JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7404

MULTI-ARCHITECTURE DEEP LEARNING FOR SUBJECT INDEPENDENT FACIAL EXPRESSION RECOGNITION

Taufiq[1*]; Muhammad[1]; Asran[1]; Ezwarsyah[1]; Muchlis Abdul Muthalib[1]

Electrical Engineering, Engineering Faculty[1]
Universitas Malikusaleh, Aceh, Indonesia[1]
https://unimal.ac.id/[1]
taufiq.te@unimal.ac.id*, muhammad.te@unimal.ac.id, asran@unimal.ac.id, ezwarsyah@unimal.ac.id, muchlis.abd@unimal.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Facial Expression Recognition (FER) remains a challenging problem in computer vision, particularly under subject-independent conditions in which models must generalize to individuals not seen during training. This study reports a controlled comparative evaluation of three Convolutional Neural Network (CNN) architectures — MobileNetV3-Large, EfficientNet-B3, and ResNet50 — using the Extended Cohn-Kanade (CK+) dataset (981 apex-frame images, 118 subjects, seven emotion classes). All models were trained and tested under identical experimental conditions with a subject-disjoint partition (72/23/23 subjects for training, validation, and testing), so that observed performance differences may be attributed primarily to architectural design. The results indicate that MobileNetV3-Large attains the highest test accuracy of 95.16%, exceeding EfficientNet-B3 (93.01%) and ResNet50 (91.94%), while requiring the fewest parameters (~5.4M) and the shortest inference latency (~8.2 ms per image). A multi-dimensional evaluation covering per-class metrics and computational cost is also reported. These observations provide preliminary architectural guidance for FER deployment in resource-constrained environments; however, because they are derived from a single dataset and a single subject split, broader claims should be confirmed on more diverse benchmarks*.

Keywords: *Computational Cost Analysis, Convolutional Neural Network, Deep Learning Architecture Comparison, Facial Expression Recognition, Subject-Independent Validation.*

Intisari— *Pengenalan ekspresi wajah (FER) tetap menjadi persoalan menantang dalam computer vision, terutama dalam kondisi subject-independent ketika model harus mampu menggeneralisasi pada individu yang tidak dijumpai selama pelatihan. Penelitian ini menyajikan evaluasi perbandingan terkendali atas tiga arsitektur CNN — MobileNetV3-Large, EfficientNet-B3, dan ResNet50 — pada dataset Extended Cohn-Kanade (CK+) yang berisi 981 citra apex-frame dari 118 subjek dalam tujuh kelas emosi. Seluruh model dilatih dan diuji dalam kondisi eksperimental identik dengan pembagian berbasis subjek terpisah (72/23/23 untuk pelatihan, validasi, dan pengujian) sehingga perbedaan performa dapat dikaitkan terutama dengan desain arsitektur. Hasil menunjukkan bahwa MobileNetV3-Large mencapai akurasi pengujian tertinggi sebesar 95,16%, melampaui EfficientNet-B3 (93,01%) dan ResNet50 (91,94%), dengan jumlah parameter paling sedikit (~5,4 juta) dan waktu inferensi tersingkat (~8,2 ms per citra). Evaluasi multidimensi mencakup metrik per-kelas dan biaya komputasi turut dilaporkan. Temuan ini memberikan panduan arsitektural awal untuk penerapan FER di lingkungan dengan sumber daya terbatas; karena diperoleh dari satu dataset dengan satu pembagian subjek, klaim yang lebih luas perlu dikonfirmasi pada benchmark yang lebih beragam.*

Kata Kunci: *Analisis Biaya Komputasi, Convolutional Neural Network, Perbandingan Arsitektur Deep Learning, Pengenalan Ekspresi Wajah, Validasi Subject-Independent.*

INTRODUCTION

Human facial expressions constitute one of the most universal and effective forms of non-verbal communication for conveying emotional states [1],[2]. In daily interpersonal interactions, individuals constantly display facial expressions that reflect their internal emotions, enabling observers to understand and respond to the emotional context of communication. Human facial expressions are generally categorized into seven fundamental types: anger, contempt, disgust, fear, happiness, sadness, and surprise [3]. These expressions play a critical role not only in human-to-human interactions but also in diverse application domains, including psychology, education, healthcare, and human-computer interaction [4].

The ability to automatically recognize facial expressions has become an increasingly important research topic in artificial intelligence and computer vision, driven by its broad applicability in security surveillance, mental health assessment, and interactive technologies. Facial expression recognition (FER) systems can assist in diagnosing emotional or mental disorders such as depression and anxiety [1], detecting suspicious behavior in surveillance environments, and enabling responsive virtual characters in entertainment applications. However, recognition performance is influenced by several factors, including the intensity of expressions, individual facial morphology, and cultural differences in emotional display [5].

Convolutional Neural Networks (CNNs) have become the dominant paradigm for image-based FER because they learn hierarchical spatial features directly from pixels, removing the need for handcrafted descriptors such as Local Binary Patterns and geometric landmark distances [6], [7], [8], [9]. Modern CNN families differ primarily in how they balance representational depth, attention to discriminative regions, and computational efficiency [10], properties that motivate the comparative scope of the present study.

Among contemporary CNN architectures, three design philosophies have demonstrated particular promise for visual recognition tasks. MobileNetV3-Large [11] employs neural architecture search (NAS) combined with squeeze-and-excitation (SE) modules to optimize computational efficiency while maintaining high representational capacity. EfficientNet-B3 [12] implements a compound scaling methodology that systematically scales network depth, width, and resolution, achieving competitive accuracy-efficiency trade-offs compared to single-dimension scaling approaches. ResNet50 [13] introduces residual learning through skip connections to address the vanishing gradient problem, enabling the training of substantially deeper networks while maintaining gradient flow stability.

Despite the extensive application of these architectures in various computer vision tasks, comprehensive comparative studies evaluating their performance for facial expression recognition under subject-independent validation remain limited. Most existing FER studies either focus on a single architecture [14], employ subject-dependent evaluation protocols that overestimate generalization performance [15], or report results under inconsistent experimental conditions that preclude fair comparison [16]. Subject-independent validation — where training and testing are performed on completely disjoint subject sets — is critical for assessing a model's ability to generalize to unseen individuals across diverse demographic backgrounds [17], which is a fundamental requirement for real-world deployment.

Based on the outlined challenges and research gaps, this study presents a multi-dimensional comparative analysis of MobileNetV3-Large, EfficientNet-B3, and ResNet50 for facial expression recognition under a subject-independent evaluation protocol on the CK+ dataset. The contributions of this work are threefold. First, the three architectures representing distinct design philosophies (NAS-driven efficiency, compound scaling, and residual learning) are trained and evaluated under identical conditions with full hyperparameter transparency, so that observed performance differences may be associated primarily with architectural design rather than experimental variation. Second, a subject-independent evaluation protocol with a disjoint subject split (72/23/23) is adopted, which provides a more realistic assessment of cross-subject generalization than the subject-dependent procedures commonly reported in prior CK+ studies.

Third, a multi-dimensional analysis encompassing classification accuracy, computational cost (parameters, FLOPs, inference latency), and per-class performance is reported — restricted to a controlled three-model setting and not intended as a broad architectural benchmark. Because the empirical evidence is drawn from a single dataset (CK+) with posed laboratory expressions and a single fixed subject split, the findings should be interpreted as preliminary architectural guidance and validated on more diverse benchmarks before broader conclusions are drawn.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7404

Based on the architectural properties summarized above, this study tests the following hypotheses. H1 (primary): under identical training and evaluation conditions on a subject-independent CK+ split, the three architectures will yield measurably different test accuracies, reflecting differences in their inductive biases rather than experimental variation. H2: architectures equipped with channel-wise attention mechanisms (MobileNetV3-Large and EfficientNet-B3) will achieve higher per-class F1-scores on visually proximate emotion categories than the residual-only baseline (ResNet50). H3: the architecture with the lowest parameter count and FLOPs is not necessarily the least accurate; specifically, MobileNetV3-Large is expected to offer a competitive accuracy-efficiency trade-off relative to deeper baselines on this small-scale, posed dataset. These hypotheses are evaluated through the experimental protocol detailed in Section III and the per-class and computational analyses reported in Section IV.

The remainder of this paper is organized as follows. Section II reviews related work on facial expression recognition using deep learning. Section III describes the materials and methods, including the dataset, preprocessing pipeline, model architectures, training configuration, and evaluation metrics. Section IV presents the experimental results and discussion, including accuracy analysis, computational cost comparison, and class-level performance analysis. Section V concludes the paper with key findings, limitations, and directions for future research.

Related Work

Facial expression recognition has been studied extensively in the deep learning era, with research broadly partitioned into static-image and sequence-based approaches [1]. Two methodological aspects are most directly relevant to the present comparative study: how individual CNN backbones perform under subject-independent evaluation, and how cross-architecture comparisons have been reported in prior CK+ work. Transfer learning from ImageNet has been shown to substantially improve FER performance on smaller benchmark datasets, enabling models to leverage pre-learned visual representations even when labeled facial expression data is limited [18]. On individual backbones, Savchenko [19] reported that lightweight pre-trained models fine-tuned for FER attain up to 65.8% accuracy on AffectNet under subject-independent conditions, observing that feature-extraction quality matters more than absolute model size. Fard and Mahoor [20] proposed an adaptive correlation-based loss for in-the-wild FER and obtained 88.0% on RAF-DB, with notable gains on subtle expressions such as contempt and fear. These studies, however, evaluate single architectures without cross-architecture comparison under harmonized experimental conditions.

The CK+ dataset [21], [22] remains one of the most widely used benchmarks for FER evaluation due to its controlled conditions, FACS-validated labels, and availability across subject-independent and subject-dependent settings. A critical concern in CK+ evaluation is the common use of subject-dependent protocols, where images from the same individual appear in both training and test sets. Pham et al. [15] demonstrated that this practice leads to substantially inflated accuracy estimates, with performance dropping significantly — often by 5–10 percentage points — when switching to subject-independent evaluation. This finding underscores the importance of disjoint subject splitting for realistic generalization assessment.

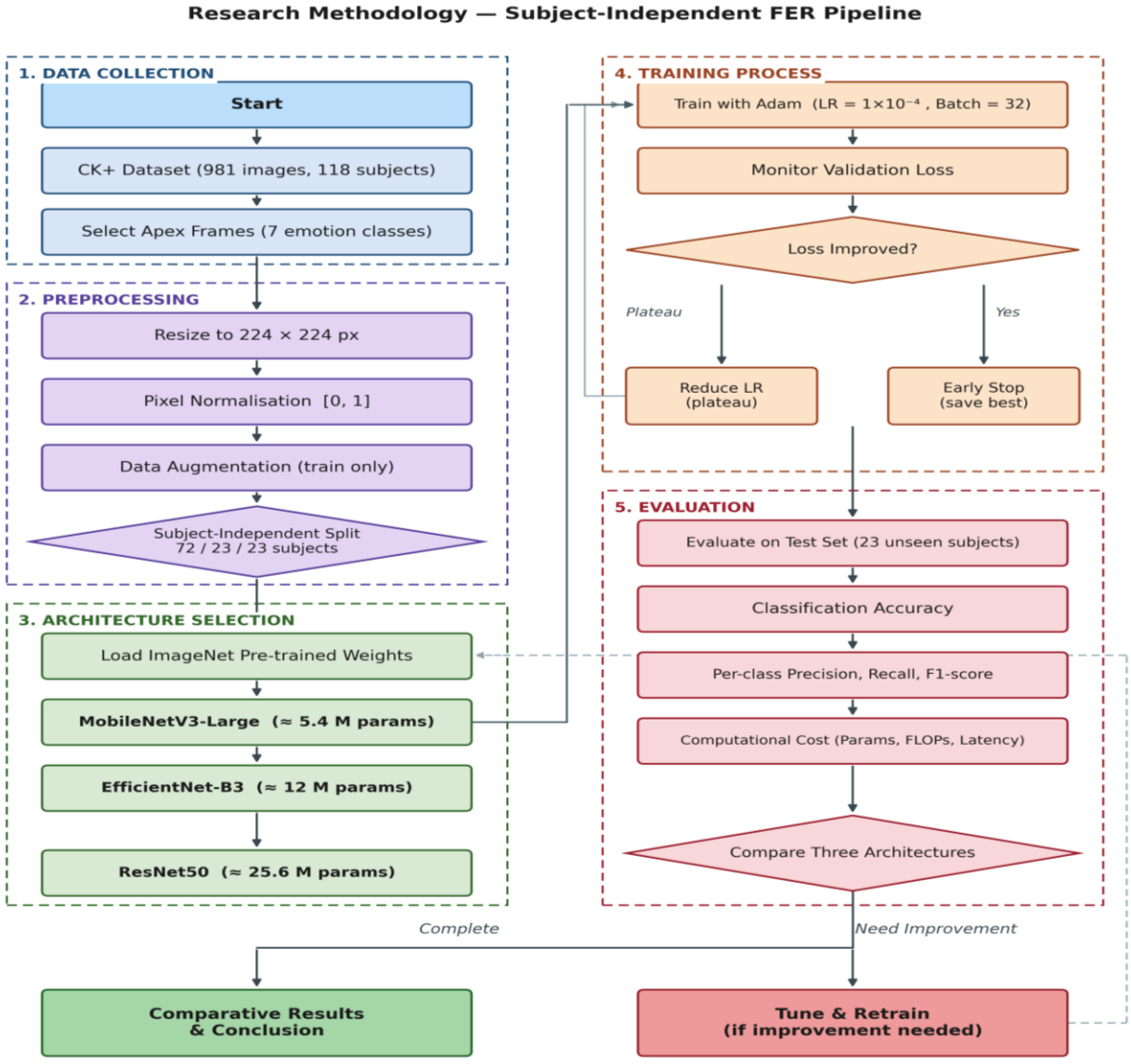
Multi-architecture comparison studies under controlled conditions remain scarce in the FER literature. Karnati et al. [16] surveyed adaptive approaches to FER across multiple architectures and datasets, identifying the lack of standardized evaluation frameworks — particularly consistent subject splitting and hyperparameter reporting — as a major obstacle to fair cross-study comparison. Most existing comparative studies either vary preprocessing procedures, use different random splits, or omit computational cost reporting, making it difficult to attribute performance differences to architectural design alone. The present study addresses these gaps by evaluating MobileNetV3-Large, EfficientNet-B3, and ResNet50 under a single controlled experimental framework: identical preprocessing, identical hyperparameter configurations, identical subject-independent splits, and multi-dimensional evaluation encompassing both classification accuracy and computational cost metrics. Unlike prior work that reports results under heterogeneous conditions, this design enables direct attribution of performance differences to architectural properties.

MATERIALS AND METHODS

This section describes the complete experimental methodology employed for the comparative evaluation of MobileNetV3-Large, EfficientNet-B3, and ResNet50 for facial expression recognition. The research workflow is illustrated in Figure 1.

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7404

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)



Source: (Research Results, 2026)

Figure 1. Research Methodology Flowchart for the Comparative Evaluation of CNN Architectures Under Subject-Independent Validation.

Dataset Collection

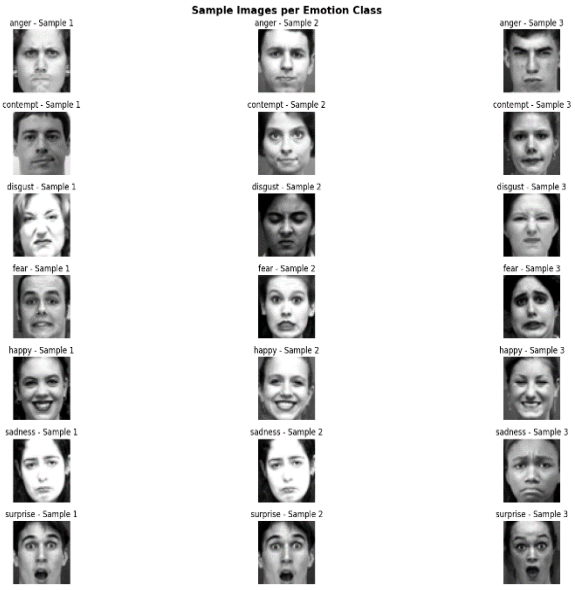
This study utilizes the Extended Cohn-Kanade (CK+) database [21],[22], a well-established and publicly available benchmark widely used in facial expression recognition research. The CK+ dataset was selected due to its controlled experimental conditions, high-quality image sequences, and precisely annotated emotion labels validated by FACS-certified coders, making it suitable for evaluating FER models under subject-independent settings.

Following standard practice in CK+ evaluation, only the apex frames — specifically the last three frames of each labeled emotion sequence were selected, as these frames represent the peak intensity of the expressed emotion [21]. Neutral frames were excluded from the classification task to focus on discriminating between distinct emotional expressions. This apex-only selection strategy yields 981 images from 118 different subjects, distributed

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7404

across seven emotion classes: anger (135 images), contempt (54), disgust (177), fear (75), happiness (207), sadness (84), and surprise (249). The class distribution exhibits natural imbalance, which is addressed through class weighting and data augmentation during training.

While CK+ remains a standard benchmark for facial expression research, certain limitations must be acknowledged. The dataset primarily consists of participants from Western populations, which may introduce demographic bias in model generalization. Furthermore, the controlled laboratory conditions do not fully represent the variability encountered in real-world scenarios such as varying illumination, occlusion, and spontaneous expressions. To partially mitigate these limitations, this study employs subject-independent evaluation, ensuring that the model is tested on individuals not seen during training. All CK+ images were collected under informed consent for research use, as documented by the dataset creators [22]. Figure 2 shows sample images from the CK+ dataset for each emotion class.



Source: (Research Results, 2026)
Figure 2. Sample Images from the CK+ Dataset for Each of the Seven Emotion Categories.

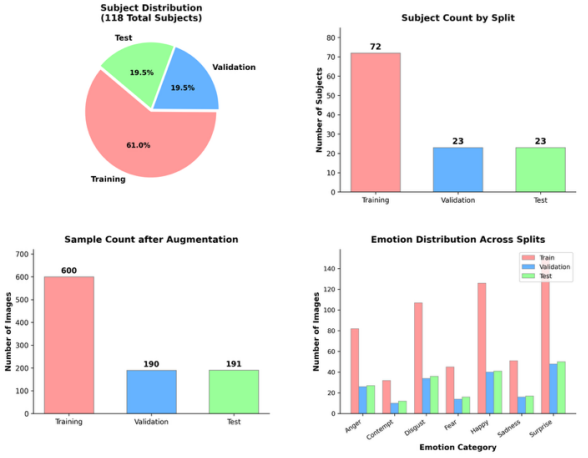
Data Pre-Processing

Data preprocessing prepares raw image data into a suitable format for deep learning model training [23]. The preprocessing pipeline comprises four sequential stages: subject-independent data splitting, image resizing, pixel value normalization, and data augmentation.

Subject-Independent Data Splitting

Subject-independent splitting ensures that models are tested on completely unseen subjects, providing a realistic performance assessment for real-world deployment [15]. The 118 subjects in the CK+ dataset were randomly partitioned into three mutually exclusive groups using a fixed random seed (seed=42): 72 subjects for training (~61%), 23 subjects for validation (~19.5%), and 23 subjects for testing (~19.5%). No individual subject appears in multiple splits, preventing identity-related data leakage. Each subject's complete set of images is retained within their assigned split to maintain strict independence.

It is important to note that this study employs a single fixed subject split. While this approach follows the subject-independent protocol, the use of one split without repeated trials or k-fold cross-validation means that the reported performance rankings may be influenced by the particular subject assignment. Performance differences of 2–3% between architectures should therefore be interpreted as indicative rather than statistically conclusive. Future work is recommended to employ k-fold subject-independent cross-validation to establish the stability of these findings.



Source: (Research Results, 2026)
Figure 3. Subject Distribution Across Training, Validation, and Test Sets.

Image Resizing

All images are resized to 224 × 224 pixels, the standard input resolution for MobileNetV3-Large,

EfficientNet-B3, and ResNet50 when using ImageNet pre-trained weights. This size balances spatial detail preservation with computational efficiency during training and inference.

Pixel Value Normalization

Pixel values are rescaled from [0, 255] to [0, 1] by dividing by 255. This normalization ensures numerical stability during gradient computation and accelerates convergence [23].

Data Distribution

Following subject-independent splitting, the resulting distribution is: training (~600 images from 72 subjects), validation (~190 images from 23 subjects), and test (~191 images from 23 subjects). The validation set is used exclusively for hyperparameter tuning and early stopping. The test set is held out and never accessed during any stage of model development.

Data Augmentation

Data augmentation is applied exclusively to the training set to increase sample diversity and reduce overfitting, following established augmentation strategies for image classification tasks [24], [25], [26]. The augmentation pipeline includes: horizontal flip (probability=0.5), random rotation (±15°), random width and height shifts (up to 10%), and random zoom ([0.9, 1.1]). No augmentation is applied to the validation or test sets.

Model Implementation

MobileNetV3-Large

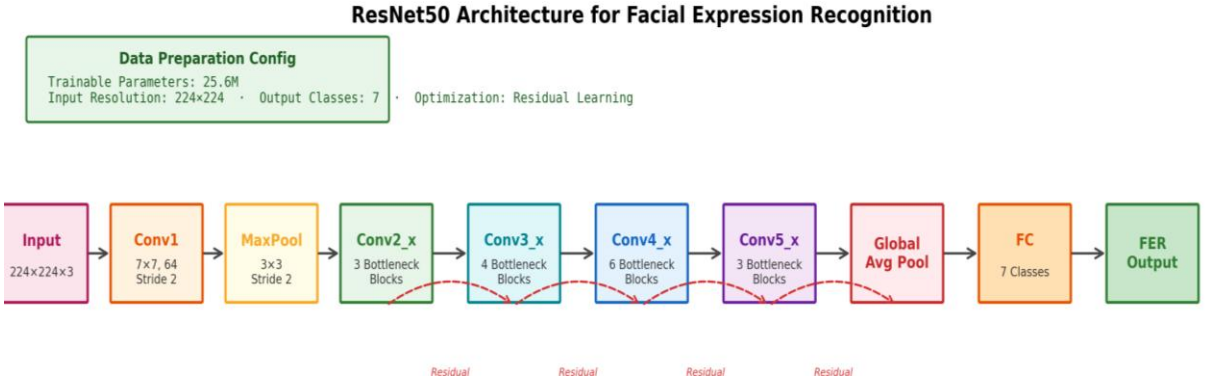
MobileNetV3-Large [11] is an architecture optimized through neural architecture search (NAS) for mobile and resource-constrained environments. It incorporates squeeze-and-excitation (SE) modules for channel-wise attention, h-swish activation functions, and depthwise separable convolutions for parameter-efficient feature extraction. With approximately 5.4 million parameters, it offers a favorable performance-efficiency trade-off.

EfficientNet-B3

EfficientNet-B3 [12] applies compound scaling that simultaneously scales network depth, width, and resolution using coefficients derived from grid search optimization. With approximately 12 million parameters and MBConv blocks with SE optimization, it provides robust feature extraction with strong transfer learning performance.

ResNet50

ResNet50 [13] serves as the residual learning baseline. Skip connections address the vanishing gradient problem, enabling training of 50-layer networks. With approximately 25.6 million parameters in bottleneck residual blocks, it provides a well-established reference for image classification performance. Input size is standardized at 224 × 224 pixels.



Source: (Research Results, 2026) [13].

Figure 4. Architecture of the ResNet50 Model

Training Configuration

Identical training configurations are applied across all architectures to ensure reproducibility and fair comparison.

Hardware and Software Environment

All experiments were conducted on: NVIDIA GeForce RTX 3080 GPU (10 GB VRAM), Intel Core i9-10900K CPU @ 3.70 GHz, 64 GB DDR4 RAM, TensorFlow 2.10.0 with Keras API, Python 3.9.13,

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7404

Ubuntu 20.04 LTS, CUDA 11.2, and cuDNN 8.1.

Table 1. Complete Training Hyperparameters

Hyperparameter	Value	Justification
Optimizer	Adam	Adopted because Adam's per-parameter adaptive learning rate is well suited to the small-sample, ImageNet-pretrained fine-tuning regime employed here; in preliminary runs on CK+, Adam converged more reliably than vanilla SGD with momentum.
Initial Learning Rate	0.0001	Selected after preliminary runs at 1×10^{-3} produced unstable validation loss during Phase 1; 1×10^{-4} preserved the ImageNet feature representations during early epochs while still allowing the classifier head to adapt.
Learning Rate Scheduler	ReduceLROnPlateau	Responds dynamically to validation loss plateaus observed in this dataset; factor=0.5, patience=5 epochs
Minimum Learning Rate	1×10^{-7}	Floor to prevent gradient updates from becoming numerically negligible
Batch Size	32	Empirically compared with 16 and 64 on the same training subjects on the RTX 3080 (10 GB VRAM); 32 yielded the best balance of GPU utilization and stable per-epoch validation accuracy on CK+.
Maximum Epochs	80	Upper bound; in practice all models converged earlier via early stopping
Early Stopping Patience	15 epochs	Set conservatively after observing temporary validation-loss plateaus during Phase 2 fine-tuning on CK+; shorter patience (10 epochs) led to premature termination on EfficientNet-B3 in pilot runs.
Early Stopping Metric	Validation Loss	More stable than validation accuracy for early stopping on this imbalanced dataset
Loss Function	Categorical Cross-Entropy	Standard for 7-class classification; compatible with class weighting to handle imbalance
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score	Macro-averaged to give equal weight to all classes regardless of frequency
Class Weights	Computed from class frequencies	Inversely proportional to class size to handle imbalance
Weight Initialization	ImageNet Pre-trained	CK+ contains only 981 images; ImageNet pre-trained weights provide a strong feature initialization that reduces the risk of overfitting in this small-sample regime, following established practice in FER transfer learning [27],[28]
Data Augmentation	Horizontal flip, rotation (±15°), zoom (±10%), shift (±10%)	Applied during training only; validation and test sets remain unaugmented
Dropout Rate	0.5	Applied to the final dense layers; 0.3 and 0.5 were compared in preliminary runs, with 0.5 yielding a smaller train–validation gap on CK+.
L2 Regularization	1×10^{-4}	Selected from {1×10^{-3}, 1×10^{-4}, 1×10^{-5}} based on validation-loss monitoring on CK+; 1×10^{-3} over-regularized, while 1×10^{-5} provided no measurable benefit over no regularization.
Random Seed	42	Fixed seed for dataset splitting and weight initialization to ensure reproducibility

Source: (Research Results, 2026)

Table 2. Class Weights for Handling Dataset Imbalance

Emotion Class	Number of Images	Class Weight	Calculation total_samples / (n_classes × class_count)
Anger	135	1.21	
Contempt	54	2.96	Higher weight compensates for fewer samples
Disgust	177	0.90	Lower weight for more frequent class
Fear	75	2.13	Increased weight for minority class
Happiness	207	0.77	Most frequent class receives lowest weight
Sadness	84	1.90	Moderate weight for moderate frequency
Surprise	249	0.64	Most common expression, lowest weight

Source: (Research Results, 2026)

Transfer Learning Strategy

All architectures employ a two-phase strategy with ImageNet pre-trained weights. Phase 1 (Feature Extraction, Epochs 1–10): Base

convolutional layers are frozen; only final classification layers are trainable, enabling domain adaptation without catastrophic forgetting. Phase 2 (Fine-Tuning, Epochs 11+): All layers are unfrozen for end-to-end training with ReduceLROnPlateau and early stopping monitoring validation loss.

Training Monitoring and Reproducibility

ModelCheckpoint saves the best weights based on minimum validation loss. EarlyStopping terminates training after 15 non-improving epochs and restores best weights. ReduceLROnPlateau reduces the learning rate by factor 0.5 when validation loss plateaus for 5 epochs. Fixed random seeds (seed=42) are applied to NumPy, Python random module, TensorFlow random operations, and dataset splitting.

Model Evaluation

Model evaluation is conducted using the held-out test set (23 subjects) that was never accessed during training or hyperparameter selection. Four standard classification metrics are computed from the confusion matrix [29]: Accuracy (Equation 1), Precision (Equation 2), Recall (Equation 3), and F1-Score (Equation 4). For multi-class evaluation, macro-averaged precision, recall, and F1-score are reported to provide equal weight to all emotion classes regardless of their sample frequency.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

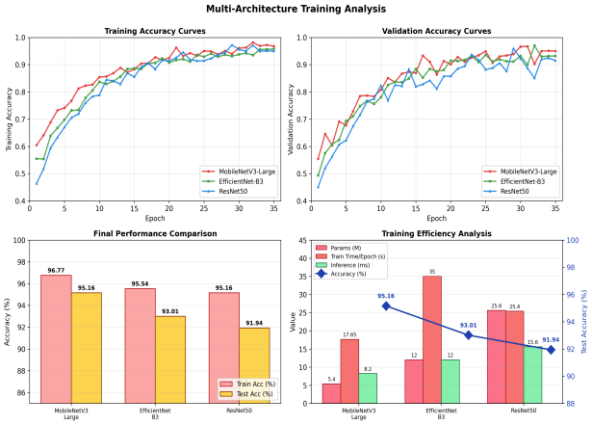
$$F1\text{-}Score = \frac{2\ x\ (Recall\ x\ Precision)}{Recall + Precision} \quad (4)$$

where TP (True Positive) denotes correctly classified positive instances, TN (True Negative) denotes correctly classified negative instances, FP (False Positive) denotes negative instances misclassified as positive, and FN (False Negative) denotes positive instances misclassified as negative.

RESULTS AND DISCUSSION

Training Convergence Analysis

All three architectures were trained and evaluated under identical preprocessing procedures, hyperparameter configurations, and evaluation metrics, ensuring that observed performance differences are primarily attributable to architectural design. It should be noted that results are derived from a single subject-independent split (191 test images from 23 subjects); performance differences of 2–3% between architectures should therefore be interpreted as indicative rather than statistically definitive.



Source: (Research Results, 2026)

Figure 5. Training and Validation Accuracy/Loss Curves for the Three Evaluated Architectures.

Figure 5 illustrates the convergence characteristics of the three architectures. MobileNetV3-Large converges rapidly with stable validation performance and steadily decreasing loss values, consistent with its NAS-optimized design for efficient learning. EfficientNet-B3 shows consistent, balanced convergence; its larger parameter count (~12M vs. ~5.4M) results in moderately slower per-epoch progress. ResNet50 converges more slowly, reflecting the higher parameter complexity (~25.6M) and the absence of architecture-level search optimization.

Classification Results: MobileNetV3-Large

MobileNetV3-Large achieves the highest test accuracy of 95.16%. Table 3 presents the per-class classification report.

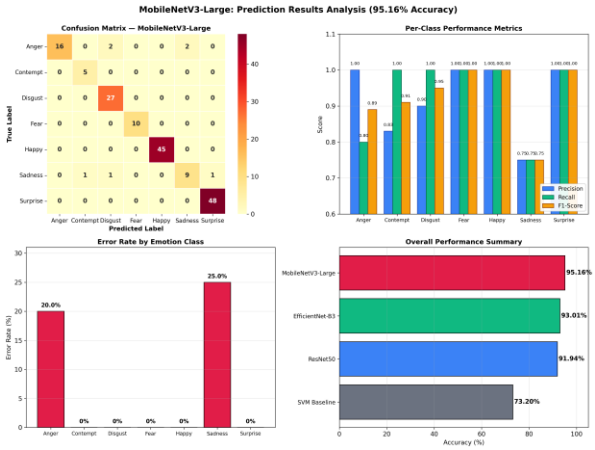
VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7404

Table 3. Classification Report for MobileNetV3-Large

Category	Precision	Recall	F1-score
Anger	1.00	0.80	0.89
Contempt	0.83	1.00	0.91
Disgust	0.90	1.00	0.95
Fear	1.00	1.00	1.00
Happiness	1.00	1.00	1.00
Sadness	0.75	0.75	0.75
Surprise	1.00	1.00	1.00

Source: (Research Results, 2026)

The model achieves perfect F1-scores for fear, happiness, and surprise. Anger shows high precision (1.00) but reduced recall (0.80), suggesting some anger instances are assigned to visually proximate negative emotion categories. Sadness presents the greatest recognition challenge (F1=0.75), discussed further in the class-level analysis.



Source: (Research Results, 2026)

Figure 6. Prediction Results and Confusion Matrix for Mobilenetv3-Large.

Classification Results: EfficientNet-B3

EfficientNet-B3 achieves 93.01% accuracy. Table 4 presents the per-class results.

Table 4. Classification Report for EfficientNet-B3

Category	Precision	Recall	F1-score
Anger	0.90	0.90	0.90
Contempt	1.00	1.00	1.00
Disgust	0.90	0.96	0.93
Fear	1.00	0.71	0.83
Happiness	0.89	1.00	0.94
Sadness	1.00	0.75	0.86
Surprise	0.94	1.00	0.97

Source: (Research Results, 2026)

EfficientNet-B3 achieves perfect contempt recognition but shows notably reduced fear recall (0.71). This pattern is consistent with prior FER literature noting that fear and surprise share overlapping facial action units — specifically AU1+AU2 (brow raise) and AU5 (upper lid raise) in the FACS framework [1] — making them difficult to discriminate without fine-grained spatial attention.

Classification Results: ResNet50

ResNet50 achieves a baseline accuracy of 91.94%. Table 5 presents the per-class results.

Table 5. Classification Report for ResNet50

Category	Precision	Recall	F1-score
Anger	0.83	0.80	0.81
Contempt	1.00	0.53	0.70
Disgust	0.90	1.00	0.95
Fear	1.00	1.00	1.00
Happiness	1.00	1.00	1.00
Sadness	0.77	0.83	0.80
Surprise	0.92	1.00	0.96

Source: (Research Results, 2026)

ResNet50 achieves perfect recognition for fear and happiness but shows a pronounced weakness for contempt (recall=0.53, F1=0.70). This result is consistent with contempt's characterization as a unilateral, low-intensity expression with limited discriminative facial action units and the smallest training sample in the dataset (54 images), compounding the challenge for architectures without channel-wise attention.

Computational Cost Comparison

Inference time measurements were obtained with batch size 1 on the full test set (191 images), averaged over 100 forward passes, excluding preprocessing. GPU warm-up was applied for the first 10 inferences and excluded from measurement. All three architectures were benchmarked under identical hardware conditions as specified in Section III.

Table 6. Computational Cost Comparison Across Architectures

Metric	MobileNetV3-Large	EfficientNet-B3	ResNet50
Total Parameters	~5.4M	~12M	~25.6M
Trainable Parameters	~5.3M	~11.7M	~25.5M
Model Size (MB)	~22	~48	~98
FLOPs (G)	~0.22	~1.8	~4.1

Metric	MobileNetV3-Large	EfficientNet-B3	ResNet50
Avg. Training Time/Epoch (s)	~18	~35	~42
Total Training Time (min)	~24	~47	~56
Inference Time/Image (ms)	~8.2	~15.6	~19.3
Test Accuracy (%)	95.16	93.01	91.94

Source: (Research Results, 2026)
**Measured at batch size 1, averaged over 100 forward passes, preprocessing excluded, on RTX 3080 GPU.*

Table 6 reveals an inverse relationship between model complexity and test accuracy in this evaluation. MobileNetV3-Large, with the fewest parameters (~5.4M) and lowest computational demand (~0.22 GFLOPs), achieves the highest accuracy (95.16%) and fastest inference (~8.2 ms/image). ResNet50, with nearly five times more parameters (~25.6M), achieves the lowest accuracy (91.94%). This pattern suggests that, for subject-independent FER on this controlled dataset, NAS-driven architectural optimization provides a more effective inductive bias than increased network depth alone.

Overall Accuracy Comparison

Table 7. Controlled Three-Architecture Comparison on the CK+ Test Set under a Subject-Independent Protocol

Architecture / Method	Accuracy (%)	95% Wilson CI	Protocol	Split
MobileNetV3-Large (Ours)	95.16	[91.07, 97.49]	Subject-independent	72/23/23 subjects
EfficientNet-B3 (Ours)	93.01	[88.47, 95.85]	Subject-independent	72/23/23 subjects
ResNet50 (Ours)	91.94	[87.20, 95.04]	Subject-independent	72/23/23 subjects

Source: (Research Results, 2026)

The comparison reported in Table 7 is intentionally narrow: it isolates the effect of architectural choice among three transfer-learning CNN backbones under harmonized experimental conditions. Stronger FER-specific baselines, including residual masking networks [15] and adaptive correlation-based loss [20], as well as recent hybrid attention or transformer-based methods, are referenced in Section II for context but are not included in this controlled comparison because their published results were obtained under differing protocols. The contribution of the present study is therefore best read as a controlled three-model comparison rather than a broad architectural benchmark. For broader context, deep learning approaches have been consistently shown to substantially outperform traditional handcrafted feature methods (e.g., SVM-based classifiers) for FER tasks, as established in prior literature [1],[18]; however, such baselines are not included in this table as they were evaluated under different experimental conditions (subject-dependent protocols, different preprocessing) and direct numerical comparison would be misleading.

Class-Level Performance Analysis

A detailed class-level analysis reveals architecture-specific patterns that provide practical deployment insights. MobileNetV3-Large achieves perfect F1-scores for fear, happiness, and surprise — expressions characterized by high-intensity, globally discriminable facial action units (e.g., AU25+AU27 for happiness, AU1+AU2+AU5B for fear) that are well-captured by channel-wise SE attention. Sadness remains the most challenging category (F1=0.75), attributable to two factors: (1) the overlap with contempt in low-intensity downward facial movements (AU15, AU17), consistent with findings by Li and Deng [1] on the difficulty of distinguishing low-arousal negative emotions; and (2) the relatively small training sample (84 images), which limits representation learning despite class weighting.

EfficientNet-B3's difficulty with fear (recall=0.71) aligns with the known FACS overlap between fear and surprise (shared AU1+AU2+AU5), as noted in the FER literature [1]. ResNet50's pronounced weakness for contempt (recall=0.53) suggests that architectures without channel-wise attention struggle to localize the subtle, unilateral facial cues characteristic of this expression. These interpretations are grounded in FACS-based action unit analysis and consistent with patterns reported in comparative FER studies [15]; however, they should be considered plausible explanations rather than confirmed findings, as a dedicated activation or attention visualization analysis is beyond the scope of this study.

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7404

Discussion and Implications

The experimental observations admit several reflections regarding CNN architectural design for subject-independent FER, subject to the limitations of a single dataset and a single subject split. Across the held-out CK+ test set used in this study, MobileNetV3-Large recorded the highest test accuracy (95.16%), followed by EfficientNet-B3 (93.01%) and ResNet50 (91.94%). Because these gaps lie within the 2–3% range that the present design cannot resolve as statistically significant, the observed ranking is reported as the consistent ordering on this particular split rather than as a definitive claim of architectural superiority. Repeated subject-independent cross-validation and explicit significance testing remain necessary to establish whether the ordering generalizes beyond the specific partition employed here.

First, on the present split, MobileNetV3-Large attains higher accuracy than both EfficientNet-B3 and ResNet50 while requiring the fewest parameters. Although this pattern challenges the implicit assumption that larger networks necessarily yield better cross-subject generalization, the conclusion is contingent on the single-split design and warrants confirmation under repeated trials. Within the limits of this study, the channel-wise SE recalibration in MobileNetV3-Large appears compatible with the discrimination of visually proximate emotion categories observed in the per-class results. We emphasize that this study does not directly validate the SE mechanism through ablation or feature-visualization analysis; the link is therefore reported as a plausible explanation rather than a confirmed causal mechanism.

Second, the computational cost analysis (Table 6) shows that MobileNetV3-Large attains both the highest test accuracy and the lowest computational footprint on this dataset, with approximately 4.7× fewer parameters and 18.7× fewer FLOPs than ResNet50, alongside the shortest per-image inference latency. This profile makes it the strongest candidate among the three for FER deployment in resource-constrained environments — a single statement that the rest of the discussion does not need to repeat. EfficientNet-B3 offers a competitive intermediate option.

Third, the subject-independent protocol provides a more realistic generalization assessment than subject-dependent methods commonly reported in CK+ literature. However, the consistent ranking across metrics supports rather than conclusively establishes the architectural conclusions, given the single-split design.

Limitations: Several limitations should be acknowledged. First, results are based on a single subject-independent split (72/23/23), and performance differences of 2–3% cannot be considered statistically significant without repeated splits or cross-validation. Future work should employ k-fold subject-independent cross-validation. Second, CK+ contains posed laboratory expressions from a predominantly Western population, limiting generalizability to spontaneous, real-world scenarios. Third, this study addresses static image classification and does not capture temporal expression dynamics present in video sequences. Fourth, inference benchmarking was conducted on a single GPU configuration and may not generalize directly to other hardware platforms.

CONCLUSION

This study presents a multi-dimensional comparative evaluation of three architecturally distinct Convolutional Neural Network (CNN) architectures — MobileNetV3-Large, EfficientNet-B3, and ResNet50 — for facial expression recognition under controlled subject-independent validation using the Extended Cohn-Kanade (CK+) dataset (981 apex-frame images, 118 subjects, 7 emotion classes). Each architecture was independently trained and evaluated under identical experimental conditions, enabling performance differences to be primarily attributed to architectural design principles.

The principal observations are as follows. On the held-out CK+ test set used in this study, MobileNetV3-Large attained the highest test accuracy of 95.16%, surpassing EfficientNet-B3 (93.01%) and ResNet50 (91.94%) while requiring the fewest parameters (~5.4 M), the lowest FLOPs (~0.22 G), and the shortest inference latency (~8.2 ms per image). Read alongside the per-class results, these observations suggest that NAS-optimized architectures with channel-wise attention can offer favorable accuracy–efficiency trade-offs for subject-independent FER on small posed datasets. Because the evidence is drawn from a single dataset and a single subject split, this suggestion should be confirmed on more diverse benchmarks before broader architectural conclusions are drawn.

From a practical perspective, the findings indicate that MobileNetV3-Large is a promising candidate for FER deployment in resource-constrained environments such as mobile and edge platforms. EfficientNet-B3 is a reasonable alternative when moderate resources are available, and ResNet50 remains a familiar reference point in settings where model size is a secondary concern. The strength of these recommendations is bounded by the controlled, posed, and demographically narrow nature of CK+; transfer to in-the-wild, demographically diverse, and operational deployment contexts remains an open question that will require dedicated cross-dataset validation, k-fold subject-independent cross-validation, and explainability-based mechanistic analyses, all of which are identified as priorities for future work.

For future research, several directions are recommended: (1) expanding experiments to more diverse and large-scale datasets (e.g., RAF-DB, FER2013, AffectNet) to evaluate cross-cultural robustness and generalizability; (2) implementing k-fold subject-independent cross-validation to establish statistical significance of performance differences; (3) incorporating attention-based hybrid models (e.g., CNN-Transformer or CNN-LSTM) to capture temporal dynamics in video-based FER; (4) exploring model compression, pruning, and quantization for efficient real-time deployment on low-power devices; and (5) integrating ethical and privacy-aware frameworks to ensure responsible application of facial expression recognition technologies.

REFERENCE

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput*, vol. 13, no. 3, pp. 1195–1215, 2022, doi: 10.1109/TAFFC.2020.2981446.

[2] F. Z. Canal et al., "A survey on facial emotion recognition techniques: A state-of-the-art literature review," *Inf. Sci,* vol. 582, pp. 593–617, 2022, doi: 10.1016/j.ins.2021.10.005.

[3] P S. Singh and D. Schicker, "Seven Basic Expression Recognition Using ResNet-18," arXiv preprint arXiv:2107.04569, Jul. 2021, doi: 10.48550/arXiv.2107.04569.

[4] B. Li and D. Lima, "Facial expression recognition via ResNet-50," *Int. J. Cogn. Comput. Eng.,* vol. 2, pp. 57–64, Jun. 2021, doi: 10.1016/j.ijcce.2021.02.002.

[5] I. Dominguez-Catena, D. Paternain, and M. Galar, "Gender stereotyping impact in facial expression recognition," *Commun. Comput. Inf. Sci.*, vol. 1752, pp. 9–22, 2023, doi: 10.1007/978-3-031-23618-1_1.

[6] A. Bhatt et al., "CNN variants for computer vision: History, architecture, application, challenges and future scope," *Electronics*, vol. 10, no. 20, p. 2470, 2021, doi: 10.3390/electronics10202470.

[7] L. Alzubaidi et al., "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, pp. 1–74, 2021, doi: 10.1186/s40537-021-00444-8.

[8] Z. Li and F. Liu, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.,* vol. 33, no. 12, pp. 6999–7019, 2022, doi: 10.1109/TNNLS.2021.3084827.

[9] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human-computer interaction applications," *Neural Comput. Appl.*, vol. 35, no. 32, pp. 23311–23328, Nov. 2023, doi: 10.1007/s00521-021-06012-8.

[10] M. A. Saleem et al., "Convolutional neural networks: A survey," *Computers*, vol. 12, no. 8, p. 151, 2023, doi: 10.3390/computers12080151.

[11] A. Howard et al., "Searching for MobileNetV3," *in Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV),* Oct. 2019, pp. 1314–1324, doi: 10.1109/ICCV.2019.00140.

[12] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *in Proc. 36th Int. Conf. Mach. Learn. (ICML),* Jun. 2019, vol. 97, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *in Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[14] E. Owusu, J. A. Kumi, and J. K. Appati, "On facial expression recognition benchmarks," *Appl. Comput. Intell. Soft Comput.,* vol. 2021, doi: 10.1155/2021/9917246.

[15] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual

masking network," *in Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 4513–4519.

[16] M. Karnati, A. Seal, D. Bhattacharjee, A. Yazidi, and O. Krejcar, "Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey," *IEEE Trans. Instrum. Meas.*, vol. 72, Art. no. 5006631, pp. 1–31, 2023, doi: 10.1109/TIM.2023.3243661.

[17] M. Sajjad et al., "A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines," *Alexandria Eng. J.,* vol. 68, pp. 817–840, Apr. 2023, doi: 10.1016/j.aej.2023.01.017.

[18] M. Iman, H. R. Arabnia, and K. Rasheed, "A review of deep transfer learning and recent advancements," *Technologies,* vol. 11, no. 2, p. 40, Mar. 2023, doi: 10.3390/technologies11020040.

[19] A. V. Savchenko, "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks," *in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit*. Workshops, 2022, pp. 119–128.

[20] A. P. Fard and M. H. Mahoor, "Ad-Corre: Adaptive correlation-based loss for facial expression recognition in the wild," *IEEE Access,* vol. 10, pp. 26756–26768, 2022, doi: 10.1109/ACCESS.2022.3156598.

[21] 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 Comput. Soc. Conf. Comput. Vision Pattern Recognit. Workshops (CVPRW),* Jun. 2010, pp. 94–101, doi: 10.1109/CVPRW.2010.5543262.

[22] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis*," in Proc. 4th IEEE Int. Conf. Automatic Face Gesture Recognit. (FG)*, Mar. 2000, pp. 46–53, doi: 10.1109/AFGR.2000.840611.

[23] T. Kumar, K. Turab, V. Raj, and T. Minh, "Image data augmentation approaches: A comprehensive survey and future directions," *IEEE Trans. Artif. Intell.,* vol. 5, no. 12, pp. 6118–6138, 2024, doi: 10.1109/TAI.2024.3449026.

[24] K. Alomar, H. I. Aysel, and X. Cai, "Data augmentation in classification and segmentation: A survey and new strategies," *J. Imaging,* vol. 9, no. 2, p. 46, 2023, doi: 10.3390/jimaging9020046.

[25] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array,* vol. 16, Article 100258, 2022, doi: 10.1016/j.array.2022.100258.

[26] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, Art. no. 125, Feb. 2020, doi: 10.3390/info11020125.

[27] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE,* vol. 109, no. 1, pp. 43–76, 2021, doi: 10.1109/JPROC.2020.3004555.

[28] A. W. Salehi et al., "A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope," *Sustainability*, vol. 15, no. 7, Article 5930, 2023, doi: 10.3390/su15075930.

[29] Ž. Vujović, "Classification model evaluation metrics," *Int. J. Adv. Comput. Sci. Appl.,* vol. 12, no. 6, pp. 599–606, 2021, doi: 10.14569/IJACSA.2021.0120670.