Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Improved EfficientNet-B0 Model for Facial Emotion Recognition

Nnaemeka Ugwumba

nnaemeka_ugwumba@uniport.edu.ng

University of Port Harcourt https://orcid.org/0009-0000-2493-9846

Peter Sunday Jaja

peter_jaja@uniport.edu.ng

University of Port Harcourt, Port Harcourt, Nigeria

Ugochi A. Okengwu

ugochi.okengwu@uniport.edu.ng

University of Port Harcourt, Port Harcourt, Nigeria

Promise Enyindah

promise.enyindah@uniport.edu.ng

University of Port Harcourt, Port Harcourt, Nigeria

Research Article

Keywords: facial emotion recognition, EfficientNet-B0, EfficientNet-B1, EfficientNet-B2, transfer learning, FER2013, RAF-DB, McNemar's test, web deployment

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11679881/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

An Improved EfficientNet-B0 Model for Facial Emotion Recognition

Peter Sunday Jaja[1] · Ugochi A. Okengwu[1] · Promise Enyindah[1] · Nnaemeka Kingsley Ugwumba[1]

[1]Department of Computer Science, University of Port Harcourt, Port Harcourt, Nigeria

Corresponding author: Peter Sunday Jaja, peter_jaja@uniport.edu.ng

ugochi.okengwu@uniport.edu.ng · promise.enyindah@uniport.edu.ng · nnaemeka_ugwumba@uniport.edu.ng

Abstract

Facial emotion recognition (FER) is increasingly built into interactive software, but reported accuracies are hard to compare because datasets, emotion sets and evaluation splits differ between studies, and few studies examine whether a larger backbone is worth its extra computational cost on the same data. This study develops and evaluates an EfficientNet-B0 model for five-class FER (happiness, sadness, anger, surprise and neutral) and compares it with EfficientNet-B1 and EfficientNet-B2 trained on the same images. FER2013 and the Real-world Affective Faces Database (RAF-DB) were combined, 1,470 duplicate or label-conflicting images and 19 unreadable files were removed, and a balanced set of 25,000 grayscale images (5,000 per class) was divided into 20,000 training, 2,500 validation and 2,500 test images, with no image shared between subsets. The model was fine-tuned in two stages from ImageNet weights using label smoothing, geometric and photometric augmentation, Mixup and CutMix. On the independent test set, EfficientNet-B0 achieved an accuracy of 78.68%, a macro F1-score of 78.83% and a macro-averaged ROC-AUC of 0.948, with class F1-scores ranging from 70.13% (neutral) to 87.78% (surprise). EfficientNet-B1 reached 79.76% accuracy, a difference that McNemar's test did not find significant ($p = 0.107$), while EfficientNet-B2, trained under a reduced schedule, reached 75.44% and was significantly less accurate ($p < 0.001$). EfficientNet-B0 used 36.6% fewer parameters than B1 (4.38 million against 6.90 million), needed 8.5 ms per image for single-image inference against 12.4 ms, required 40% less fine-tuning time and recorded the highest macro AUC of the three backbones. The trained model was deployed in a FastAPI web application that returns the predicted emotion with a confidence score. Within the conditions tested, EfficientNet-B0 gave the best balance between recognition performance and computational cost, while sadness and neutral remained the expressions most often confused with each other.

Keywords: facial emotion recognition; EfficientNet-B0; EfficientNet-B1; EfficientNet-B2; transfer learning; FER2013; RAF-DB; McNemar's test; web deployment

1 Introduction

Facial emotion recognition (FER) is the automatic identification of emotional states such as happiness, sadness, anger, fear, surprise, disgust and neutrality from facial images or video (Li & Deng, 2022). It supports applications in human-computer interaction, adaptive learning, driver monitoring, healthcare support and behavioural research (Siddiqui & Javaid, 2020). Early systems described the face with handcrafted descriptors such as Local Binary Patterns and classified them with Support Vector Machines (Shan et al., 2009). These descriptors worked well on posed laboratory images but degraded under changes in illumination, head pose and occlusion, and convolutional neural networks (CNNs), which learn features directly from pixels, have since become the standard approach.

Deeper CNNs improved accuracy once residual connections (He et al., 2016) and dense connectivity (Huang et al., 2017) made very deep networks trainable, but at the cost of more parameters and computation. For real-time use this cost matters, and two families of efficient architectures have become common. MobileNet replaces standard convolutions with depthwise separable convolutions (Howard et al., 2017; Sandler et al., 2018). EfficientNet, obtained through neural architecture search, scales network depth, width and input resolution together using a single compound coefficient (Tan & Le, 2019). Its smallest member, EfficientNet-B0, has about 5.3 million parameters with its ImageNet classifier, and the larger members B1 to B7 are scaled versions of it.

The baseline for this study is the real-time MobileNet-based FER system of Sowmiya et al. (2026). That system was trained on FER2013 and CK+ with a supplementary image set, classifies seven emotions, reports a test accuracy of 70.7% on its native FER2013 test samples, and processes about 30 to 33 frames per second on an NVIDIA GPU with an inference time of about 18 ms per frame. Its recognition was uneven across emotions: happiness and surprise reached 97% and 95%, while fear and disgust reached 82% and 80%. Occlusion reduced its accuracy by about 7% to 8% and moderate lighting changes by about 3%, and the authors recommended evaluating other datasets and architectures in future work.

This study follows that recommendation. It develops an EfficientNet-B0 model trained on a combination of FER2013 (Goodfellow et al., 2013) and RAF-DB (Li et al., 2017) for five emotions, evaluates it on a cleaned,

balanced and leakage-free test set, and asks a question that most single-model studies leave open: whether the larger EfficientNet-B1 and EfficientNet-B2 backbones, trained on exactly the same split, give enough additional accuracy to justify their cost. The paper makes four contributions: (i) a reproducible data preparation procedure that removes duplicate and label-conflicting images and balances the five classes; (ii) a two-stage EfficientNet-B0 model evaluated with class-level metrics, ROC analysis and accuracy by source dataset; (iii) a controlled comparison of EfficientNet-B0, B1 and B2 on the same 2,500 test images, including McNemar's significance test and inference cost; and (iv) a working web application built around the trained model. Section 2 reviews related work, Section 3 describes the materials and methods, Section 4 presents the results, Section 5 discusses them, and Section 6 concludes the paper.

2 Related Work

2.1 From Handcrafted Features to Deep CNNs

The theory of basic emotions holds that a small set of emotions are associated with recognisable configurations of facial muscle movement (Ekman, 1993), and the Facial Action Coding System described these configurations as discrete action units. Early automated FER followed this structure by extracting geometric or texture descriptors and passing them to statistical classifiers. Shan et al. (2009) showed that Local Binary Pattern features could recognise expressions effectively on posed datasets, but such features cannot adapt to the full variability of faces captured in uncontrolled settings. CNNs removed the need to design features by hand: lower layers learn edges and textures and deeper layers learn patterns associated with particular expressions. Li and Deng (2022) review this transition and identify limited training data, label noise and variation in pose and illumination as the main obstacles that remain for deep FER. On FER2013 alone, a carefully tuned single VGG-style network has reached 73.28% accuracy on the seven-class task (Khaireddin & Chen, 2021), which indicates the difficulty of the benchmark.

2.2 Efficient Architectures for Real-Time FER

Real-time FER requires models that are accurate and inexpensive to run. MobileNet factorises a standard convolution into a per-channel spatial filter followed by a pointwise combination, which greatly reduces the number of multiply-accumulate operations (Howard et al., 2017), and MobileNetV2 added inverted residual blocks with linear bottlenecks (Sandler et al., 2018). Haq et al. (2024) applied MobileNet to FER2013 supplemented with 350 additional images across seven classes and reported a training accuracy of 97.9%, but did not report accuracy on a separate held-out test set. EfficientNet builds on the inverted residual (MBConv) block, adds squeeze-and-excitation channel attention (Hu et al., 2018), and scales the resulting baseline network through compound scaling (Tan & Le, 2019). Lightweight EfficientNet models have performed close to the state of the art for facial expression recognition on AffectNet (Savchenko, 2021), and an EfficientNet-B2 model trained on FER2013 at 260 × 260 pixels recently reported a seven-class test accuracy of 68.78% (Naik et al., 2026). Chowdary et al. (2023) compared several ImageNet-pretrained networks for FER and found that transfer learning consistently improved accuracy over training from scratch. These studies usually evaluate one backbone at a time, so the trade-off between the smaller and larger members of the EfficientNet family on the same FER data is rarely measured directly.

2.3 Datasets, Data Quality and Bias

FER2013 contains 35,887 grayscale images of 48 × 48 pixels in seven categories, collected from web searches (Goodfellow et al., 2013). It is unbalanced, contains label noise, and includes exact duplicate images. RAF-DB contains about 15,000 images with basic-emotion labels, crowd-annotated and captured under varied real-world conditions of age, ethnicity, pose and illumination (Li et al., 2017). Larger corpora such as AffectNet (Mollahosseini et al., 2019) show that scale alone does not resolve inconsistent annotation or limited demographic diversity. Duplicate images matter in particular because a duplicate shared between training and test data inflates measured accuracy, and class imbalance biases a classifier toward the most frequent emotions. Dewi et al. (2026) found that most disparities in data-driven classifiers originate in the data rather than the model, and that mitigation applied at the data stage, before training, gives more stable outcomes than adjustments made afterwards. Cleaning, de-duplicating and balancing the training data, as done in this study, is a data-stage intervention of this kind, although it does not replace a formal fairness audit.

2.4 Regularisation for Small and Noisy Datasets

FER datasets are small compared with general image benchmarks, so regularisation is important when fine-tuning large pretrained networks. Label smoothing replaces one-hot targets with slightly softened targets and reduces overconfident predictions (Szegedy et al., 2016). Mixup trains on convex combinations of pairs of images and their labels (Zhang et al., 2018), and CutMix pastes a rectangular patch from one image into another and mixes the labels in proportion to the patch area (Yun et al., 2019). Decoupled weight decay, as implemented in AdamW, regularises the weights independently of the adaptive gradient update (Loshchilov & Hutter, 2019). This study combines all four techniques with conventional geometric and photometric augmentation.

2.5 Multimodal and Temporal Extensions

Several studies combine the face with other modalities. Zadeh et al. (2017) proposed the Tensor Fusion Network to model interactions between language, visual and acoustic features, and Mittal et al. (2020) re-weighted

modalities according to their reliability. Recurrent architectures combined with convolutional features capture how expressions change over time in video (Tzirakis et al., 2017). These extensions can improve recognition but require synchronised multimodal data and more computation. This study keeps to static facial images, the input most consistently available in a web application.

2.6 Research Gaps

Three gaps motivate this study. First, the baseline system of Sowmiya et al. (2026) recognises subtle emotions less reliably than distinct ones and relies principally on FER2013, so its evaluation reflects a limited range of faces and capture conditions. Second, many FER studies do not report whether duplicate images were removed or whether the test images are independent of those used for training and model selection, which makes reported accuracies difficult to interpret. Third, studies that adopt EfficientNet for FER rarely compare the smallest backbone with its larger variants on the same split, so the benefit of scaling for this task is not well established. This study addresses these gaps with a de-duplicated, balanced combination of FER2013 and RAF-DB, a strictly independent test set, and a controlled comparison of EfficientNet-B0, B1 and B2.

3 Materials and Methods

3.1 Research Design

The study adopts a quantitative experimental design within a Design Science Research framework, with the Cross-Industry Standard Process for Data Mining (CRISP-DM) as the operational process for the machine learning component. Design Science Research structures the identification of the problem, the design and development of the artefact, an EfficientNet-B0 FER system integrated into a web application, and its evaluation against the existing MobileNet-based system. CRISP-DM structures the data understanding, data preparation, modelling, evaluation and deployment stages. All experiments were run in a Kaggle notebook with GPU acceleration, using TensorFlow 2.20 with mixed precision and a fixed random seed of 42.

3.2 Data Sources and Preparation

Images from FER2013 and RAF-DB were collected for the five target emotions. Fear and disgust were excluded from both datasets, and the numbered RAF-DB folders were mapped to their emotion names. The five classes contained 44,325 images, 30,219 from FER2013 and 14,106 from RAF-DB. Every file was hashed with MD5, and 1,470 exact duplicates and label-conflicting copies were removed; among them, 22 images appeared under more than one emotion label and were discarded entirely. A further 19 files that could not be decoded were removed, leaving 42,836 usable images (Table 1).

Table 1. Images per emotion class collected from each source and usable after cleaning.

Class	FER2013	RAF-DB	Collected	Usable
Happiness	8,989	5,957	14,946	14,740
Sadness	6,077	2,460	8,537	8,380
Anger	4,953	867	5,820	5,592
Surprise	4,002	1,619	5,621	4,878
Neutral	6,198	3,203	9,401	9,246
Total	**30,219**	**14,106**	**44,325**	**42,836**

Usable images remain after removing 1,470 duplicate or label-conflicting images and 19 unreadable files. Each class then contributed 4,000 training, 500 validation and 500 test images.

A balanced dataset of 5,000 images per class (25,000 in total) was then drawn with a fixed seed and split into 4,000 training, 500 validation and 500 test images per class, giving 20,000, 2,500 and 2,500 images, or 80%, 10% and 10%. The validation and test sets contain unique images only. The surprise class had 3,878 unique images available for training after the validation and test images had been set aside, so 122 training images were repeated to reach 4,000; augmentation makes each repeated copy appear differently to the model. An explicit check confirmed that no image appears in more than one subset. The test set contains 1,667 FER2013 images and 833 RAF-DB images.

3.3 Preprocessing and Augmentation

All images were converted to grayscale. FER2013 is natively grayscale, and leaving the RAF-DB images in colour would have allowed the model to identify the source dataset from colour rather than learning facial expression. Images were resized to 224 × 224 pixels with bilinear interpolation and the single channel was replicated to three channels to match the ImageNet-pretrained input. Pixel values were kept in the 0 to 255 range, because EfficientNet in Keras performs its own input normalisation. Labels were one-hot encoded and label smoothing of 0.1 was applied in the categorical cross-entropy loss.

During training only, each batch was augmented with horizontal flipping, rotation of up to about ±14°, zoom of up to 12%, translation of up to 8% in each direction, and random contrast (25%) and brightness (15%) changes. In addition, half of the training batches were mixed: a quarter with Mixup ($\alpha = 0.2$) and a quarter with CutMix ($\alpha = 1.0$). Validation and test images were not augmented. At test time, the prediction for each image was the average of the softmax outputs for the image and its horizontal mirror (flip test-time augmentation).

3.4 Proposed System Architecture

Fig. 1 shows the architecture of the proposed system. Images from FER2013 and RAF-DB are processed and labelled into the five emotion classes, passed to the EfficientNet-B0 model, and the resulting emotion prediction is produced for images supplied either through

a live webcam or as uploaded images. The classifier uses the EfficientNet-B0 convolutional base, pretrained on ImageNet (Deng et al., 2009), followed by global average pooling, dropout of 0.3, a 256-unit dense layer with swish activation and L2 regularisation (10^{-4}), a second dropout of 0.3, and a five-unit softmax output. The model has 4,378,792 parameters, of which 4,294,753 are trainable during fine-tuning because the batch normalisation layers are kept frozen.

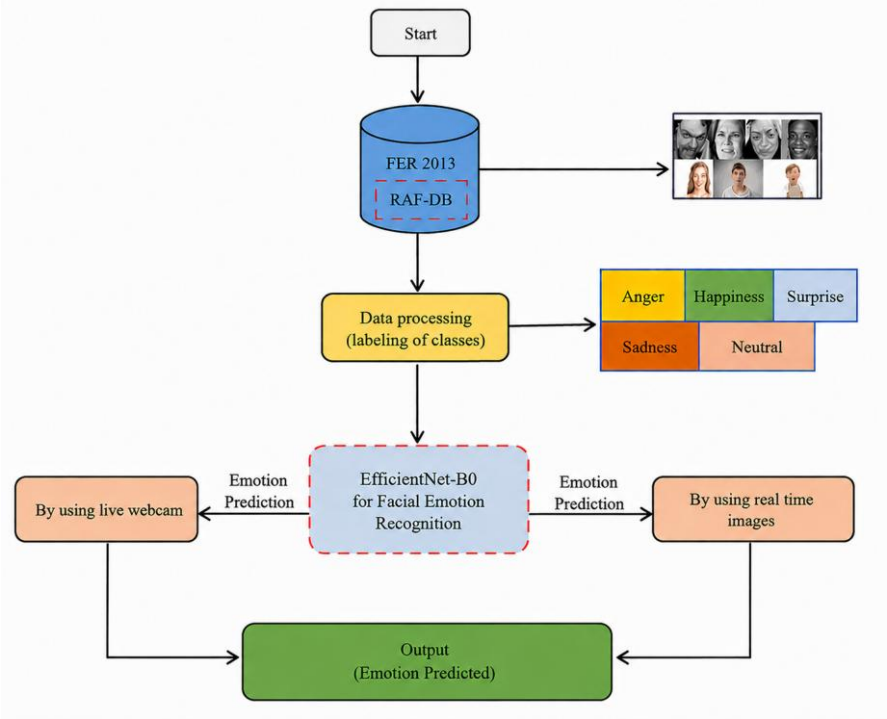


Fig. 1. Architecture of the proposed EfficientNet-B0 facial emotion recognition system.

3.5 Model Training Configuration

Training proceeded in two stages. In the first stage the convolutional base was frozen and only the new classification head was trained for six epochs with the Adam optimiser (Kingma & Ba, 2015) at a learning rate of 10^{-3}. In the second stage all layers except batch normalisation were unfrozen and the whole network was fine-tuned with AdamW and a weight decay of 10^{-4}. The learning rate rose linearly from 10^{-6} to 10^{-4} over the first two epochs and then followed a cosine decay to 1% of its peak over the planned 50 epochs. The batch size was 64 (313 steps per epoch). The checkpoint with the highest validation accuracy was saved, and early stopping ended fine-tuning after 15 epochs without improvement. Table 2 summarises the configuration.

Table 2. EfficientNet-B0 experimental and training configuration.

Parameter	Configuration
Backbone	EfficientNet-B0, ImageNet-pretrained
Input	224 × 224 × 3 (grayscale replicated to three channels)
Classification head	Global average pooling, dropout 0.3, dense 256 (swish, L2 10^{-4}), dropout 0.3, softmax (5)
Parameters	4,378,792 total; 4,294,753 trainable
Stage 1	6 epochs, backbone frozen, Adam, learning rate 10^{-3}
Stage 2	Up to 50 epochs, AdamW (weight decay 10^{-4}), 2-epoch warm-up to 10^{-4}, cosine decay
Batch size	64 (313 steps per epoch)
Loss	Categorical cross-entropy, label smoothing 0.1
Augmentation	Flip, rotation ±14°, zoom 12%, shift 8%, contrast 25%, brightness 15%
Mixed-sample augmentation	Mixup ($\alpha = 0.2$) and CutMix ($\alpha = 1.0$) on 50% of batches
Model selection	Best validation accuracy; early stopping with patience 15
Test-time augmentation	Mean of original and horizontally flipped image
Environment	Kaggle GPU, TensorFlow 2.20, mixed precision, seed 42

3.6 Comparative Backbones: EfficientNet-B1 and EfficientNet-B2

To test whether a larger backbone improves recognition, EfficientNet-B1 and EfficientNet-B2 were trained on the same saved split, with the same classification head, augmentation, Mixup, CutMix, label smoothing and evaluation code. EfficientNet-B1 used its native input size of 240 × 240 pixels and the same schedule as EfficientNet-B0: six head epochs followed by up to 50 fine-tuning epochs with early stopping. EfficientNet-B2 could not be trained with the full schedule within the session time available on Kaggle. It was therefore trained at 224 × 224 pixels rather than its native 260 × 260, with one head epoch and up to six fine-tuning epochs under a 24-minute budget, a half-epoch warm-up and the same peak learning rate. Its results therefore reflect a short training schedule and are reported with that qualification.

3.7 Baseline System

The real-time FER system of Sowmiya et al. (2026) was adopted as the comparative baseline. It uses a MobileNet-based CNN trained on FER2013 and CK+ with a supplementary image set, classifies seven emotions (anger, disgust, fear, happiness, sadness, surprise and neutral), reports a test accuracy of 70.7% on its native FER2013 test samples, an inference time of about 18 ms per frame and 30 to 33 frames per second on an NVIDIA GPU, and is deployed as a mobile application, FACEBOOTH, with a Flask backend. Because the baseline differs in its emotion set, datasets, evaluation protocol and hardware, its figures are used as contextual benchmarks rather than as a strictly equivalent comparison.

3.8 Evaluation Metrics

Performance was assessed with overall accuracy, class-level precision, recall and F1-score, their macro averages, the confusion matrix and one-versus-rest ROC analysis with the area under the curve (AUC) for each class. For a class c with true positives TP_c, false positives FP_c and false negatives FN_c, the precision P_c, recall R_c and F1-score $F1_c$ are computed as below, where N is the number of test images and K = 5 is the number of classes:

$$Accuracy = \Sigma \text{ correct predictions} / N \quad (1)$$

$$P_c = TP_c / (TP_c + FP_c) \quad (2)$$

$$R_c = TP_c / (TP_c + FN_c) \quad (3)$$

$$F1_c = 2 P_c R_c / (P_c + R_c) \quad (4)$$

$$Macro\text{-}F1 = (1/K) \Sigma F1_c \quad (5)$$

Because each class has exactly 500 test images, weighted averages equal macro averages and macro recall equals accuracy. The 95% confidence interval of a test accuracy p was estimated as $p \pm 1.96 \sqrt{(p(1 - p)/N)}$, which gives about ±1.6 percentage points for N = 2,500. Pairs of models were compared with McNemar's test (McNemar, 1947; Dietterich, 1998), computed as an exact two-sided binomial test on the b and c test images that only one of the two models classified correctly (the value is capped at 1):

$$p = 2 \cdot P(X \leq min(b, c)), \; X \sim Binomial(b + c, 0.5) \quad (6)$$

Computational cost was measured by the number of parameters, the mean time per image for batches of 32 (30 repetitions), and the median latency for a single image (200 repetitions), after a warm-up and on the same GPU for all three backbones.

3.9 Deployment

The trained model was integrated into a client-server web application. The frontend, written in HTML, CSS and JavaScript, lets a user upload or capture a facial image. The Python backend receives the image through a FastAPI RESTful endpoint, applies the same preprocessing used in training, runs EfficientNet-B0 inference and returns the predicted emotion and the probability of each class as a JSON response, which the interface displays with a confidence score.

4 Results

4.1 Exploratory Data Analysis

Before balancing, the five classes were unevenly represented. Happiness accounted for 14,946 of the 44,325 collected images (33.7%) and surprise for 5,621 (12.7%). The two sources also differed in composition: anger came mainly from FER2013 (4,953 of 5,820

images), while RAF-DB contributed 40% of the happiness images. Balancing to 5,000 images per class removed the influence of class frequency, so differences between classes in the results reflect how distinguishable each expression is rather than how often it occurs. Fig. 2 shows the final allocation of images across the three subsets.

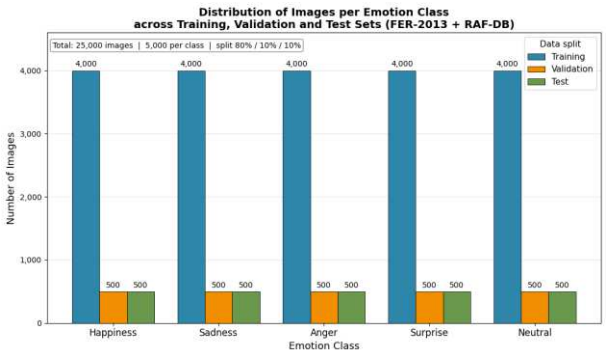


Fig. 2. Images per emotion class in the training, validation and test sets of the balanced dataset (25,000 images).

4.2 Training Convergence

Fig. 3 shows accuracy and loss for the training and validation sets across both stages. During the six head-only epochs validation accuracy rose to 54.96%. Unfreezing the convolutional base produced a sharp improvement, to 62.12% after the first fine-tuning epoch, and validation accuracy reached its best value of 79.04% at epoch 39 overall (the 33rd fine-tuning epoch). At that epoch the training accuracy was 87.94%, the training loss 0.7550 and the validation loss 0.9157. Early stopping ended training after 54 epochs in total, and fine-tuning took 124.6 minutes. The loss values are higher than would be expected with plain cross-entropy because label smoothing, Mixup and CutMix keep the targets soft, and the training accuracy was measured on augmented and mixed batches. On 2,500 clean training images without augmentation, the model reached 97.36%.

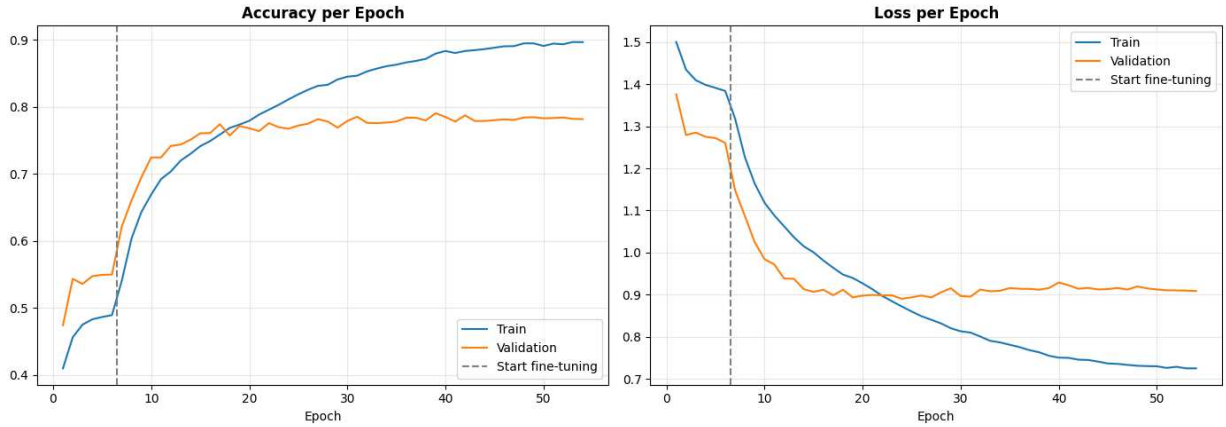


Fig. 3. Training and validation accuracy (left) and loss (right) per epoch for EfficientNet-B0. The dashed line marks the start of fine-tuning.

4.3 Overall and Class-Level Performance

On the 2,500-image test set the selected model achieved an accuracy of 78.68% (Table 3), with a 95% confidence interval of about ±1.6 percentage points; without flip test-time augmentation the accuracy was 78.24%. The best validation accuracy (79.04%) and the test accuracy differ by only 0.36 percentage points, which indicates that selecting the checkpoint on the validation set did not overfit to it. The macro-averaged precision, recall and F1-score were 79.16%, 78.68% and 78.83%.

Table 3. Overall performance of the EfficientNet-B0 model.

Metric	Result
Best validation accuracy (epoch 39)	79.04%
Test accuracy (flip TTA)	78.68%
Test accuracy (no TTA)	78.24%
95% confidence interval of test accuracy	77.07% to 80.29%
Training accuracy, augmented batches (epoch 39)	87.94%
Clean training accuracy	97.36%
Training / validation loss (epoch 39)	0.7550 / 0.9157
Macro precision / recall / F1	79.16% / 78.68% / 78.83%
Weighted precision / recall / F1	79.16% / 78.68% / 78.83%
Macro AUC	0.9478
Accuracy on FER2013 / RAF-DB test images	75.70% / 84.63%

Table 4 gives the class-level results. Surprise and happiness were recognised best, with F1-scores of 87.78% and 87.50%. Anger reached 77.15%, while sadness (71.60%) and neutral (70.13%) were the most difficult. Neutral had the lowest precision (66.49%) but a higher recall (74.20%), so the model assigned the neutral label too often. Anger showed the opposite pattern, with a precision of 81.06% and a recall of 73.60%: anger predictions were usually correct, but about a quarter of the genuine anger images were assigned to other classes.

Table 4. Class-level performance of EfficientNet-B0 on the test set (500 images per class).

Class	Precision	Recall	F1-score	AUC
Happiness	88.21%	86.80%	87.50%	0.9807
Sadness	70.62%	72.60%	71.60%	0.9241
Anger	81.06%	73.60%	77.15%	0.9331
Surprise	89.42%	86.20%	87.78%	0.9783
Neutral	66.49%	74.20%	70.13%	0.9229
Macro average	**79.16%**	**78.68%**	**78.83%**	**0.9478**

4.4 Confusion Matrix and ROC Analysis

Fig. 4 shows the confusion matrix in counts and as row-normalised percentages. Of the 2,500 test images, 1,967 were classified correctly: 434 happiness, 431 surprise, 371 neutral, 368 anger and 363 sadness. The largest source of error was mutual confusion between sadness and neutral: 88 sadness images (17.6%) were predicted as neutral and 83 neutral images (16.6%) as sadness. Of the 558 images predicted as neutral only 371 were correct, which explains the low precision of that class. Anger was most often misclassified as sadness (46 images, 9.2%) and neutral (39, 7.8%), and 36 happiness images (7.2%) were predicted as neutral. Missed surprise images went mainly to neutral (24) and anger (23). These patterns are consistent with low-intensity expressions of sadness, anger and happiness resembling a neutral face.

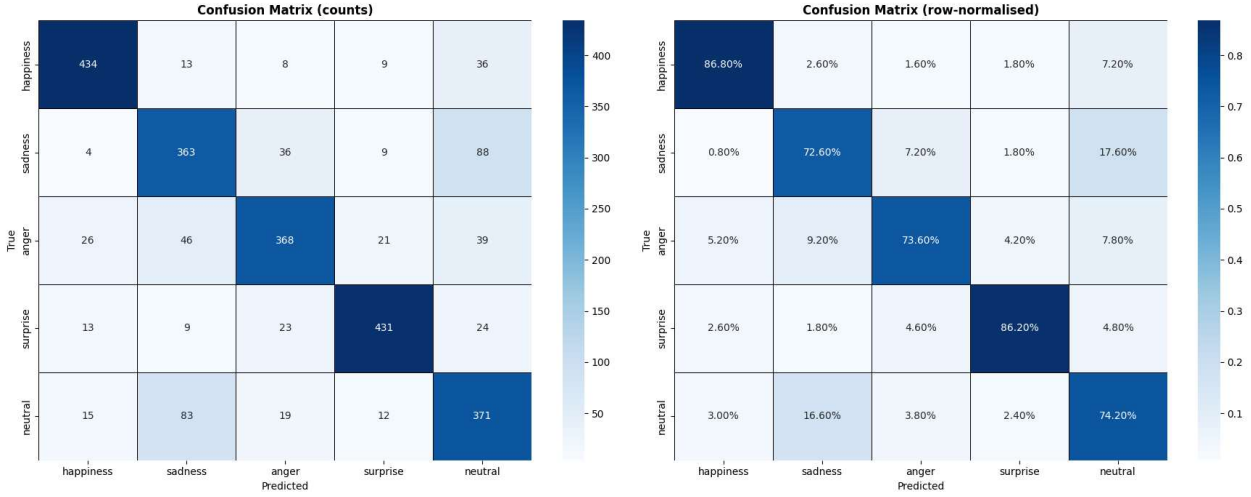


Fig. 4. Confusion matrix of EfficientNet-B0 on the 2,500 test images, in counts (left) and row-normalised (right).

The one-versus-rest ROC curves in Fig. 5 measure how well each class is separated from the others across all decision thresholds. The AUC was 0.9807 for happiness, 0.9783 for surprise, 0.9331 for anger, 0.9241 for sadness and 0.9229 for neutral, giving a macro-averaged AUC of 0.9478. The ordering of the classes by AUC mirrors their F1-scores, and all five values are above 0.92, so the model ranks the correct class highly even for the expressions it finds hardest to label at a fixed threshold.

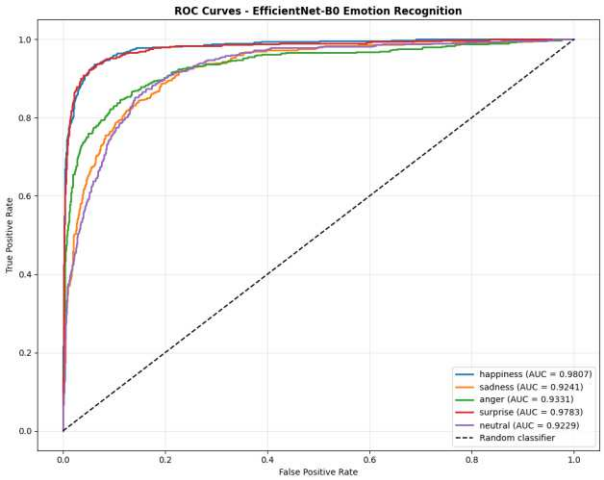


Fig. 5. One-versus-rest ROC curves and AUC values of EfficientNet-B0 for the five emotion classes.

Accuracy also depended on the source of the image. The model classified 75.70% of the 1,667 FER2013 test images correctly and 84.63% of the 833 RAF-DB test images, a difference of about nine percentage points. The lower-resolution FER2013 images were harder to classify, although the experiment does not establish whether the gap is due to image quality, labelling consistency or both.

4.5 Comparison of EfficientNet-B0, B1 and B2

Table 5 compares the three backbones on the same 2,500 test images. EfficientNet-B1 achieved the highest test accuracy (79.76%) and macro F1-score (79.73%), 1.08 and 0.90 percentage points above EfficientNet-B0. This difference is smaller than the ±1.6-point confidence interval of either model, and McNemar's test on the paired predictions did not find it significant ($p = 0.107$; Table 6). EfficientNet-B0 had the highest macro AUC (0.9478 against 0.9431 for B1) and the highest best validation accuracy (79.04% against 78.72%). EfficientNet-B1 also fitted its training data more closely, with a clean training accuracy of 99.36% against 97.36%, so most of its additional capacity was absorbed by the training set rather than converted into test accuracy.

Table 5. Comparison of EfficientNet-B0, B1 and B2 trained on the same split and evaluated on the same 2,500 test images.

Metric	B0	B1	B2
Input size (pixels)	224	240	224[a]
Parameters (millions)	4.38	6.90	8.13
Head + fine-tuning epochs	6 + 48	6 + 50	1 + 6[a]
Best epoch	39	52	7
Best validation accuracy	79.04%	78.72%	74.84%
Test accuracy	78.68%	79.76%	75.44%
Test accuracy, no TTA	78.24%	79.68%	75.68%
Macro precision	79.16%	79.73%	75.47%
Macro recall	78.68%	79.76%	75.44%
Macro F1-score	78.83%	79.73%	75.40%
Macro AUC	0.9478	0.9431	0.9436
FER2013 test accuracy	75.70%	76.36%	72.23%
RAF-DB test accuracy	84.63%	86.55%	81.87%
Clean training accuracy	97.36%	99.36%	80.60%
Single-image latency (ms)	8.50	12.42	12.70
Batch latency (ms per image)	2.10	3.37	3.13
Fine-tuning time (min)	124.6	206.7	23.6

[a] *EfficientNet-B2 was trained at 224 × 224 pixels (native 260 × 260) under a reduced schedule because of the session time limit.*

Table 6. McNemar's test on the paired predictions for the 2,500 test images.

Comparison	Test accuracy	*p*-value	Result
B0 vs B1	78.68% vs 79.76%	0.107	Not significant
B0 vs B2	78.68% vs 75.44%	< 0.001	Significant
B1 vs B2	79.76% vs 75.44%	< 0.001	Significant

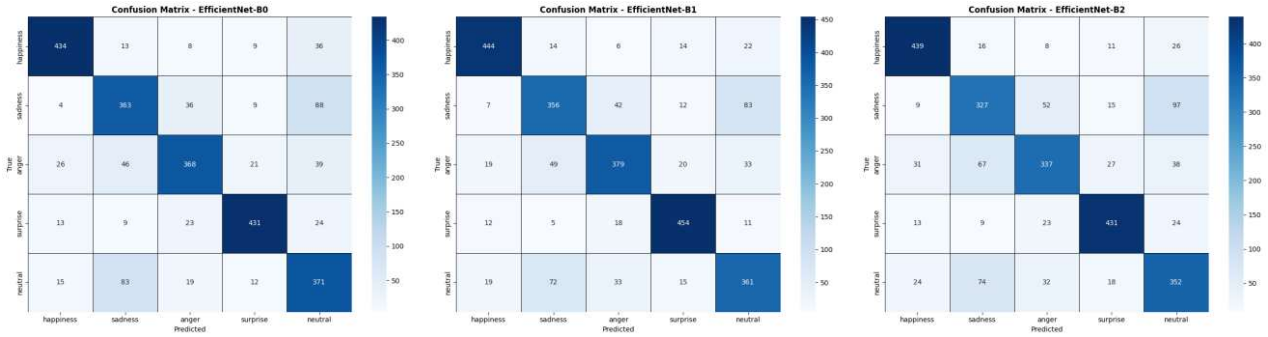
The cost of EfficientNet-B1 was substantially higher. It has 6.90 million parameters, 57.7% more than EfficientNet-B0, its single-image latency was 12.42 ms against 8.50 ms (about 81 against 118 images per second), its batch throughput was 3.37 ms per image against 2.10 ms, and its fine-tuning took 206.7 minutes against 124.6 minutes. Put the other way, EfficientNet-B0 used 36.6% fewer parameters, was 31.6% faster for single images and needed 39.7% less fine-tuning time, for an accuracy difference that was not statistically significant.

EfficientNet-B2 reached 75.44% accuracy and a macro F1-score of 75.40%, significantly below both EfficientNet-B0 and EfficientNet-B1 ($p < 0.001$ in both comparisons). Its clean training accuracy of only 80.60% shows that it was still underfitting when its reduced schedule of seven epochs ended, so this result reflects the training budget it received rather than the capacity of the architecture. Even at 224 × 224 pixels it was the slowest backbone for single images (12.70 ms), and it has 8.13 million parameters.

Table 7. Per-class F1-scores of the three backbones.

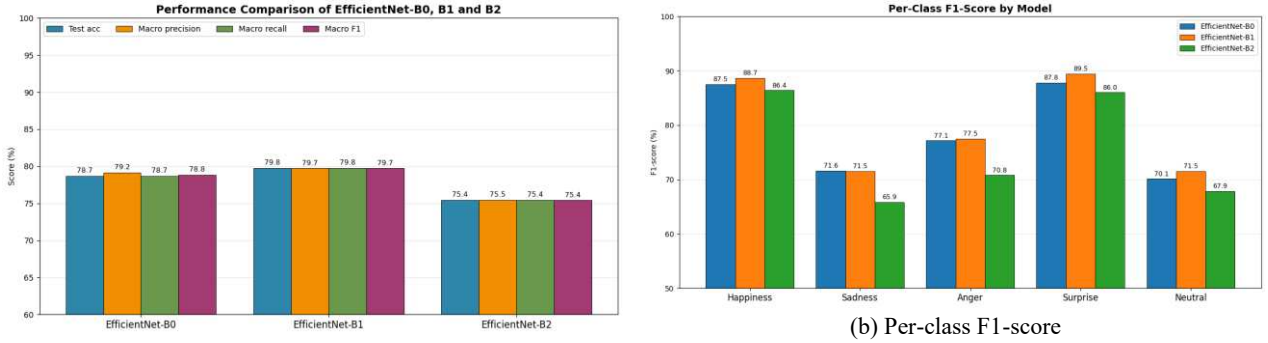
Class	B0	B1	B2
Happiness	87.50%	88.71%	86.42%
Sadness	71.60%	71.49%	65.86%
Anger	77.15%	77.51%	70.80%
Surprise	87.78%	89.46%	86.03%
Neutral	70.13%	71.49%	67.89%
Macro average	**78.83%**	**79.73%**	**75.40%**

Table 7 shows that the three backbones agree on which emotions are easy and which are hard. EfficientNet-B1 improved the F1-score of surprise (89.46% against 87.78%), happiness (88.71% against 87.50%) and neutral (71.49% against 70.13%), while EfficientNet-B0 was marginally better on sadness (71.60% against 71.49%). EfficientNet-B2 was lowest in every class, most clearly for sadness (65.86%) and anger (70.80%). The confusion matrices in Fig. 6 show the same structure for all three models, with sadness and neutral forming the largest off-diagonal pair. Fig. 7 summarises the aggregate metrics and per-class F1-scores.



(a) EfficientNet-B0 (b) EfficientNet-B1 (c) EfficientNet-B2

Fig. 6. Confusion matrices of (a) EfficientNet-B0, (b) EfficientNet-B1 and (c) EfficientNet-B2 on the same 2,500 test images.



(a) Aggregate metrics (b) Per-class F1-score

Fig. 7. Comparison of EfficientNet-B0, B1 and B2: (a) test accuracy and macro precision, recall and F1-score; (b) F1-score per emotion class.

4.6 Comparison with the Existing System

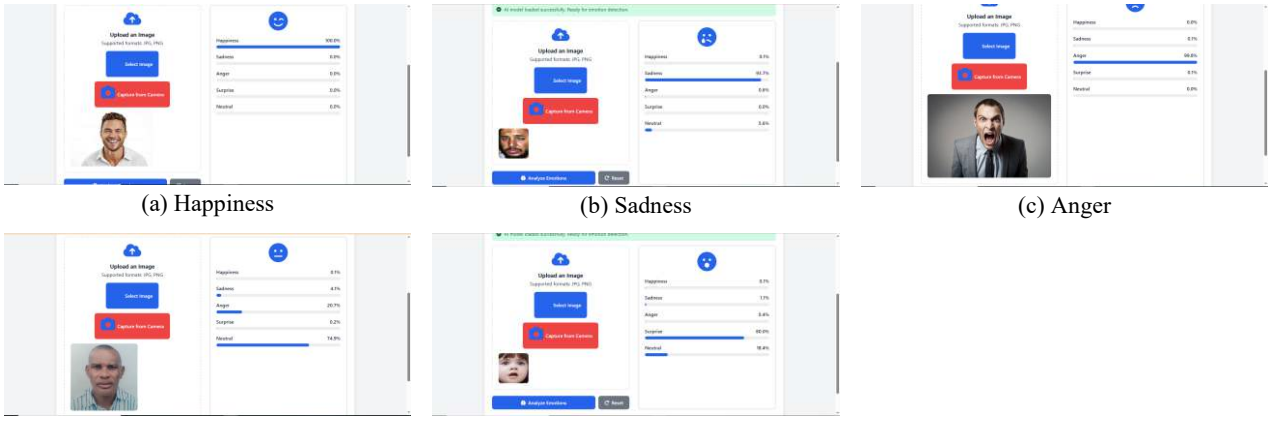
Table 8 compares the proposed system with the MobileNet-based system of Sowmiya et al. (2026). The proposed model achieved a higher test accuracy (78.68% against 70.7%) and a lower inference time (8.50 ms per image for single-image prediction against about 18 ms per frame), corresponding to about 118 frames per second against 30 to 33. It also reports precision, recall, F1-score, AUC and parameter counts, which the baseline does not. The two systems solve different problems, five emotions against seven, using different datasets, evaluation protocols and hardware, so the comparison indicates the relative position of the proposed system rather than a like-for-like ranking. In particular, the baseline's weakest classes, fear and disgust, are not part of the five-class task studied here.

Table 8. Comparison of the proposed system with the existing MobileNet-based system (Sowmiya et al., 2026).

Metric	Existing system	Proposed system
Architecture	MobileNet-based CNN	EfficientNet-B0 with transfer learning
Datasets	FER2013 and CK+ with supplementary images	FER2013 and RAF-DB, de-duplicated and balanced
Emotion classes	7	5
Test accuracy	70.7% (native FER2013)	78.68%
Macro F1-score	Not reported	78.83%
Macro AUC	Not reported	0.948
Inference time	About 18 ms per frame	8.50 ms (single image); 2.10 ms (batch of 32)
Frames per second	30 to 33 (NVIDIA GPU)	About 118 (single image)
Parameters	Not reported	4,378,792
Deployment	FACEBOOTH mobile application, Flask backend	Web application, FastAPI backend

4.7 Deployed System Performance

The deployed web application was tested with one image for each of the five emotions (Fig. 8). In each case the system located the face, applied the preprocessing pipeline and returned the correct emotion with its confidence score (Table 9). Confidence was highest for happiness (100.0%) and anger (99.8%) and lowest for neutral (74.9%), where anger received the second-highest probability (20.7%), and for surprise (80.0%), where neutral received 18.4%; both observations agree with the confusions seen on the test set. End-to-end response times, from submission to displayed result, ranged from 2.1 to 3.2 seconds. Since model inference takes about 8.5 ms, almost all of this time is spent on network transfer, image upload and application processing rather than on the network itself.



(a) Happiness (b) Sadness (c) Anger

(d) Neutral (e) Surprise

Fig. 8. Prediction results from the deployed web application for (a) happiness, (b) sadness, (c) anger, (d) neutral and (e) surprise.

Table 9. Sample predictions from the deployed web application.

Test case	Predicted emotion	Confidence	Response time (s)
Image 1	Happiness	100.0%	2.3
Image 2	Sadness	93.7%	3.2
Image 3	Anger	99.8%	2.1
Image 4	Neutral	74.9%	3.0
Image 5	Surprise	80.0%	2.5

5 Discussion

The results show that EfficientNet-B0, trained on a cleaned and balanced combination of FER2013 and RAF-DB, recognises five emotions with a test accuracy of 78.68% and a macro F1-score of 78.83% on images that were never used for training or model selection. Because duplicates were removed and the subsets were checked for overlap, this figure is a more trustworthy estimate than accuracies obtained on splits in which the same image can appear on both sides. The result is not directly comparable with seven-class FER2013 results such as the 73.28% of Khaireddin and Chen (2021) or the 68.78% of an EfficientNet-B2 model trained on FER2013 (Naik et al., 2026), because the present task excludes fear and disgust and includes RAF-DB images, but it places the model in a reasonable range for in-the-wild FER.

The class-level results identify where the model fails. Happiness and surprise, which involve pronounced changes around the mouth, eyes and brows, were recognised with F1-scores above 87%. Sadness and neutral were confused with each other in both directions, and the neutral label was over-assigned, which suggests that the model treats low-intensity expressions as neutral when the visual evidence is weak. Anger was predicted reliably when it was predicted but was often missed. Since no attribution method such as Grad-CAM was applied, these observations describe the model's behaviour rather than establishing which facial regions drive its decisions.

The backbone comparison is the main new finding. Scaling from EfficientNet-B0 to EfficientNet-B1 increased the number of parameters by 57.7% and single-image latency by 46%, yet produced a gain of 1.08 percentage points in accuracy that was not statistically significant and slightly lower ranking quality (macro AUC 0.9431 against 0.9478). One plausible explanation is that compound scaling increases input resolution along with depth and width, but most FER2013 images are only 48 × 48 pixels, so a larger input mostly upsamples information that is already present. The higher clean training accuracy of EfficientNet-B1 (99.36%) also suggests that its extra capacity was used mainly to fit the training images. For a five-class task of this size, EfficientNet-B0 therefore offers the best balance between accuracy and cost of the three backbones tested. The lower accuracy of EfficientNet-B2 should not be read as evidence that the architecture is unsuitable, because it was trained for only seven epochs and was still underfitting; a full-length run at its native resolution would be needed to settle its position.

The nine-point gap between FER2013 (75.70%) and RAF-DB (84.63%) test images shows that performance depends on image source, and the overall accuracy therefore depends on the mixture of sources in the test set. In line with the observation of Dewi et al. (2026) that most disparities originate in the data, balancing the classes and removing duplicates addressed two forms of data-level bias, but no demographic fairness audit was carried out, and differences across age, gender or ethnicity remain untested.

The deployment results distinguish model speed from system speed. The model's single-image latency of 8.5 ms is well within real-time requirements, but the end-to-end response time of 2.1 to 3.2 seconds is dominated by network and application overhead. Further gains in responsiveness are therefore more likely to come from image compression, asynchronous request handling or on-device inference than from a smaller network.

Several limitations qualify these findings. All models were trained once, with a single seed, so the variability between runs was not measured, and McNemar's test was applied to a single split. The test images come from the same two datasets as the training images, so performance on images from other sources, cameras or populations is unknown. The task is limited to five emotions in static images and does not address fear, disgust, compound expressions or temporal dynamics in video. EfficientNet-B2 was trained under a reduced schedule. Inference times were measured on a GPU, and the baseline's figures were obtained on unspecified hardware with a different protocol. Finally, no controlled tests of occlusion, pose or lighting were performed, so the robustness weaknesses reported for the baseline system remain untested for the proposed one.

6 Conclusion

This study developed and evaluated an EfficientNet-B0 model for five-class facial emotion recognition on a de-duplicated, balanced combination of FER2013 and RAF-DB with a strictly independent test set. The model achieved a test accuracy of 78.68%, a macro F1-score of 78.83% and a macro AUC of 0.948, recognising surprise and happiness best and sadness and neutral least well. Trained on the same split, EfficientNet-B1 was not significantly more accurate (79.76%, $p = 0.107$) despite 57.7% more parameters, 46% higher single-image latency and 66% longer fine-tuning, while a reduced-schedule EfficientNet-B2 was significantly less accurate (75.44%). Compared with the existing MobileNet-based system, the proposed model was faster (8.5 ms against about 18 ms per image) and more accurate on its own task (78.68% against 70.7%), although the two systems were evaluated on different emotion sets and data. The model was deployed in a working FastAPI web application.

The contributions of this study are: (i) a reproducible preparation procedure for combined FER2013 and RAF-DB data that removes duplicates and conflicting labels and prevents leakage between subsets; (ii) a two-stage EfficientNet-B0 model with class-level, ROC and per-source evaluation; (iii) evidence from a controlled comparison that EfficientNet-B0 gives the best balance between accuracy and computational cost among EfficientNet-B0, B1 and B2 for this task; and (iv) a functional real-time web application. Future work should train EfficientNet-B2 with a full schedule at its native resolution, repeat all runs with several seeds, validate the model on independently collected images, add attribution methods such as Grad-CAM, test robustness to occlusion, pose and lighting, extend the task to fear, disgust and video sequences, and carry out a demographic fairness audit.

Declarations

Funding: The authors declare that no funding was received for the conduct of this research.

Conflict of Interest: The authors declare that they have no conflict of interest.

Data Availability: FER2013 and RAF-DB are publicly available benchmark datasets that can be obtained from their providers for research use. The data split, trained model files and evaluation results are available from the corresponding author upon reasonable request.

Ethical Approval: This study used only publicly available, previously collected facial expression datasets and did not involve the recruitment of human participants or the collection of new biometric data, so formal ethical approval was not required. Any deployment that processes user-submitted facial images should include appropriate consent, privacy and data protection safeguards.

Author Contributions: P.S.J. conceptualised the study, prepared the data, developed and evaluated the models, implemented the deployment system and drafted the manuscript. U.A.O., P.E. and N.K.U. supervised the research, contributed to the methodology, and reviewed and edited the manuscript. All authors read and approved the final manuscript.

References

Chowdary, M. K., Nguyen, T. N., & Hemanth, D. J. (2023). Deep learning-based facial emotion recognition for human-computer interaction applications. *Neural Computing and Applications, 35*(32), 23311–23328. https://doi.org/10.1007/s00521-021-06012-8

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848

Dewi, D. A., Okengwu, U., & Rizqi, Z. U. (2026). Bias detection and mitigation techniques in data science pipelines: An empirical evaluation. *International Journal of Advances in Artificial Intelligence and Machine Learning, 3*(1), 34–43. https://doi.org/10.58723/ijaaiml.v3i1.655

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(4), 384–392. https://doi.org/10.1037/0003-066X.48.4.384

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., … Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013* (pp. 117–124). Springer. https://doi.org/10.1007/978-3-642-42051-1_16

Haq, H. B. U., Akram, W., Irshad, M. N., Kosar, A., & Abid, M. (2024). Enhanced real-time facial expression recognition using deep learning. *Acadlore Transactions on AI and Machine Learning, 3*(1), 24–35. https://doi.org/10.56578/ataiml030103

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint* arXiv:1704.04861. https://doi.org/10.48550/arXiv.1704.04861

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7132–7141). https://doi.org/10.1109/CVPR.2018.00745

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4700–4708). https://doi.org/10.1109/CVPR.2017.243

Khaireddin, Y., & Chen, Z. (2021). Facial emotion recognition: State of the art performance on FER2013. *arXiv preprint* arXiv:2105.03588. https://doi.org/10.48550/arXiv.2105.03588

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1412.6980

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

Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2852–2861). https://doi.org/10.1109/CVPR.2017.277

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1711.05101

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153–157. https://doi.org/10.1007/BF02295996

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(2), 1359–1367. https://doi.org/10.1609/aaai.v34i02.5492

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10*(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

Naik, S., Bagayatkar, S., & Singh, P. (2026). Facial emotion recognition on FER-2013 using an EfficientNetB2-based approach. *arXiv preprint* arXiv:2601.18228. https://doi.org/10.48550/arXiv.2601.18228

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4510–4520). https://doi.org/10.1109/CVPR.2018.00474

Savchenko, A. V. (2021). Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *Proceedings of the IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*. https://doi.org/10.1109/SISY52375.2021.9582508

Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing, 27*(6), 803–816. https://doi.org/10.1016/j.imavis.2008.08.005

Siddiqui, M. F. H., & Javaid, A. Y. (2020). A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images. *Multimodal Technologies and Interaction, 4*(3), 46. https://doi.org/10.3390/mti4030046

Sowmiya, N., Anitharani, A., et al. (2026). Real time facial emotion recognition using deep learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 12*(2), 237–246. https://doi.org/10.32628/CSEIT213181

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2818–2826). https://doi.org/10.1109/CVPR.2016.308

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)* (Vol. 97, pp. 6105–6114). PMLR.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing, 11*(8), 1301–1309. https://doi.org/10.1109/JSTSP.2017.2764438

Yun, S., Han, D., Chun, S., Oh, S. J., Yoo, Y., & Choe, J. (2019). CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*

(ICCV) (pp. 6023–6032). https://doi.org/10.1109/ICCV.2019.00612

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1103–1114). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1115

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1710.09412