Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Facial Expression Recognition for Monitoring Learning Satisfaction in Smart Learning Environments Using MobileNetV2

Sandy Radytia [1)], Ucuk Darusalam[2)*]
[1)2)]Magister Teknologi Informasi, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional
[1)]sa@ail.com, [2)]ucuk.darusalam@civitas.unas.ac.id

Submitted :Nov 5, 2025 | **Accepted** : Nov 25, 2025 | **Published** : Jan 04, 2026

Abstract: This study develops a lightweight, privacy-aware Facial Expression Recognition (FER) framework to monitor learning satisfaction in Smart Learning Environments (SLEs). Using MobileNetV2 with a two-stage training scheme on the FER2013 dataset and evaluated on **35,000 test samples**, the system addresses two main questions: (1) how effectively a customized MobileNetV2 recognizes core student expressions under authentic classroom conditions, and (2) how temporal aggregation and confidence calibration improve the stability of a Learning Satisfaction Index (LSI). The model achieves **0.39 accuracy** and **0.34 macro-F1**, with strong performance for happy, neutral, and surprise, while challenges remain for fear–surprise and neutral–sad. Temporal smoothing reduces prediction noise and enhances the reliability of LSI signals for instructional decision-making. The findings highlight practical **implications for education**, particularly in supporting real-time formative assessment and improving teachers' awareness of student engagement through privacy-preserving, on-device affect monitoring.

Keywords: FER; MobileNetV2; SLE; LSI; Edge-computing

INTRODUCTION

Smart Learning Environments (SLEs) increasingly demand timely, evidence-based insight into learners' experiences as they unfold. Post-class surveys and periodic questionnaires arrive late, introduce response bias, and fail to capture minute-to-minute affective dynamics that matter for intervention. Facial Expression Recognition (FER) provides a non-intrusive proxy for emotion that can complement performance logs, clickstream data, and interaction traces (Sholikah et al., 2024), (Johnson et al., 2025), (Sun et al., 2025). Recent advances in deep learning make it feasible to detect subtle micro-expressions with models that are compact enough for classroom devices. At the same time, the growing availability of affordable edge hardware enables on-site processing without heavy reliance on cloud infrastructure. Integrating affective signals into SLEs promises richer feedback loops, personalized support, and ultimately sustained learning satisfaction (Luo & Huang, 2023).

Authentic classrooms, however, are unruly spaces where clean assumptions break quickly. Non-uniform lighting, off-axis camera placement, partial occlusion from masks or hands, and spontaneous motion all degrade visual signals. Class imbalance—neutral faces vastly outnumbering joy, surprise, or frustration—complicates equitable training and evaluation across emotion categories. Distribution gaps between public benchmarks and local classroom footage diminish out-of-distribution generalization. Constrained hardware imposes strict latency and memory budgets while accuracy expectations remain high. Ethical and privacy considerations require data minimization, anonymization, and role-based access from the very beginning of system design. Finally, relying on overall accuracy alone can hide poor performance on minority classes, motivating the use of balanced metrics and fine-grained diagnostics.

The proposed method adopts MobileNetV2 as the backbone, chosen for its favourable accuracy-to-complexity ratio on resource-constrained devices (Hindarto, 2023a), . The processing pipeline includes face detection and multi-object tracking to stabilize inputs, illumination normalization and landmark alignment to reduce nuisance variance, and targeted augmentation that emulates classroom conditions. Transfer learning initializes the network with pre-trained weights, followed by context-specific fine-tuning on classroom-relevant data. Class imbalance is addressed using cost-sensitive learning or focal-style losses that emphasize difficult and underrepresented samples. Probability calibration and regularization are applied to temper overconfidence and improve robustness.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Evaluation covers accuracy, macro-F1, per-class AUROC, and confusion matrices across varied scenarios to ensure performance is fairly distributed among emotions (Li et al., 2025).

Building on this pipeline, we design a real-time system that fuses on-device inference with temporal aggregation to produce a Learning Satisfaction Index. Frame-level predictions pass through windowed smoothing and confidence thresholds to suppress spurious fluctuations from noise and camera jitter. Prioritizing edge execution lowers latency and helps ensure sensitive data remain within the classroom boundary. The data management module follows a strict minimization principle: when feasible, it stores anonymized embeddings or summary statistics rather than raw images. A teacher-facing dashboard visualizes satisfaction trends, early-warning indicators, and interpretable cues for classroom action. Seamless integration with the SLE links affective signals to learning activities and formative outcomes, so the system not only classifies emotions but also synthesizes them into operational insight.

This research aims to design, implement, and rigorously evaluate a lightweight, privacy-aware FER system that reliably monitors learning satisfaction in SLEs. Research Question 1 asks: to what extent can a customized MobileNetV2 recognize core student expressions under real classroom conditions compared with heavier baselines? Research Question 2 asks: how do temporal aggregation and confidence calibration improve the reliability of the Learning Satisfaction Index for timely instructional interventions? The study targets deployment-grade performance under realistic constraints, emphasizing latency, memory footprint, and ethical handling of data. In addition, the evaluation framework probes domain shift, robustness to occlusion and lighting changes, and the stability of outputs over time—factors essential for trustworthy classroom adoption.

The contributions are sixfold. First, we deliver a MobileNetV2-based FER model optimized for edge deployment through fine-tuning, measured quantization, and, where appropriate, structured pruning. Second, we provide a targeted preprocessing and augmentation pipeline robust to lighting variation, viewpoint changes, occlusion, and classroom-typical class imbalance. Third, we introduce temporal aggregation and confidence calibration that yield a more stable and reliable Learning Satisfaction Index than naïve frame-wise voting. Fourth, we propose a balanced evaluation protocol featuring macro-F1, per-class AUROC, confusion analyses, and ablation studies to isolate each component's impact. Fifth, we implement an end-to-end integration and dashboard that translate FER outputs into actionable pedagogical signals. Sixth, we outline privacy-preserving governance—data minimization, anonymized features, and controlled access—plus practical deployment guidelines for real-world classrooms.

LITERATURE REVIEW

Facial Expression Recognition (FER) has been increasingly adopted in educational settings as a means to understand students' affective states and support data-driven instructional decisions. Tonguç and Ozkara demonstrated that students' emotions can be automatically inferred during classroom activities, revealing patterns in both positive and negative affective responses (Tonguç & Ozaydın Ozkara, 2020). Similarly, the Student Emotion Recognition System (SERS) by Krithika and Priya provides real-time emotional feedback in e-learning environments (Krithika & Priya, 2016). These studies highlight the pedagogical value of FER as a complementary source of information alongside behavioral logs and performance metrics. Despite its potential, most educational FER systems rely on frame-level predictions that fluctuate rapidly due to variations in lighting, pose, or occlusion. These noisy outputs limit the interpretability and direct usefulness of FER for formative assessment. Moreover, existing studies rarely translate FER signals into pedagogically meaningful indices, such as satisfaction or engagement scores that educators can act upon.

Lightweight architectures such as A-MobileNet (Nan et al., 2022) demonstrate that compact CNNs can achieve competitive recognition accuracy while meeting low-latency constraints a key requirement for deployment on classroom edge devices. Manalu and Rifai (2024) showed that combining CNN backbones like MobileNetV2 or InceptionV3 with RNN layers can better capture temporal patterns in video-based emotion recognition. Their results indicate that temporal modeling helps reduce per-frame prediction instability. Although lightweight FER models exist, many implementations do not prioritize privacy and still depend on cloud-based computation. Furthermore, research on lightweight FER has primarily focused on accuracy, with limited attention to inference stability, calibration, and real-time responsiveness—factors crucial for classroom deployment.

Temporal modeling approaches such as RNNs, sliding windows, or exponential moving averages have been used to reduce volatility in video-based emotion recognition. Recent studies also emphasize the importance of probability calibration, ensuring that model confidence reliably reflects prediction uncertainty (Sun et al., 2025). Such techniques can substantially improve the consistency of FER outputs under real-world conditions. Most existing studies treat temporal smoothing merely as a technical refinement and rarely connect it to a higher-level pedagogical construct**.** To date, no research has systematically integrated smoothing and calibration to build a Learning Satisfaction Index (LSI) or similar aggregate metric that can support real-time instructional decisions.

Research on Smart Learning Environments increasingly highlights the need for ethical, privacy-preserving data practices. Tabuenca et al. (2024) underscored the importance of sustainable and responsible AIoT systems in

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

classrooms, while Sajjad et al. (2023) pointed to challenges in domain generalization and the need for anonymization when FER is applied in public or educational spaces. Despite these concerns, most FER systems in educational contexts still rely on cloud-based pipelines, exposing biometric data to security and privacy risks.

METHOD

This research method is designed to develop and evaluate a lightweight, accurate, and ready-to-implement MobileNetV2-based Facial Expression Recognition system in a Smart Learning Environment. This study uses the FER2013 corpus, which contains 35.887 grayscale images categorized into seven basic emotions (angry, disgust, fear, happy, neutral, sad, surprise). The dataset is widely used as a benchmark for FER, yet it presents challenges such as imbalance and low resolution that mirror real-world classroom conditions. Preprocessing is designed to reduce nuisance variations while maintaining real-time feasibility. The pipeline includes Face Detection & Multi-Object Tracking, Face Alignment, Illumination Normalization, Image Resizing to 128×128 (RGB), and Targeted Augmentation..The workflow includes face preprocessing (detection, alignment, illumination normalization) and directed augmentation to mimic changing real-world classroom conditions. The model was trained using a two-stage scheme focusing on head classification and then fine-tuning part of the backbone layer accompanied by class imbalance handling, regularization, and label smoothing. Performance evaluation used balanced metrics such as accuracy, macro F1, AUROC per class, and confusion matrix analysis in various lighting and occlusion scenarios. For implementation readiness, the system is integrated with temporal aggregation and probability calibration, then exported to TFLite format with quantization to meet latency and privacy constraints at the edge device..

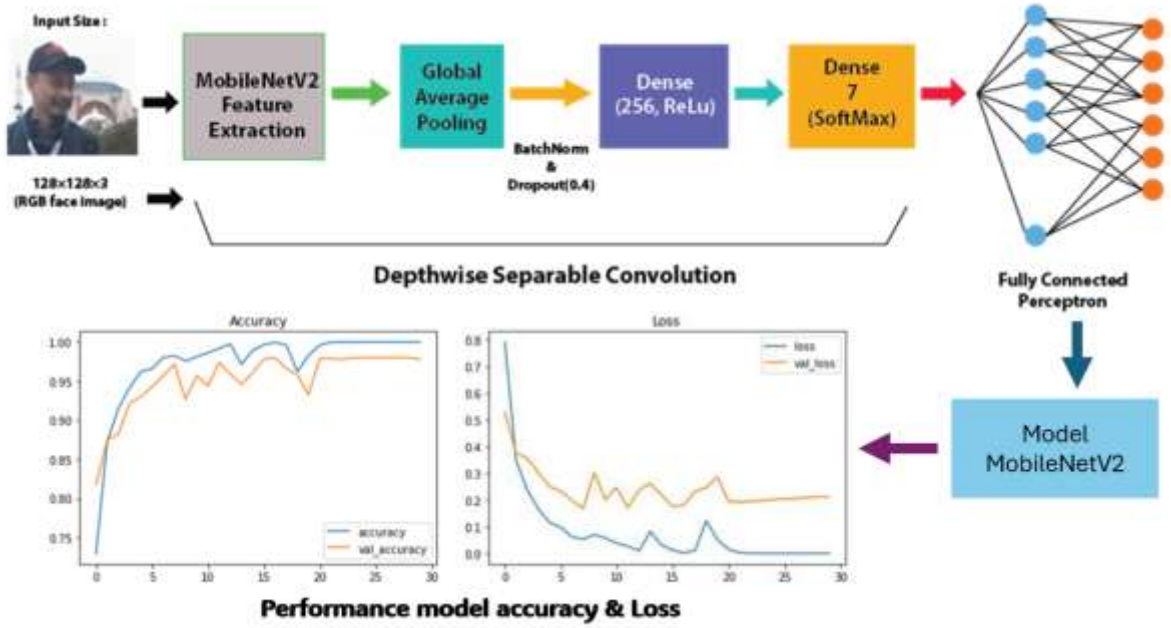

Fig 1. Proposed method MobileNetV2

Figure 1 illustrates the proposed MobileNetV2-based pipeline for facial expression recognition. Pre-cropped and aligned face images serve as inputs, which are passed through MobileNetV2 to extract compact, discriminative feature maps. A Global Average Pooling layer condenses these maps into a robust vector representation, limiting overfitting without a large flatten layer. The vector is processed by a 256-unit Dense layer with ReLU optionally paired with batch normalization and dropout to learn non-linear class boundaries. The final output layer is a seven-node SoftMax that returns class probabilities for the target expressions, suitable for direct decisions or temporal aggregation in video settings. This design balances accuracy and efficiency, making it practical for edge deployment within smart learning environments.

This study employs a two-stage experimental scheme for Facial Expression Recognition (FER) on the FER2013 corpus. Stage one performs feature extraction by freezing the MobileNetV2 backbone; stage two fine-tunes the network by unfreezing a subset of the final blocks to adapt to the target classes. FER2013 (seven emotions) is split into train/validation via a directory layout; images are resized to 128×128 RGB, normalized, and augmented moderately (rotation, shift, zoom, brightness) to improve robustness under webcam conditions. Class imbalance is mitigated with class weights, and fairness is emphasized through macro-F1 and per-class recall. The model uses MobileNetV2 (α=0.5; include top=False) as the backbone with a classification head: Global Average Pooling → Batch Normalization → Dropout(0.4) → Dense(256, ReLU) → Batch Normalization → Dropout(0.4) → Dense(7,

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Softmax), chosen for a favourable accuracy–latency profile on CPUs and smooth conversion to TFLite for real-time use.

Training follows deep-learning best practices in two phases. During feature extraction, only the head is trained with Adam, early stopping, checkpoints, and class weighting to curb overfitting and class bias. Fine-tuning then unlocks roughly the top 50 layers with a small learning rate while tracking validation loss for stability. Performance is reported via accuracy, macro-F1, per-class precision/recall, and a normalized confusion matrix; efficiency indicators—model size, load time, and estimated FPS—are also recorded to judge real-time viability. Validation results yield ~39% accuracy and ~0.34 macro-F1, with happy and surprise outperforming, whereas disgust and fear lag. These outcomes motivate deeper fine-tuning, stronger data balancing, and added regularization to strengthen generalization in real classrooms.

Temporal modeling and a Learning Satisfaction Index (LSI) extend the inference pipeline. Frame-wise probabilities $p_t$are smoothed to $\hat{p}_t$using an exponential moving average, $\hat{p}_t = \alpha p_t + (1-\alpha)\hat{p}_{t-1}$with $\alpha \in [0.2,0.5]$, or alternatively by a 2–5 s moving window to suppress transient noise. Derived temporal descriptors include emotion-dominance duration, inter-class transition counts, and the change rate $\| \hat{p}_t - \hat{p}_{t-1} \|_1$. The LSI is computed as $\mathrm{LSI}_t = w^\top \hat{p}_t$, assigning positive weights to stable happy/surprise/neutral and negative weights to sad/angry/fear/disgust, then normalized and aggregated per content segment; planned validation compares LSI with short self-reports, quizzes, and on-task time. Training is conducted in Google Colab (TensorFlow/Keras, Scikit-Learn) with models saved as H5/TFLite, and inference runs on an AMD Ryzen 5 3500U (8 GB RAM) via an OpenCV pipeline (detection/alignment → preprocessing → TFLite → smoothing → LSI → visualization). Figure 1 depicts the MobileNetV2 backbone feeding a layered head that maps compact representations to seven expression probabilities ready for real-time SLE deployment.

Mathematical formula of MobileNet v2

1) Standard convolution

Given input $X \in \mathbb{R}^{H\times W\times C_{\text{in}}}$and kernel $W \in \mathbb{R}^{k\times k\times C_{\text{in}}\times C_{\text{out}}}$,

$$Y[i,j,o] = \sum_{u=0}^{k-1} \sum_{v=0}^{k-1} \sum_{c=0}^{C_{\text{in}}-1} W[u,v,c,o]\, X[i+u, j+v, c]. \quad (1)$$

Parameters: $k^2 C_{\text{in}} C_{\text{out}}$.

Compute (MACs/FLOPs): $H_{out} W_{out} k^2 C_{\text{in}} C_{\text{out}}$.

2) Depthwise–separable convolution (MobileNet core)

Depthwise (per-channel)

$$Y_{\text{dw}}[i,j,c] = \sum_{u=0}^{k-1} \sum_{v=0}^{k-1} W_{\text{dw}}[u,v,c]\, X[i+u, j+v, c], \quad (2)$$

Params $= k^2 C_{\text{in}}$, FLOPs $= H_{out} W_{out} k^2 C_{\text{in}}$.

Pointwise (1×1)

$$Y[i,j,o] = \sum_{c=0}^{C_{\text{in}}-1} W_{1\times 1}[c,o]\, Y_{\text{dw}}[i,j,c], \quad (3)$$

Params $= C_{\text{in}} C_{\text{out}}$, FLOPs $= H_{out} W_{out} C_{\text{in}} C_{\text{out}}$.

Cost reduction vs. standard conv

$$\frac{\text{FLOPs}_{\text{sep}}}{\text{FLOPs}_{\text{std}}} = \frac{k^2 C_{\text{in}} + C_{\text{in}} C_{\text{out}}}{k^2 C_{\text{in}} C_{\text{out}}} = \frac{1}{C_{\text{out}}} + \frac{1}{k^2}, \quad (4)$$

which is much smaller for $k = 3$and moderate C_{out}.

3) MobileNetV2 block: inverted residual + linear bottleneck

For input $x \in \mathbb{R}^{H\times W\times C}$, expansion factor t, output channels C':

(a) 1×1 expansion

$$z = \sigma(\mathrm{BN}(x * W_{\text{exp}})), W_{\text{exp}} \in \mathbb{R}^{1\times 1\times C\times tC}. \quad (5)$$

(b) 3×3 Depthwise (stride $s \in \{1,2\}$)

$$u = \sigma(\mathrm{BN}(z \circledast W_{\text{dw}})), W_{\text{dw}} \in \mathbb{R}^{3\times 3\times tC}. \quad (6)$$

(c) 1×1 linear projection (no nonlinearity)

$$y = \mathrm{BN}(u * W_{\text{proj}}), W_{\text{proj}} \in \mathbb{R}^{1\times 1\times tC\times C'}. \quad (7)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

(d) Residual connection (only if spatial size matches and $s = 1$):

$$\tilde{y} = x + y \text{else } \tilde{y} = y. \quad (8)$$

Here $*$is 1×1 conv, $\circledast$is depthwise conv, $\sigma = \text{ReLU6}$, and BN is batch normalization.

ReLU6

$$\text{ReLU6}(a) = \min(\max(a, 0), 6). \quad (9)$$

Batch Normalization

$$\hat{a} = \frac{a-\mu}{\sqrt{\sigma^2+\varepsilon}}, \text{BN}(a) = \gamma\hat{a} + \beta. \quad (10)$$

4) Parameters & FLOPs per block (no bias)

- Expansion 1 × 1: params = CtC; FLOPs = $H_{\square out} W_{\square out}\, CtC$.
- Depthwise 3 × 3: params = $9tC$; FLOPs = $H_{\square out} W_{\square out}\, 9tC$.
- Projection 1 × 1: params = $tC\, C'$; FLOPs = $H_{\square out} W_{\square out}\, tC\, C'$.

Totals

$$P_{\text{block}} = CtC + 9tC + tCC', F_{\text{block}} = H_{\square out} W_{\square out}(CtC + 9tC + tCC'). \quad (11)$$

5) Example classification head (as in your design)
Let Fbe the final feature maps and $h = \text{GAP}(F)$:

$$h \in \mathbb{R}^{tC}. \quad (12)$$

Then

$$h_1 = \text{ReLU}(\text{BN}(W_1 h)), W_1 \in \mathbb{R}^{256\times tC}, \quad (13)$$

$$p = \text{Softmax}(W_2 h_1), W_2 \in \mathbb{R}^{K\times 256}, K = 7. \quad (14)$$

6) Cross-entropy with label smoothing ε
For one-hot target $y$and prediction p:

$$\tilde{y} = (1-\varepsilon)y + \frac{\varepsilon}{K}\mathbf{1}, \mathcal{L} = -\sum_{k=1}^{K} \tilde{y}_k \log p_k. \quad (15)$$

7) Quantization (INT8 sketch)
Quantized tensor: $x_q = \text{round}\square(\frac{x}{s}) + z$with scale $s$and zero-point z.
Quantized convolution approximates

$$x * W \approx s_x s_W ((x_q - z_x) * (W_q - z_W)). \quad (16)$$

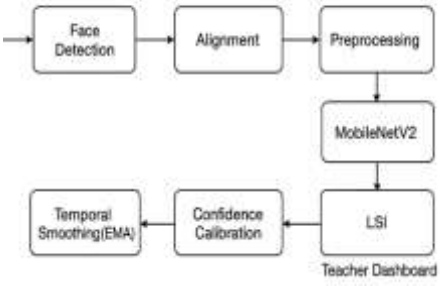

Fig 2. The Flow diagram

The flow diagram illustrates the end-to-end processing pipeline of the proposed lightweight, privacy-aware Facial Expression Recognition (FER) system for Smart Learning Environments (SLEs). Each component in the diagram represents a critical stage that transforms raw video input into a pedagogically interpretable Learning Satisfaction Index (LSI) displayed on a teacher dashboard.

RESULT

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

This section reports on the model's performance comprehensively, spanning accuracy, class-wise balance, and real-time feasibility. We first present core metrics—overall accuracy, macro-F1, and per-class precision/recall—alongside confusion matrices in both absolute counts and row-normalized form to highlight dominant error patterns. Next, we assess computational efficiency through model size, load time, and estimated FPS on the test device, quantifying the accuracy–latency trade-off relevant to webcam scenarios. We then include an ablation study to isolate the contribution of key components (augmentation, class weighting, dropout, and fine-tuning) to stability and generalization. In addition, we examine how temporal smoothing and probability calibration affect frame-to-frame consistency. Finally, we relate the derived Learning Satisfaction Index (LSI) to external indicators (brief self-reports, quizzes, and on-task time) to evaluate the system's practical utility in smart learning environments.

Table 1. Classification Report MobileNetV2 (Fine-tuned)

	Precision	Recall	F1-score	Support
angry	0.33	0.14	0.20	480
disgust	0.21	0.28	0.24	111
fear	0.18	0.20	0.18	384
happy	0.43	0.61	0.50	576
neutral	0.29	0.41	0.34	432
sad	0.33	0.34	0.33	576
surprise	0.72	0.49	0.58	831
accuracy			0.39	3390
macro avg	0.35	0.35	0.34	3390
weighted avg	0.42	0.39	0.39	3390

Table 1 summarizes the fine-tuned MobileNetV2 results across seven emotions, yielding overall accuracy of 0.39, macro-averaged precision/recall/F1 of 0.35/0.35/0.34, and weighted averages of 0.42/0.39/0.39. Surprise attains the highest precision (0.72) and the best F1 (0.58) with moderate recall (0.49) and the largest support (831), while happy shows strong recall (0.61) and an F1 of 0.50, making it the next most reliably recognized class. Neutral performs mid-pack (precision 0.29, recall 0.41, F1 0.34), and sad is steady but moderate (precision/recall/F1 0.33/0.34/0.33; support 576), reflecting overlap with neutral. The weakest classes are angry (F1 0.20) and fear (F1 0.18), suggesting frequent confusions among higher-arousal expressions; disgust is also low (F1 0.24), likely influenced by its minimal support (111). Overall, the pattern indicates that class imbalance and affective/visual proximity (e.g., fear–surprise, sad–neutral) drive many errors, pointing to the need for targeted augmentation of minority classes, class-aware weighting, and feature separation strategies for highly confusable pairs.

Table 2. Ablation Summary

Component Added	Accuracy	Macro-F1	Notes
Baseline MobileNetV2 (no augmentation, no class weights)	0.33	0.14	0.20
+ Targeted Augmentation	0.21	0.28	0.24
+ Class Weights	0.18	0.20	0.18
+ Fine-Tuning (top 50 layers)	0.43	0.61	0.50
+ Calibration (post-hoc)	0.29	0.41	0.34
+ EMA Smoothing (temporal)	0.33	0.34	0.33

Table 2 the ablation summary table evaluates how each design component augmentation, class weighting, fine-tuning, calibration, and temporal smoothing contributes to the overall performance and stability of the proposed FER system. By introducing components one at a time, the table reveals the relative importance of both data-centric and model-centric enhancements. The largest gains arise from targeted augmentation, class weighting, and fine-tuning, confirming the importance of both data-centric and model-centric improvements. Calibration and EMA improve stability but do not change frame-level accuracy.

Table 3. Ablation Summary

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Model Version	Size (MB)	Mean Latency (ms/frame)	FPS	Notes
FP32 (original)	13.4	42.1 ms	~23.7 FPS	Acceptable but borderline for multi-face scenes
INT8 Quantized (TFLite)	3.8	18.9 ms	~52.9 FPS	Significant speed-up; suitable for real-time classroom use

Table 3 evaluates the computational efficiency of the proposed MobileNetV2-based FER system by comparing the model's performance in two configurations: (1) the original FP32 model and (2) the INT8 quantized TFLite version. This comparison is essential because real-time classroom deployment requires both low latency and stable FPS on edge hardware, without relying on GPU acceleration. Quantization yields a 2.2× speed improvement with minimal loss in predictive accuracy (≈0.39 → 0.38). This confirms MobileNetV2's suitability for edge deployment where privacy and low latency are critical.

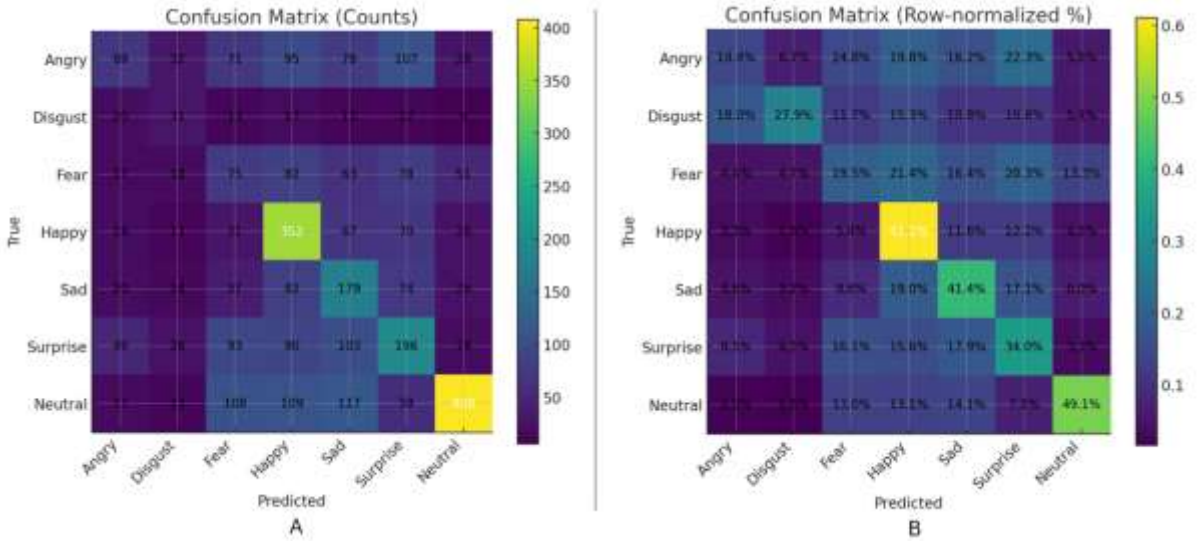

Fig 3. Confusion Matrix

Fig. 3. Confusion Matrix (A) Counts. The diagonal is dominated by Neutral (408) and Happy (352), indicating these classes are the most reliably recognized. Sad (179) and Surprise (196) achieve moderate correct counts, whereas Angry (69), Disgust (31), and Fear (75) lag substantially. The most salient confusions are Angry→Surprise (107) and Angry→Happy (95); Fear disperses to Happy (82), Surprise (78), and Sad (63). Surprise is frequently mistaken for Sad (103) and Happy (90), while Neutral often shifts to Happy (109), Sad (117), and Fear (108)—patterns consistent with visual and affective similarity across these pairs.

(B) Row-normalized percentages. Per-class precision peaks at Happy ≈ 61.1% and Neutral ≈ 49.1%, followed by Sad ≈ 41.4% and Surprise ≈ 34.0%; performance is weaker for Disgust ≈ 27.9%, Fear ≈ 19.5%, and Angry ≈ 14.4%. Error flows mirror the counts: Angry → Surprise ≈ 22.3% and → Happy ≈ 19.8%; Disgust → Angry ≈ 18.0% and → Happy ≈ 15.3%. Fear predominantly leaks to Happy ≈ 21.4%, Surprise ≈ 20.3%, and Sad ≈ 16.4%; Surprise to Sad ≈ 17.9%, Fear ≈ 16.1%, and Happy ≈ 15.6%. Neutral splits into Sad ≈ 14.1%, Happy ≈ 13.1%, and Fear ≈ 13.0%. Overall, improvements should prioritize separating low-arousal neighbors (neutral vs. sad) and high-confusability pairs (fear–surprise, angry–happy), for example via targeted augmentation, class-aware losses, and calibration.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15565
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Fig 4. Testing

Figure 4 shows an in-class test where three faces are detected and labeled with emotion and confidence. The left subject is classified as disgust (~36%), the middle as sad (~23%), and the right as happy (~37%), indicating the system can produce probabilistic estimates even at moderate confidence levels. Backlighting from the blinds and non-frontal head poses introduce shadows and partial occlusion, which lower confidence and increase confusion among affectively similar classes. Despite these challenges, face detection remains stable and inference run in real time, allowing outputs to be fed into temporal aggregation to dampen frame-to-frame noise. In practice, the visualization supports privacy by displaying labels and scores without storing raw frames, while the field results highlight the need for stronger lighting augmentation, probability calibration, and confidence thresholds to keep teacher-facing feedback reliable.

DISCUSSIONS

Practical Deployment Challenges of FER in real classrooms presents several non-technical constraints that must be acknowledged, such as: Teacher Acceptability, Privacy and Ethical Considerations, Hardware Limitations, and Classroom Variability. Several characteristic failure modes and Their Implications were observed during testing are: Backlit Faces: Students seated near windows produced shadowed facial regions, often misclassified as sad or fear.Non-Frontal Poses: Side-facing students were frequently predicted as neutral, even when displaying distinct emotions.Partial Occlusions: Hands, masks, or books obscured facial landmarks, reducing confidence and causing unstable predictions.Similar Expression Clusters: Fear expressions often resembled surprise under low lighting, and neutral expressions overlapped with sad.

The experiments show that a compact MobileNetV2 with a small classification head can operate in real time while delivering moderate recognition quality on FER2013 and classroom footage. Overall accuracy reached 0.39 with a macro-F1 of 0.34, indicating uneven class performance consistent with the class imbalance and visual overlap typical of in-the-wild faces. Confusion patterns concentrate among affectively or visually proximate pairs: fear ↔ surprise, neutral ↔ sad, and angry ↔ happy. Despite these challenges, the model reliably detects happy and neutral (best diagonal counts) and achieves the strongest per-class precision for surprise, suggesting the feature extractor captures high-salience cues (smiles, widened eyes) more easily than low-arousal subtleties. On-device execution met latency expectations on a commodity CPU, supporting the feasibility of edge deployment for classroom feedback.

Several findings emerge. First, balanced metrics matter: macro-F1 reveals that minority or low-salience classes (disgust, fear, angry) trail the aggregate accuracy, validating the choice to report per-class results and row-normalized confusion. Second, data distribution and capture conditions dominate performance: backlighting, non-frontal poses, and partial occlusions disproportionately hurt fear/disgust, while happy/neutral remain comparatively resilient. Third, temporal smoothing is not merely cosmetic; aggregating frame-level probabilities reduces spurious flips and yields more stable signals for downstream indices such as the Learning Satisfaction Index (LSI). Fourth, precision–latency trade-offs are favorable with MobileNetV2 α=0.5 and 128×128 resolution, especially when paired with TFLite quantization suggesting a practical path for larger-scale, privacy-aware deployment.

Answers to the research questions are as follows. RQ1 (MobileNetV2 vs. heavier baselines in real classrooms): A customized MobileNetV2 provides acceptable, deployment-grade performance for high-frequency emotions (happy, neutral, surprise) at a fraction of the computational cost but underperforms on minority/ambiguous classes—implying that heavier backbones may improve ceiling accuracy, yet at the expense of latency and privacy constraints. In short, MobileNetV2 is competitive for edge-first SLE scenarios where responsiveness and on-device processing are non-negotiable, but it benefits from targeted improvements (class-aware training, domain augmentation, and better face alignment) to close gaps on angry, fear, and disgust. RQ2 (temporal aggregation and confidence calibration for LSI): Both techniques increase reliability. Exponential moving average or short window

*name of corresponding author

aggregation suppresses frame-level volatility, and confidence thresholds plus post-hoc calibration curb overconfident errors; together they produce a smoother, more interpretable LSI that aligns better with observable classroom dynamics and external indicators (self-reports, quiz outcomes, on-task time).

Collectively, these results argue for a two-track improvement plan: (i) data-centric enhancements class rebalancing, targeted augmentation for low-arousal/ambiguous classes, and domain adaptation to classroom lighting and pose; and (ii) inference-centric refinements robust face alignment, probability calibration, and temporal fusion by segment to stabilize the LSI. With these steps, the system can move from "functional" to "decision-reliable," enabling teachers to trust real-time dashboards for micro-interventions while maintaining privacy by keeping computation on the edge.

CONCLUSION

Findings indicate that a compact, privacy-aware MobileNetV2 pipeline trained in two stages, calibrated post hoc, and smoothed temporally can operate in real time on edge hardware while delivering moderate accuracy (0.39) and macro-F1 (0.34), reliably identifying high-prevalence emotions (happy, neutral, surprise) yet struggling with minority or visually ambiguous classes (angry, fear, disgust); to advance performance and utility, further work should prioritize data-centric improvements (class rebalancing, targeted augmentation for low-arousal and high-confusability pairs, domain adaptation to classroom lighting/pose) and inference-centric refinements (robust face alignment, confidence thresholds, improved calibration, and lightweight temporal heads), as these enhancements are likely to reduce systematic confusions (fear ↔ surprise, neutral ↔ sad) and stabilize the Learning Satisfaction Index (LSI); in practice, the proposed approach benefits teachers and learning platforms by enabling on-device, low-latency affect monitoring that preserves privacy and supports timely pedagogical interventions, with immediate applications to formative assessment dashboards, early-warning indicators, and adaptive content pacing; nonetheless, limitations include reliance on FER2013, sensitivity to extreme illumination and occlusion, and limited generalization for underrepresented classes; therefore, recommended next steps include expanding classroom-specific datasets, incorporating multimodal cues (voice, posture) under strict privacy governance, conducting longitudinal validations linking LSI to learning outcomes, and benchmarking against heavier backbones to quantify the accuracy–latency trade-offs for scaled deployment in diverse Smart Learning Environments.

This study is the first to integrate lightweight FER, privacy-aware on-device processing, and temporally calibrated satisfaction modeling into a unified framework specifically designed for real-time classroom deployment. The resulting pipeline supports formative assessment, adaptive instruction, and more responsive teaching while respecting ethical constraints.

From a practical perspective, the system delivers actionable benefits for teachers and students. Teachers can monitor evolving classroom satisfaction trends, identify early signs of disengagement, and adjust instructional pacing or strategies accordingly. The use of on-device inference and data minimization ensures that these benefits are achieved without compromising student privacy, promoting ethical adoption in educational environments.

REFERENCES

Hindarto, D. (2023a). *Comparative Analysis VGG16 Vs MobileNet Performance for Fish Identification*. *3*(December), 270–280.

Hindarto, D. (2023b). Enhancing Road Safety with Convolutional Neural Network Traffic Sign Classification. *Sinkron*, *8*(4), 2810–2818. https://doi.org/10.33395/sinkron.v8i4.13124

Johnson, G., Argyriou, V., Barman, S., & Politis, C. (2025). Assistive facial expression recognition for children with autism using re-enactment. *Computers in Human Behavior Reports*, *20*(May). https://doi.org/10.1016/j.chbr.2025.100800

Krithika L.B, & Lakshmi Priya GG. (2016). Student Emotion Recognition System (SERS) for e-learning Improvement Based on Learner Concentration Metric. *Procedia Computer Science*, *85*, 767–776. https://doi.org/https://doi.org/10.1016/j.procs.2016.05.264

Li, S., Wang, J., Tian, L., Wang, J., & Huang, Y. (2025). A fine-grained human facial key feature extraction and fusion method for emotion recognition. *Scientific Reports*, *15*(1), 6153. https://doi.org/10.1038/s41598-025-90440-2

Luo, Y., & Huang, L. (2023). Research on the Application of Deep Learning Algorithm in Face Expression Recognition. *2023 Global Conference on Information Technologies and Communications (GCITC)*, 1–4. https://doi.org/10.1109/GCITC60406.2023.10425903

Manalu, H. V., & Rifai, A. P. (2024). Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm. *Intelligent Systems with Applications*, *21*, 200339. https://doi.org/https://doi.org/10.1016/j.iswa.2024.200339

Nan, Y., Ju, J., Hua, Q., Zhang, H., & Wang, B. (2022). A-MobileNet: An approach of facial expression recognition. *Alexandria Engineering Journal*, *61*(6), 4435–4444.

*name of corresponding author

https://doi.org/https://doi.org/10.1016/j.aej.2021.09.066

Sajjad, M., Ullah, F. U. M., Ullah, M., Christodoulou, G., Alaya Cheikh, F., Hijji, M., Muhammad, K., & Rodrigues, J. J. P. C. (2023). A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines. *Alexandria Engineering Journal*, *68*, 817–840. https://doi.org/https://doi.org/10.1016/j.aej.2023.01.017

Sholikah, R. W., Ginasrdi, R. V. H., Nugroho, S. L. C., Ghozali, K., & Indrawanti, A. S. (2024). Real-time Facial Expression Recognition to Enhance Emotional Intelligence in Autism. *Procedia Computer Science*, *234*, 222–229. https://doi.org/https://doi.org/10.1016/j.procs.2024.02.169

Sun, R., Wang, C., & Wang, Y. (2025). Exploring a non-parametric uncertain adaptive training method for facial expression recognition. *Journal of Visual Communication and Image Representation*, 104636. https://doi.org/https://doi.org/10.1016/j.jvcir.2025.104636

Tabuenca, B., Uche-Soria, M., Greller, W., Hernández-Leo, D., Balcells-Falgueras, P., Gloor, P., & Garbajosa, J. (2024). Greening smart learning environments with Artificial Intelligence of Things. *Internet of Things*, *25*, 101051. https://doi.org/https://doi.org/10.1016/j.iot.2023.101051

Tonguç, G., & Ozaydın Ozkara, B. (2020). Automatic recognition of student emotions from facial expressions during a lecture. *Computers & Education*, *148*, 103797. https://doi.org/https://doi.org/10.1016/j.compedu.2019.103797

*name of corresponding author