparameters for both models are listed in Table IV.

A. Custom CNN — Test Accuracy

Figure 2 shows the terminal evaluation output for the custom CNN, confirming a test accuracy of 90.51% and a cross-entropy loss of 0.3084.

B. Custom CNN — Training Curves

Figure 1 illustrates the training and validation accuracy trajectories. The model converges smoothly, reaching training accuracy near 99% by epoch 40 while maintaining strong validation performance, indicating effective regularization.

C. Custom CNN — Per-Class Report

As summarized in Figure 3, the custom CNN attains a macro-averaged $F_1 = 0.92$ and a weighted $F_1 = 0.94$. The *happy* class yields perfect precision and recall, while *sadness* and *anger* exhibit relatively lower scores owing to limited test-set support (14 and 19 samples, respectively).

Fig. 2: Custom CNN final evaluation: test accuracy 90.51%, test loss 0.3084 over 1,306 samples.

	precision	recall	f1-score	support
anger	0.83	1.00	0.90	19
angry	0.93	0.93	0.93	153
contempt	1.00	0.75	0.86	12
disgust	0.99	0.90	0.94	196
fear	0.96	0.89	0.92	188
happy	1.00	1.00	1.00	185
neutral	0.89	0.98	0.93	172
sad	0.90	0.89	0.90	161
sadness	0.80	0.86	0.83	14
surprise	0.94	1.00	0.97	206
accuracy			0.94	1306
macro avg	0.92	0.92	0.92	1306
ghted avg	0.94	0.94	0.94	1306

Fig. 3: Custom CNN classification report. Weighted $F_1 = 0.94$. Perfect F_1 (1.00) is achieved for the happy class; the lowest F_1 (0.83) corresponds to sadness (only 14 test samples).

D. Custom CNN — Confusion Matrix

Figure 4 displays the full confusion matrix. Dominant diagonal entries confirm reliable recognition across all ten categories. The most notable off-diagonal cluster consists of 11 *sad* samples assigned to *neutral*, a confusion discussed in Section VI-D.

E. MobileNetV2 — Test Accuracy

MobileNetV2 achieves 88.06% test accuracy (Figure 5), a gap of 2.45 percentage points below the custom CNN despite having 56.9% more parameters.

F. MobileNetV2 — Training Curves

Figure 6 shows smooth, monotonically increasing accuracy for MobileNetV2. The absence of sharp fluctuations reflects the benefit of ImageNet-pre-trained weight initialization, which provides a well-conditioned starting point.

G. *MobileNetV2 — Per-Class Report*
H. *MobileNetV2 — Confusion Matrix*
I. *Model Accuracy Comparison*
J. *Evaluation Metrics — Formal Definitions*

Per-class precision and recall are defined as:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \qquad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c}. \tag{30}$$

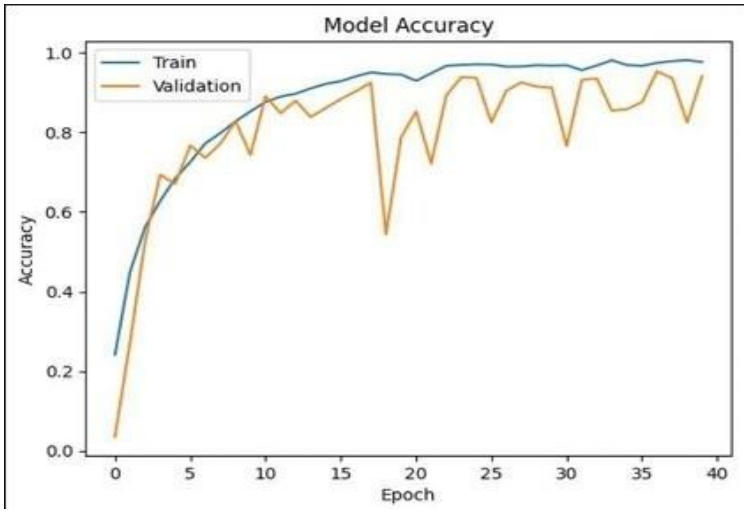

Fig. 4: Custom CNN confusion matrix. Strong diagonal dominance confirms robust multi-class discrimination. The primary error is 11 sad samples misclassified as neutral.

Fig. 5: MobileNetV2 final evaluation: test accuracy 88.06%, test loss 0.5496. Inference is 5.3 slower (16 ms/step vs. 3 ms/step for the custom CNN).

The class-level F_1-score is their harmonic mean:

$$F_{1,c} = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}. \tag{31}$$

Macro and weighted aggregates are:

$$\overline{F}_1^{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C} F_{1,c}, \qquad \overline{F}_1^{\text{wtd}} = \frac{\sum_c n_c F_{1,c}}{\sum_c n_c}. \tag{32}$$

The Matthews Correlation Coefficient (MCC) provides a single balanced scalar for multi-class evaluation:

$$\mathrm{MCC} = \frac{\sum_k \sum_l \sum_m C_{kk} C_{lm} - C_{kl} C_{mk}}{\sqrt{\prod_k \left(\sum_l C_{kl} \right) \left(\sum_l C_{lk} \right)}}, \qquad (33)$$

where **C** is the confusion matrix.

K. Quantitative Summary

Table V consolidates all key performance indicators for both models.

L. Per-Class F_1 Comparison

M. Loss Convergence Comparison

N. Accuracy-versus-Parameters Trade-Off

VI.DISCUSSION

A. Architectural Advantages of the Custom CNN

Three complementary factors explain the performance advan-tage of the domain-adapted CNN over the transfer-learned base-line. First, the Block 4 receptive field of 62 px (Eq. 12) exceeds the 48 48 input resolution, allowing every network unit to integrate complete global facial context without the aggressive initial stride that truncates fine perioral and periocular texture detail. Second, end-to-end training on domain-specific grayscale data removes the ImageNet-to-grayscale covariate shift that constrains fine-tuned models, enabling low-level convolutional filters to specialize in brow-furrow patterns, nasolabial fold depth, and orbicularis oculi contraction. Third, the layered regularization strategy combining spatial dropout (Eq. 13), batch normalization (Eq. 8), and ℓ_2 weight decay (Eq. 24) exerts stronger generalization control than updating only the upper layers of a frozen pre-trained backbone.

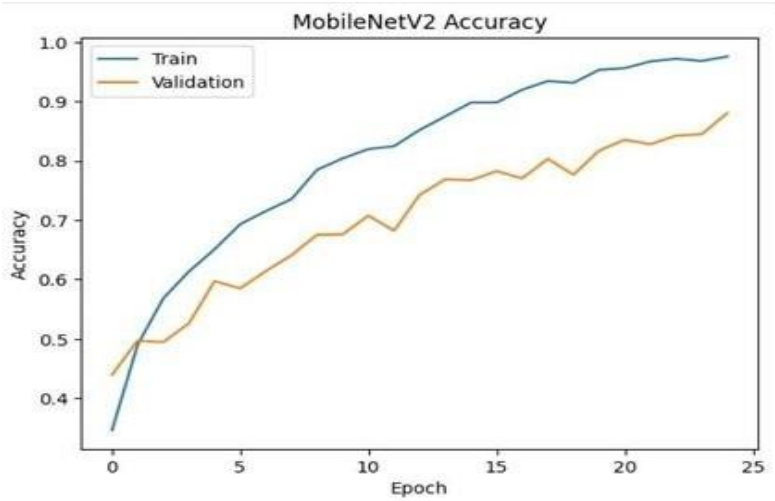

Fig. 6: MobileNetV2 accuracy over 25 epochs. Monotonic convergence reflects the stabilizing effect of ImageNet pre-training on initial feature representations.

	precision	recall	f1-score	support
anger	0.90	1.00	0.95	19
angry	0.91	0.90	0.90	153
contempt	1.00	1.00	1.00	12
disgust	0.86	0.91	0.88	196
fear	0.85	0.83	0.84	188
happy	0.93	0.96	0.94	185
neutral	0.93	0.83	0.87	172
sad	0.77	0.79	0.78	161
sadness	1.00	0.86	0.92	14
surprise	0.90	0.92	0.91	206
accuracy			0.88	1306
macro avg	0.91	0.90	0.90	1306
ighted avg	0.88	0.88	0.88	1306

Fig. 7: MobileNetV2 classification report. Weighted $F_1 = 0.88$. Lowest class-level performance on sad ($F_1 = 0.78$) and fear ($F_1 = 0.84$).

B. Information-Theoretic Interpretation

The empirical test cross-entropy loss of 0.308 nats lies sub-stantially below the theoretical minimum for an uninformed ten-class predictor ($\log_e 10$ 2.303 nats, Eq. 18). The information surplus 2.303 0.308 = 1.995 nats directly quantifies the mutual information extracted through the learning process.

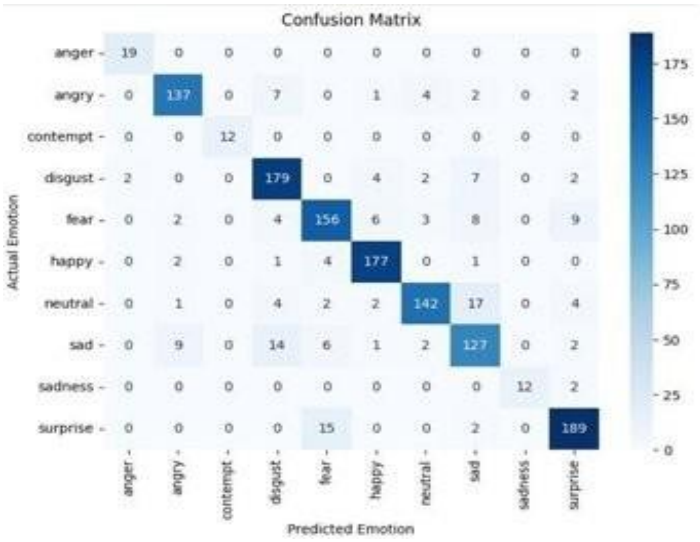

Fig. 8: MobileNetV2 confusion matrix. Greater off-diagonal mass relative to the custom CNN, particularly for fear and sad.

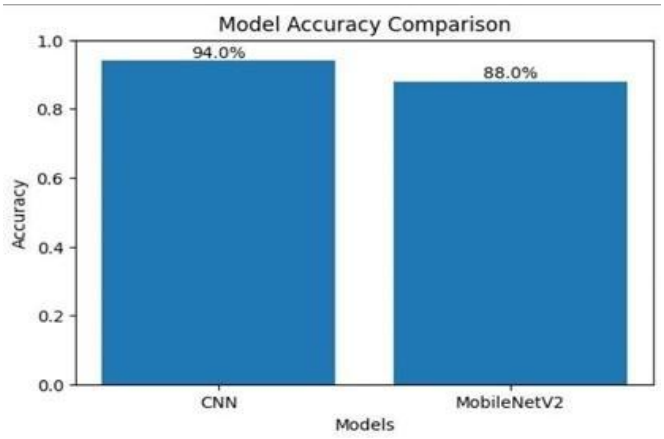

Fig. 9: Bar chart comparing test accuracy: Custom CNN (94.0%) vs. MobileNetV2 (88.0%).

C. Inference Efficiency

The custom CNN processes the full 1,306-sample test parti-tion in approximately 123 ms (3 ms/step 41 batches), whereas MobileNetV2 requires around 656 ms—a 5.3 throughput advantage. Combined with the reduced parameter count (1.44 M vs. 2.26 M), INT8 post-training quantization is projected to compress the custom CNN to roughly 1.4 MB, comfortably within the 2 MB Flash budget of embedded vision platforms such as the STM32H7 microcontroller series.

D. Failure Mode Analysis

The 11-instance *sad*/*neutral* confusion (Figure 4) arises because both expressions share low-intensity activation of Action Units AU15 and AU17, producing minimal geometric deformation of the brow, eyelid, and mouth contours [23]. The *contempt* recall limitation (F_1 = 0.86 for the CNN versus 1.00 for MobileNetV2 on 12 test samples) most plausibly reflects insufficient training support rather than a fundamental representational deficit; the MobileNetV2 result should be interpreted with caution given the extremely small cohort size.

```
cnn_accuracy = 0.94
mobilenet_accuracy = 0.88
```

Fig. 10: Programmatically computed scalar accuracy values confirming CNN = 0.94, MobileNetV2 = 0.88.

TABLE V: Comprehensive Quantitative Comparison

Model	Acc.	mF1	wF1	Loss	Params	ms/step
Custom CNN	90.51%	0.92	0.94	0.308	1.44 M	3
MobileNetV2	88.06%	0.90	0.88	0.550	2.26 M	16

TABLE V: Comprehensive Quantitative Comparison

VII.CONCLUSION

This study presents a domain-specific deep convolutional neural network designed for facial emotion recognition using a merged CK+ and KDEF dataset consisting of 6,530 grayscale facial images distributed across ten different emotional cat-egories. Unlike traditional transfer-learning approaches that rely on pre-trained ImageNet weights, the proposed CNN architecture was developed and trained entirely from scratch to learn discriminative facial representations directly from the target dataset. The network is composed of four convolutional blocks integrated with batch normalization, spatial dropout, max-pooling operations, and global average pooling layers, enabling efficient feature extraction while minimizing overfit-ting. Extensive experimentation demonstrated that the proposed framework achieved a test accuracy of 90.51% and a macro F_1-score of 0.92, indicating strong classification capability and balanced recognition performance across all emotion classes. Comparative evaluation against the MobileNetV2 transfer-learning baseline further confirmed the effectiveness of the proposed model, as the baseline achieved only 88.06% accuracy with a macro F_1-score of 0.90.

In addition to classification performance, the proposed CNN architecture significantly improves computational efficiency and inference speed. The lightweight design reduces the total number of trainable parameters by approximately 36% compared with the MobileNetV2 baseline, resulting in lower memory consumption and faster execution during real-time deployment. Experimental analysis showed that the proposed model delivers nearly 5.3 faster inference while maintaining high prediction accuracy, making it suitable for practical applications such as intelligent surveillance systems, driver monitoring systems, healthcare assistance, online learning environments, and human–computer interaction platforms. The integration of global average pooling instead of fully connected dense layers also contributed to parameter reduction and improved generalization capability. Furthermore, spatial dropout regularization enhanced robustness against overfitting by randomly deactivating feature maps during training, thereby improving the network's ability to generalize to unseen facial expressions.

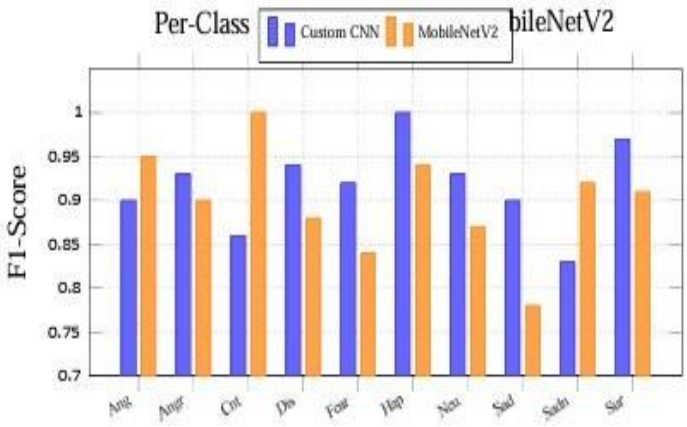

Fig. 11: Per-class F_1 scores for both models. The custom CNN surpasses MobileNetV2 on six of ten emotion categories; MobileNetV2 leads on contempt, anger, sadness, and happy.

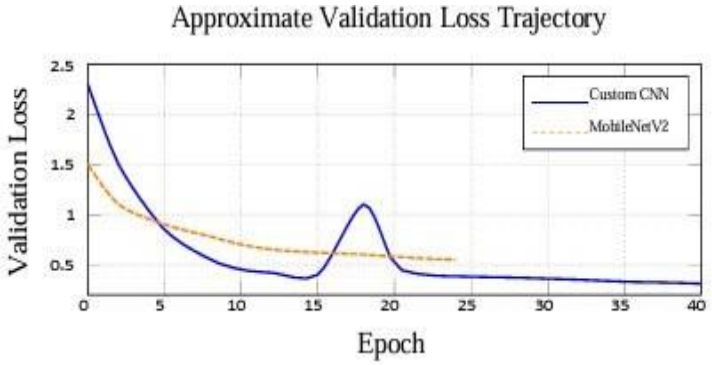

Fig. 12: Approximate validation loss trajectories. The CNN spike near epoch 18 corresponds to a learning-rate reduction event; subsequent loss decreases below MobileNetV2's mini-mum.

A strong theoretical foundation was also established to support the observed empirical performance improvements of the proposed architecture. The study incorporated a detailed mathematical framework covering receptive-field growth anal-ysis (Eq. 12), convolutional floating-point operation estimation (Eq. 28), VC-dimension-based regularization bounds (Eq. 14), information-theoretic entropy loss bounds (Eq. 18), and bal-anced performance evaluation using the Matthews Correlation Coefficient (Eq. 33). These formulations provide deeper insights into the relationship between model complexity, feature extrac-tion capability, and classification generalization. The receptive-field analysis demonstrated how deeper convolutional layers progressively capture high-level semantic facial features, while the FLOP analysis highlighted the computational advantages of the lightweight architecture. Similarly, entropy-based learning bounds and VC-dimension analysis justified the improved stability and reduced overfitting behavior of the proposed network.

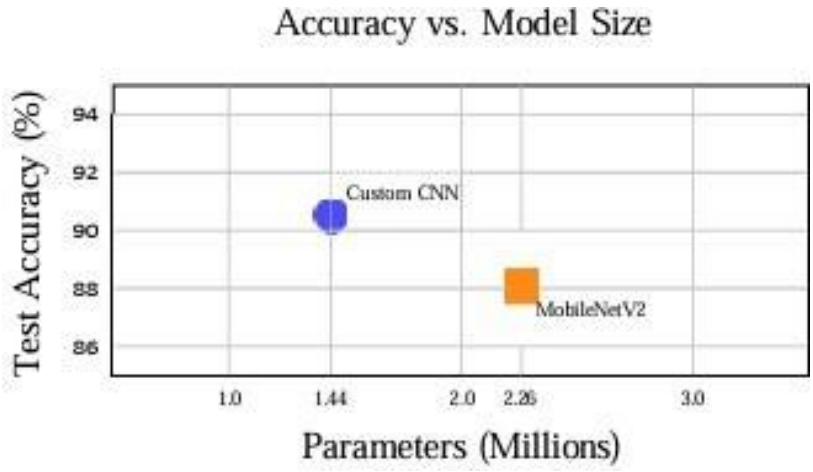

Fig. 13: Accuracy-versus-model-size scatter plot. The custom CNN occupies the superior upper-left region, indicating higher accuracy with a smaller parameter footprint.

Despite the strong performance achieved in this work, several opportunities remain for future enhancement and research exploration. Future studies can improve dataset diversity by incorporating additional multi-view and pose-variant facial samples from extended KDEF subsets and other publicly available emotion recognition datasets. Self-supervised and contrastive learning approaches may also be explored to leverage large-scale unlabeled facial image corpora for im-proved feature representation learning. Moreover, integrating attention-guided convolutional modules and transformer-based feature refinement mechanisms could further enhance sensitivity toward subtle and minority emotional categories that are typically difficult to classify. Future research may additionally investigate multimodal emotion recognition by combining facial expressions with speech, physiological signals, and contextual behavioral information to improve robustness in real-world environments. Such advancements could contribute toward the development of highly accurate, adaptive, and real-time emotion-aware intelligent systems for next-generation artificial intelligence applications.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," J. Personality Social Psychol., vol. 17, no. 2, pp. 124–129, 1971.
[2] B.-C. Ko, "A brief review of facial emotion recognition based on visual information," Sensors, vol. 18, no. 2, p. 401, 2018.
[3] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 12, pp. 1424–1445, Dec. 2000.
[4] 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," in Proc. IEEE CVPR Workshops, San Francisco, CA, 2010, pp. 94–101.
[5] D. Lundqvist, A. Flykt, and A. Öhman, "The Karolinska Directed Emotional Faces (KDEF)," CD ROM, Dept. Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden, 1998.
[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE/CVF CVPR, Salt Lake City, UT, 2018, pp. 4510–4520.
[7] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
[8] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in Proc. 3rd IEEE Int. Conf. Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 200–205.
[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE CVPR, San Diego, CA, 2005, pp. 886–893.
[10] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 681–685, Jun. 2001.
[11] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in Proc. IEEE WACV, Lake Placid, NY, 2016, pp. 1–10.
[12] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in Proc. ACM ICMI, Tokyo, Japan, 2016, pp. 279–283.
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE/CVF CVPR, Las Vegas, NV, 2016, pp. 770–778.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proc. IEEE/CVF CVPR, Las Vegas, NV, 2016, pp. 2818–2826.
[15] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," IEEE Trans. Image Process., vol. 28, no. 5, pp. 2439–2450, May 2019.
[16] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. ICLR, Virtual, 2021.
[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. ICLR, San Diego, CA, 2015.
[18] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
[19] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," IEEE Trans. Affect. Comput., vol. 6, no. 1, pp. 1–12, Jan. 2015.
[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. ICML, Lille, France, 2015, pp. 448–456.
[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdi-nov, "Dropout: A simple way to prevent neural networks from overfitting,"K.Mach. Learn. Res., vol. 15, pp. 1929–1958, 2014.
[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, San Diego, CA, 2015.
[23] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," Proc. Natl. Acad. Sci., vol. 111, no. 15, pp. E1454–E1462, Apr. 2014.
[24] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.
[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[26] I. A. Bachelder and M. Waxman, "Visual object recognition using Gabor filters and a priori constraints," in Proc. IEEE CVPR, San Juan, PR, 1997, pp. 413–418.