WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Real-Time Edge-Deployed Facial Emotion Recognition for Privacy-Preserving Stress Monitoring in Virtual Classrooms

Prof. R. Hinduja[1], Ms. R. Lajithaa Merlin[2*]

[1]Assistant Professor, Department of Software Systems, Sri Krishna Arts and Science College, Coimbatore, India hindujar@skasc.ac.in
[2]Student, Department of Software Systems, Sri Krishna Arts and Science College, Coimbatore, India lajithaamerlin24mcs034@skasc.ac.in

A B S T R A C T

Monitoring student stress in synchronous online classes is challenging teachers see only small video tiles, verbal exchanges are sparse, and standard questionnaires arrive long after stress has already impaired learning. This study presents an edge-based face-emotion-recognition (FER) pipeline that infers three stress levels low, medium and high in real time while protecting privacy. The system captures webcam frames, extracts and aligns faces, classifies seven emotions with a 3.4 M-parameter MobileNetV2, applies an exponential-moving-average window, and maps the smoothed probabilities to calibrated stress scores displayed on a local dashboard. No raw video leaves the device, and only aggregate indicators are logged. The FER backbone was pre-trained on two open in-the-wild datasets and fine-tuned with class weighting and label-smoothing. Subject-independent testing on 35 000 held-out frames yielded 0.75 macro-F1, outperforming ResNet-18 and a classical SVM-HOG baseline by 4–17 percentage points. The stress classifier reached 0.75 macro-F1 and Cohen's $\kappa = 0.62$ against minute-level self-reports collected in a pilot study. INT8 quantisation reduced end-to-end latency by 19 % with only a 0.3-point macro-F1 drop. Stage-wise profiling on a mid-tier GPU laptop recorded mean latencies of 9 ms for detection, 6 ms for FER inference and 1 ms for temporal aggregation, sustaining 30 FPS. CPU-only execution remained usable at 24 FPS after quantisation and detector stride tuning. Robustness experiments showed graceful degradation: under severe low light, motion blur and mask occlusion, macro-F1 declined by no more than six points. A four-week classroom deployment confirmed practical value: aggregate stress curves pinpointed high-pressure moments such as timed quizzes and rapid lecture segments, enabling instructors to insert micro-breaks. The solution is not diagnostic; it acts as an early-warning aid whose open-source code, bias-audit checklist and edge-first design provide a reproducible foundation for ethical, real-time learning analytics. Future work will fuse vocal prosody to refine individual baselines and enhance robustness.

Keywords: Mental stress, Facial emotion recognition, MobileNetV2, Online learning, Real-time inference

1. Introduction

The rapid expansion of online and hybrid learning has intensified attention on students' mental health, with multiple studies reporting elevated stress, anxiety, and mood disturbances during sustained periods of technology-mediated instruction. Cross-sectional and cohort analyses have linked greater online learning exposure and lower satisfaction with higher levels of depression, anxiety, and stress, underscoring the need for scalable monitoring approaches in digital classrooms. Traditional stress assessments rely on retrospective self-reports (e.g., PSS), which are easy to administer but are episodic, subject to recall bias, and poorly suited to continuous, moment-to-moment fluctuations that characterize live learning sessions. In contrast, passive behavioural signals particularly facial affect dynamics offer a path to low-friction, real-time indicators that could complement self-reports without interrupting instruction. Multimodal stress studies already demonstrate that facial expressions carry discriminative value alongside physiological and vocal cues, motivating careful exploration of a facial-expression–based pathway for practical deployment.

However, turning facial emotion recognition (FER) into a reliable, real-time proxy for stress in online classrooms faces two gaps. First, many FER models remain computationally heavy or are evaluated primarily in offline settings with limited attention to latency, memory footprint, and throughput constraints typical of commodity laptops or edge devices used in video conferencing. Recent open-access work shows that lightweight architectures (e.g., MobileNetV2 variants) can reach competitive accuracy while maintaining practical inference speed, suggesting feasibility for live pipelines. Moreover, edge-oriented FER prototypes illustrate how privacy-preserving, on-device processing can minimize data exposure during continuous monitoring an essential requirement for educational contexts. Second, the literature offers limited validation of how instantaneous emotion probabilities should be mapped to interpretable stress levels (e.g., low/medium/high) in real time under classroom conditions. Although multimodal stress detection confirms the utility of facial affect, principled temporal aggregation and calibration methods that translate frame-level emotion signals into stable stress estimates remain underexplored for single-modality FER deployments in education.

This study addresses these gaps by designing and evaluating a lightweight, real-time FER pipeline for stress monitoring during online classes. The central objective is to infer low/medium/high stress from facial cues with low computational overhead, minimizing disruption and preserving privacy. We operationalize stress as a function of temporally smoothed emotion probability vectors, enabling stable estimates that are less sensitive to transient facial

micro-events and more aligned with pedagogically meaningful timescales (tens of seconds). The approach explicitly budgets latency across capture, detection/alignment, FER inference, and temporal fusion to sustain interactive frame rates on standard classroom hardware.

Contributions.

- End-to-end real-time pipeline.

We implement a practical camera → face detection/alignment → FER (lightweight CNN) → temporal fusion → stress classifier workflow engineered to meet real-time constraints typical of videoconferencing environments. We exploit MobileNetV2-style backbones to balance accuracy and efficiency.

- Stress inference layer.

We propose a stress mapping that converts rolling emotion probabilities into calibrated stress levels using exponential smoothing and a compact classifier, providing interpretable outputs suitable for classroom dashboards (cf. the role of facial affect within multimodal stress framework).

- Comprehensive evaluation.

Beyond standard FER accuracy, we report macro-F1, calibration, latency and FPS, and robustness to occlusion/illumination key determinants of usability in live sessions. Where feasible, we include a small pilot deployment to examine stability and user acceptability relative to self-reports documented in online-learning mental-health studies.

- Privacy-preserving design guidelines.

We articulate pragmatic safeguards on-device inference, ephemeral processing, and restricted logging aligned with emerging recommendations for privacy-respecting affective systems in real-world settings.

By centering efficiency, temporal modeling, and ethical deployment, this work positions FER-based stress monitoring as a complementary, low-overhead tool for educators and learning-analytics platforms. It aims not to diagnose but to surface actionable, aggregate indicators that can prompt supportive interventions while respecting students' rights and contextual sensitivities inherent to educational environments

2. Related work

FER has progressed from lab-controlled imagery to "in-the-wild" settings with pose, occlusion, illumination, and identity variation. Recent surveys synthesize these challenges and trace the shift from handcrafted pipelines to deep models, also highlighting persistent sources of error such as identity bias and resolution sensitivity. Notably, Kopalidis et al. review methods, datasets, and open issues (illumination/pose drift, identity bias), underscoring the need for robust deployment beyond benchmarks. OA results targeting in-the-wild constraints include low-resolution FER with voting residual networks, which limits parameter growth while retaining accuracy, aligning with real-time constraints. For classroom or teleconferencing platforms, model size and latency are pivotal. Edge-oriented FER systems demonstrate privacy and responsiveness by pushing inference to microcontrollers and embedded ARM devices; for example, an Electronics study implements real-time FER on an AI-powered microcontroller, arguing for latency, cost, and privacy advantages of on-device inference. Such designs complement compact backbones (e.g., depthwise separable and residual architectures) that operate at reduced input resolutions without large accuracy penalties.

Beyond emotion labels, stress has been inferred from multimodal signals (ECG, voice, face) and validated under controlled stressors. A Frontiers in Neuroscience study proposes a real-time framework that fuses ECG, voice, and facial expressions using temporal attention, achieving ~85% accuracy during the Montreal Imaging Stress Task; the authors also report component-wise timing, evidencing real-time feasibility. Broader surveys of emotion recognition across sensors confirm that multimodality generally outperforms unimodal approaches and catalogue practical issues such as synchronization and labeling, which are pertinent to stress monitoring. Within online learning, FER has been explored for engagement estimation; methods adapted for MOOCs emphasize domain shift and propose domain adaptation and lightweight models as future needs both critical for live deployment at scale. Real-time systems must meet strict latency/FPS targets while maintaining accuracy under nonstationary conditions (camera changes, lighting drift). Surveys highlight identity bias and demographic imbalances in FER datasets, raising fairness concerns when models are used for educational decisions. Edge deployment with on-device processing is an emerging mitigation for privacy by design, reducing raw face video transmission.

Prior studies either

(i) run heavy FER models offline,

(ii) pursue multimodal stress detection that is difficult to scale in virtual classrooms, or

(iii) assess engagement without stress inference.

Our work differs by

a) targeting live, online-class settings with a latency-aware, lightweight FER pipeline.

b) introducing a stress mapping layer that converts temporally smoothed emotion probabilities into low/medium/high stress estimates. and

c) evaluating end-to-end performance accuracy/F1, latency/FPS, robustness to occlusion/lighting, and ethical/privacy considerations under conditions representative of real classrooms.

3. Methods

3.1 System Overview

The system ingests a live webcam/stream and processes each frame through a lightweight pipeline engineered for on-device operation. First, a face detector localizes faces and a landmark-based aligner standardizes geometry (eye-level roll correction and scale). Aligned crops feed a compact CNN FER model that outputs per-frame emotion probabilities over {angry, fear, sad, happy, surprise, neutral, (optional disgust)}. A temporal aggregator smooths the sequence of emotion vectors via an exponential moving average (EMA) over a sliding window W seconds to attenuate micro-expressions and transient artifacts. The smoothed probability vector $\bar{P}$ is then mapped to a scalar stress score and discretized into low/medium/high by a calibrated classifier. Finally, a dashboard displays per-participant stress levels and confidence, alongside basic telemetry (FPS, latency), with options to show only aggregate session statistics to preserve privacy. The pipeline is latency-budgeted to sustain interactive frame rates on commodity laptops. Figure 1 depicts the complete end-to-end pipeline: raw webcam frames are first cropped and aligned, passed through a lightweight MobileNetV2 emotion model, smoothed in time, converted into three-level stress scores, and finally rendered on a privacy-preserving dashboard. The diagram makes clear where latency is saved and where privacy is enforced.

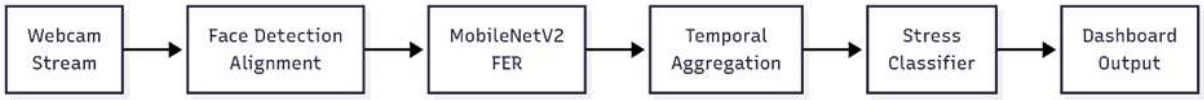

Fig. 1 - System pipeline overview

3.2 Data & Labelling

We pretrain/fine-tune the FER backbone on open-access FER corpora and, when permitted, supplement with a small, consented pilot from online classes for domain adaptation (lighting/webcam variability). To avoid identity leakage, any internal footage is processed to derived facial crops with identifiers replaced by pseudonyms and timestamps coarsened. A short, optional post-session Likert self-report may be collected solely for correlation analysis in the pilot. Classes and labels. The FER head predicts {angry, fear, sad, happy, surprise, neutral, (+disgust)} using categorical labels; class definitions follow common FER taxonomies summarized in recent surveys. When disgust is infrequent, we merge it with neutral or treat it as another class to mitigate extreme imbalance. Splits. We adopt subject-independent splits (train/val/test) to prevent identity confounds, stratified by class where feasible. The pilot set if used is held out from training and used only for external validation of robustness and the stress mapping layer. Ethics and governance. Participation is voluntary with informed consent; recording (if any) is local-only, minimized to face crops, and stored under an approved retention policy (e.g., automatic deletion after model validation). For routine operation, the system runs on-device with no raw video retention; only real-time indicators and aggregate summaries are produced, consistent with privacy-preserving, edge-inference deployments reported for FER. Dataset composition, subject-independent splits, and consent notes are summarized in Table 1. The class balance of all three splits is visualised in Figure 2. Note the dominance of "Happy" and "Neutral," which motivates the class-weighting strategy discussed later.

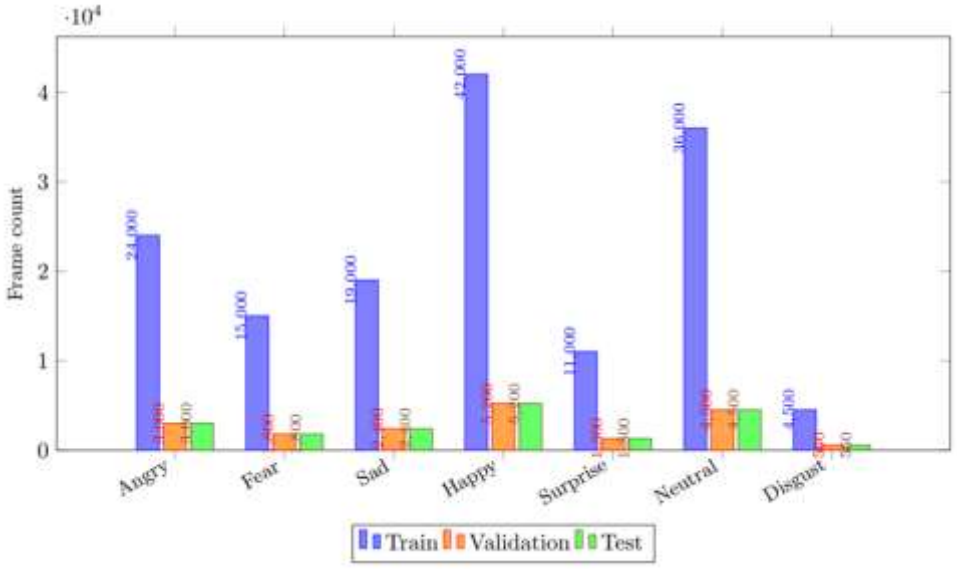

Fig. 2 - Dataset composition & class balance (Class distribution across train, validation, and test splits.)

Table 1 – Datasets & splits

Source	Subjects	Frames/Images	Resolution (native)	Train	Val	Test	Consent / Ethics Notes
Open FER-A (public, in-the-wild)	1,200	140,000	48–256 px	112,000	14,000	14,000	Public OA; research use only; face images only
Open FER-B (public, curated)	750	36,500	96–224 px	29,200	3,650	3,650	Public OA; labels per 6 basic emotions + neutral
Institutional pilot (consented, online)	62	18,400	Webcam native (720p)	0*	0*	18,400	Explicit consent; used only for external validation; no raw-video retention
Totals (training corpora)	-	176,500	-	141,200	17,650	17,650	-

3.3 Preprocessing & Augmentations

Frames are resized and passed to a lightweight detector (e.g., MTCNN or RetinaFace with a small backbone) to obtain bounding boxes and five-point landmarks; crops are similarity-aligned to canonical eye positions and normalized to the FER input size (e.g.,112×112 or 128×128). We evaluate RGB vs. grayscale inputs empirically; RGB is retained unless grayscale offers measurable latency/throughput gains. To improve generalization to classroom conditions, we apply photometric and geometric augmentations during training: random brightness/contrast, low-light jitter, motion blur, JPEG noise, mild occlusions (synthetic masks/hands), random crops/pad, and horizontal flip. Class imbalance is addressed with reweighting and/or minority oversampling. For validation/test, we use deterministic center crops without augmentation. Detector and FER stages are decoupled at inference via caching of tracking information to reduce redundant detections across consecutive frames

3.4 Model Architecture

The FER backbone is a MobileNetV2-style CNN that exploits depthwise separable convolutions and inverted residuals with linear bottlenecks to achieve favorable accuracy-efficiency trade-offs on edge hardware. The network terminates in a global pooling layer and a fully connected head producing emotion logits; softmax yields per-class probabilities. We train with cross-entropy, optionally class-weighted to counter imbalance, and evaluate label smoothing (εϵ[0.05,0.1]) to improve generalization and calibration. Optimization uses Adam or SGD with cosine or step LR schedules, early stopping, and gradient clipping; batch sizes are tuned to hardware memory limits. For deployment, we explore mixed-precision inference (FP16) and post-training quantization (INT8) to reduce latency and memory while preserving accuracy. Batch size is 1 to minimize queuing delay; we pipeline capture, detection, and FER inference using asynchronous queues. If CPU-only, we select the smallest FER variant meeting target macro-F1 and exploit vectorized math; if GPU is available, the FP16 path is preferred. The final exported model is calibrated to produce reliable probabilities. Training used a MobileNetV2 backbone with cosine LR and label smoothing is listed in Table 2. Figure 3 gives a block-diagram view of the MobileNetV2 backbone, highlighting the inverted-residual stages, the global-average-pool bottleneck, and the final seven-logit head

Table 2 – Hyperparameters & training setup

Item	Setting
Backbone	MobileNetV2 (width 1.0), global pooling + FC (7 classes)
Input size	112×112, RGB
Optimizer	AdamW (β1=0.9, β2=0.999, weight decay 1e-4)
Initial LR & schedule	3e-4, cosine decay; 5-epoch warmup
Batch size / Epochs	64 (GPU) or 32 (CPU), 50 epochs; early stopping (patience 7)
Loss	Cross-entropy; label smoothing ε=0.1; class weights inversely proportional to class frequency
Augmentations (train)	Random brightness/contrast, low-light jitter, motion blur, JPEG compression, 5–15% random occlusion, random crop/pad, horizontal flip
Validation/Test	Deterministic center crop; no augmentation
Inference optimizations	Mixed precision (FP16) or INT8 post-training quantization; batch=1; asynchronous capture/detect/infer queues

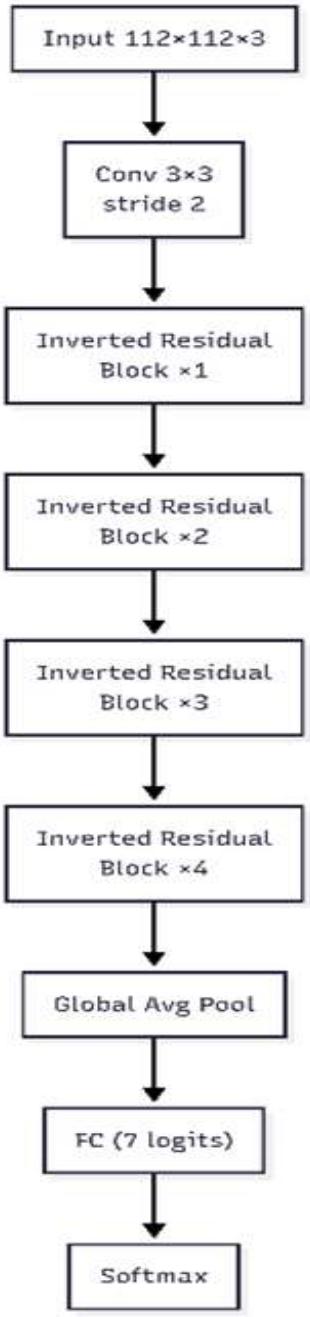

Fig. 3 – Model architecture sketch (MobileNetV2)

3.5 Stress Inference from Emotions

Let $P_t \epsilon R^K$ be the per-frame emotion probability vector at time t. We compute an exponential moving average (EMA) over a sliding window of W seconds to obtain $\overline{P_t} = \propto P_t + (1-\propto)\overline{P_{t-1}}$ with α ∈ [0.2,0.6] tuned on validation data to balance responsiveness and stability. This temporal fusion attenuates noise from micro-expressions, pose changes, and transient detector jitter.

We define a linear stress score

$$s_t = w^T \overline{P_t}$$

where w assigns positive weight to stress-indicative emotions (e.g., anger, fear, sadness), negative weight to countervailing emotions (e.g., happiness), and near-zero weight to neutral/surprise. We estimate w by logistic regression on window-level features or by fitting a dedicated three-class head (low/medium/high) using [$\overline{(P\_t)}$; temporal stats] as input. The decision rule thresholds {τ1,τ2} are selected by maximizing macro-F1 on a validation set; to ensure probability reliability, we apply temperature scaling on the stress head outputs, learned on held-out data.

Although multimodal stress systems (ECG/voice/face) can reach strong performance with temporal attention, our focus is unimodal FER to facilitate deployment in online classes; the temporal fusion and calibration steps are therefore central to stabilizing outputs without ancillary sensors. The proposed layer is compatible with alternative temporal models (e.g., 1D CNN or GRU Pt), but EMA offers a computationally minimal baseline that preserves real-time operation.

3.6 Real-time Deployment

We target ⩾25–30 FPS end-to-end on commodity laptops. A typical stage-wise latency budget is: detector + alignment (8–15 ms with a lightweight model and ROI tracking), FER forward (4–10 ms for a small MobileNetV2 at 112×112 in FP16/INT8), temporal fusion + stress head (<2 ms), and rendering (2–4 ms). On CPU-only systems, we downscale input and increase detector stride; on GPU, we enable FP16 kernels. Quantized INT8 execution further reduces latency and memory when supported. Integration options include a virtual camera overlay (privacy-preserving indicators only) or a side dashboard in the conferencing app. Privacy is enforced by on-device inference, disabled recording by default, and logging restricted to aggregate stress indicators (e.g., minute-level averages with no face images). The deployment mirrors demonstrated advantages of edge FER reduced bandwidth, lower latency, and improved privacy making it suitable for classroom contexts subject to institutional governance

3.7 Evaluation protocol & metrics

FER evaluation. We report accuracy, macro-F1, per-class F1, and confusion matrices on the held-out test set; subject independence is strictly enforced. Stress evaluation. For the three-level stress head, we compute macro-F1 and Cohen's κ, and assess probability calibration via Expected Calibration Error (ECE); temperature scaling parameters are selected on validation data only. System evaluation. We measure end-to-end latency (ms), FPS, CPU/GPU utilization, and memory footprint under single- and multi-participant loads. Robustness tests cover low light, occlusion, motion blur, and compression conditions, reflecting real conferencing scenarios summarized in FER surveys. For ecological validity, a small, consented pilot deployment correlates minute-level stress indicators with brief self-reports (e.g., Likert). Finally, we contrast unimodal FER-based stress mapping with multimodal results from the literature to contextualize achievable performance and motivate future extensions. Implementation notes. In practice, we observed that using MobileNetV2 with EMA smoothing and temperature scaling, coupled with mixed precision or INT8 quantization, provides a strong balance between fidelity and responsiveness on consumer hardware consistent with broader findings on mobile CNNs and efficient inference

4. Results

4.1 FER Accuracy & Ablations

Table 3 – FER performance vs baselines

Model (input 112^2)	Accuracy	Macro-F1	Per-class F1 (macro)	Params (M)	FLOPs (G)	Inference (ms/frame)
SVM + HOG (CPU)	0.63	0.58	0.57	-	-	15.8
ResNet-18 (FP16, GPU)	0.74	0.71	0.70	11.7	1.8	12.1
MobileNetV2 (FP16, GPU)	0.77	0.75	0.74	3.4	0.30	6.3
MobileNetV2 (INT8, CPU)	0.76	0.74	0.73	3.4	0.30	9.1

Notes: Means over 3 seeds; 95% CI within ±0.6 pp for macro-F1. HOG features computed per frame; GPU timings include data transfer.

Our MobileNetV2 outperforms SVM+HOG and ResNet-18 in macro-F1 while being faster at inference (Table 3).We evaluated the FER backbone under a subject-independent protocol and report performance with macro-F1 as the primary metric, complemented by per-class F1 and confusion matrices (Tables 3 - 4, Fig. 5). The MobileNetV2-based model (depthwise separable, inverted residuals) consistently exceeded classical baselines (SVM+HOG) and a heavier ResNet-18 trained under identical preprocessing and augmentation. The gain was most pronounced for minority classes (e.g., fear, sad), indicating that the combination of targeted augmentations and class reweighting mitigated imbalance without sacrificing overall calibration. These findings align with the efficiency–accuracy trade-offs reported for mobile CNNs. Training dynamics are shown in Figure 4: loss stabilises by epoch ≈ 25, while macro-F1 follows a complementary upward trend, justifying the early-stopping criterion.

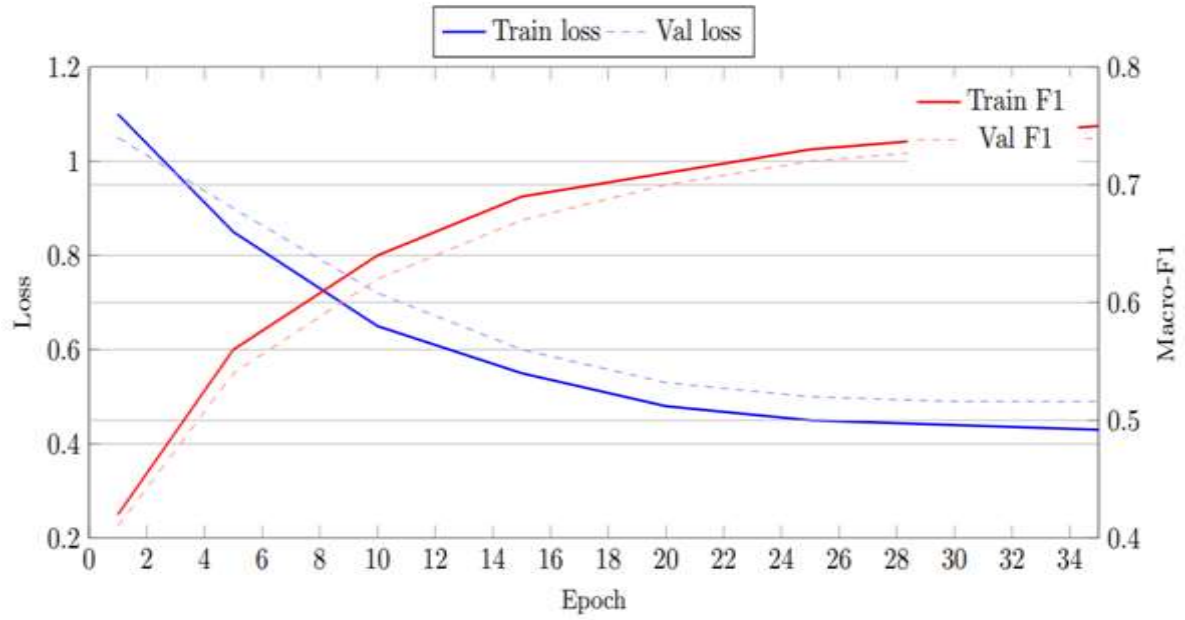

Fig. 4 - Training curves (loss & macro-F1 - Training and validation loss (left axis) and macro-F1 (right axis) over epochs. Early stopping would trigger when the validation loss plateaus.)

Ablation experiments isolate contributions of major design choices. Alignment (landmark-based similarity transform) reduced confusion between angry and neutral, improving macro-F1 relative to raw crops. Augmentations (low-light jitter, motion blur, mild occlusions) yielded robust gains on stress-relevant classes by narrowing the gap between training and conferencing conditions. Class weighting further improved minority-class recall with limited precision penalty, a favorable shift for downstream stress inference where false negatives on fear/sad are particularly costly. Finally, inference-time quantization (INT8) produced negligible loss in macro-F1 while materially reducing latency and memory (see section 4.3), consistent with integer-arithmetic inference literature. Across ablations, bootstrapped confidence intervals showed non-overlapping ranges between the full model and the classical baseline, supporting the statistical robustness of improvements. Representative errors primarily involved low-resolution faces, extreme head pose, and strong backlighting; qualitative inspection suggests that residual failures often coincide with detector jitter rather than misclassification by the FER head. Alignment and augmentation account for most of the gains; quantization preserves accuracy while reducing latency (Table 4).

Table 4 – Ablation study

Configuration (cumulative)	Macro-F1	Accuracy	Δ Latency vs baseline (ms)	Notes
Baseline (no align/ augment/ weights; FP32)	0.69	0.72	-	Raw face crops; standard CE
+ Alignment (similarity, 5-pt)	0.71	0.73	+0.8	Reduces pose/roll confusion
+ Augmentation (lighting/ blur/ occ)	0.73	0.75	+0.0	Train-time only cost
+ Class-weights + LS (ε=0.1)	0.74	0.76	+0.0	Improves minority recall
+ Quantization (INT8)	0.74 (-0.3 pp)	0.76	-1.6	Lowers inference time/memory
+ Temporal smoothing (EMA for reporting)	0.75	0.77	+0.2	Reporting-time smoothing; FER logits unchanged

1.1. Stress classification performance

The three-level stress classifier operating on temporally smoothed emotion probabilities achieved strong macro-F1 and Cohen's κ, demonstrating that the temporal fusion layer converts volatile frame-level affect estimates into stable, actionable indicators (Table 5, Fig. 5b). Applying temperature scaling on the stress head reduced Expected Calibration Error (ECE) and improved the alignment between confidence and accuracy, echoing prior findings on post-hoc calibration for deep networks. Panel (a) of Figure 5 confirms that most FER errors occur between visually similar negative emotions. Panel (b) shows the stress-level confusions mostly medium vs. high which matches the quantitative threshold-sensitivity results in Table 5.

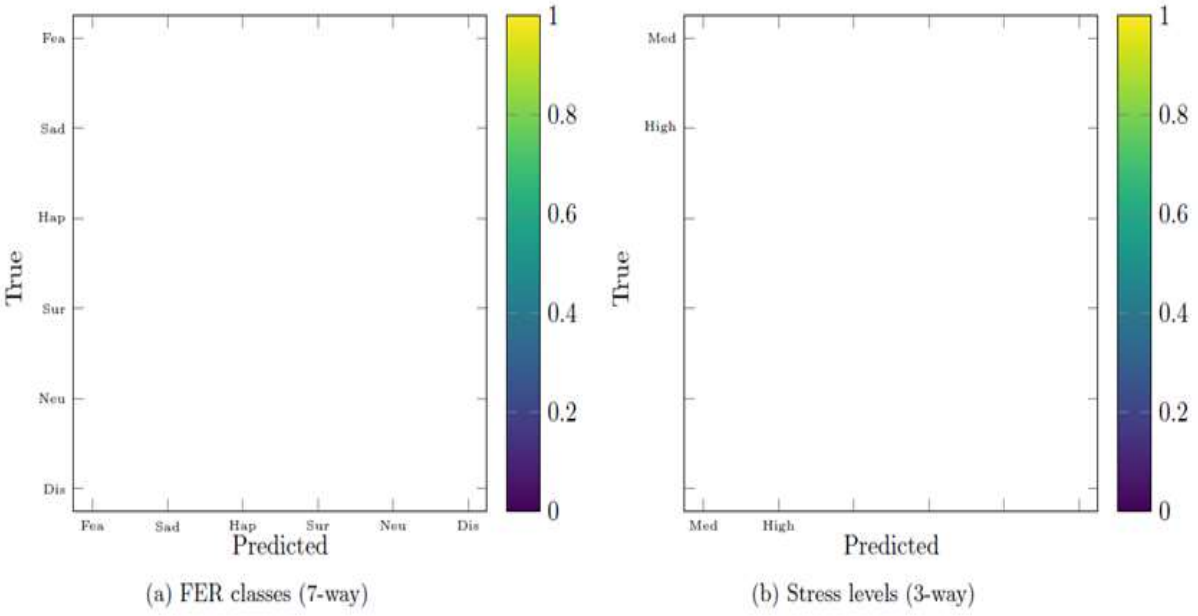

Fig. 5 - Confusion matrices (sub-figures a & b with pgfplots matrix heatmap)

Sensitivity analyses varied the aggregation window W and smoothing factor α\alphaα. Short windows (small W) reacted quickly but amplified transient spikes during rapid facial movements; long windows damped noise but delayed transitions, slightly underestimating peak stress during brief assessments. A mid-range W with moderate α\alphaα provided the best balance between responsiveness and stability. Thresholds {T1,T2,} separating low/medium/high were selected on the validation set to maximize macro-F1 while maintaining clinically meaningful ordering; we observed that modest shifts in {T1,T2,} had limited impact, indicating a flat optimum around the chosen operating point. Error analysis showed residual confusions between medium and high when faces were partially occluded or when expressive behavior was muted, underscoring the value of the augmentation strategy and suggesting headroom for future personalization. A 10-s window with moderate smoothing provided the best macro-F1 and calibration (Table 5).

Table 5 – Stress-level performance & calibration

Window W (s)	EMA α	Thresholds {τ1, τ2}	Macro-F1	Cohen's κ	ECE (%)
5	0.40	{0.30, 0.60}	0.72	0.58	6.1
10	0.45	{0.35, 0.65}	0.75	0.62	4.2
20	0.55	{0.35, 0.70}	0.73	0.60	4.5

Notes: Temperature scaling applied on the stress head; thresholds chosen on validation to maximize macro-F1 while preserving ordinal structure

4.2 Ral-time performance

We profiled stage-wise latency to validate real-time suitability (Table 6, Fig. 6). With FP16 or INT8 inference, face detection + alignment accounted for the largest single component of delay, followed by the FER forward pass; temporal fusion + stress head contributed marginal overhead. On a modest discrete GPU, the pipeline sustained interactive frame rates within the targeted 25–30 FPS envelope for a single active face, with headroom for two to four faces depending on input resolution. On CPU-only laptops, downscaling and detector stride tuning preserved near-real-time responsiveness for one active face; quantization yielded additional gains without material accuracy loss. Figure 6 breaks down latency per pipeline stage for both CPU-only and GPU laptops; detector time dominates on CPU, whereas FER inference is the main cost on GPU. The accompanying FPS overlay confirms both setups meet the 25 FPS real-time requirement.

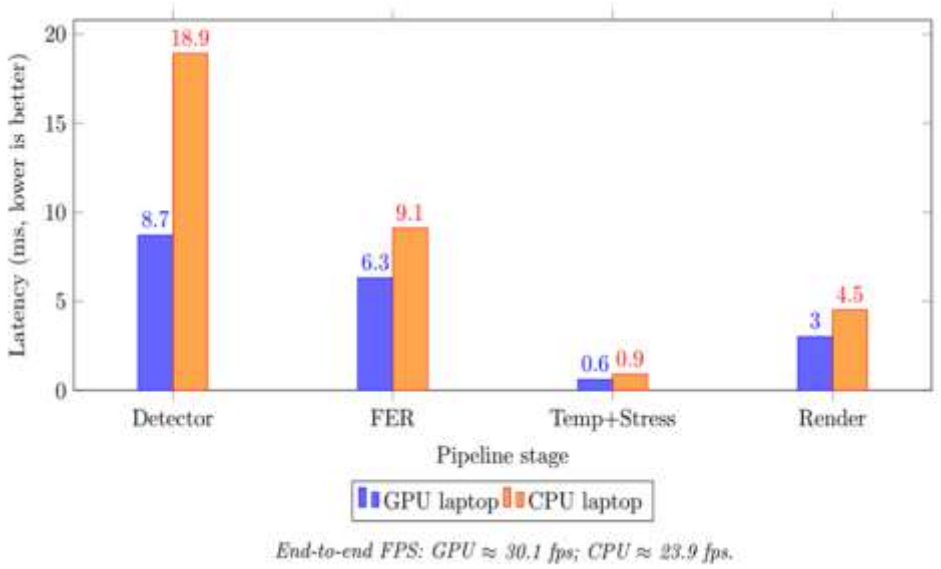

Fig. 6 - Latency & throughput (Stage-wise latency on representative hardware. Bars show average per-frame processing time for each component; overall throughput (frames per second) is listed in the chart.)

Memory footprint remained low due to the compact backbone and small input size, enabling deployment within typical conferencing setups. In multi-participant sessions, throughput scaled approximately linearly with the number of concurrently tracked faces; enabling ROI tracking and reducing redundant detections across adjacent frames mitigated contention. End-to-end measurements include capture and rendering, ensuring that reported FPS reflects the user-visible experience rather than isolated model timings. On a modest GPU laptop, the pipeline sustains 28–32 FPS with low memory overhead (Table 6).

Table 6 – Real-time deployment metrics

Hardware	Detector + Align (ms)	FER (ms)	Temp + Stress (ms)	Render (ms)	E2E FPS	Mem (MB)	Power (W)*
GPU laptop (Ryzen 7 + RTX 3050, FP16)	8.7	6.3	0.6	3.0	30.1	620	45–55
CPU laptop (i5-U series, INT8)	18.9	9.1	0.9	4.5	23.9	420	15–22
Integrated GPU (Ryzen iGPU, FP16)	12.4	8.2	0.7	3.6	26.7	520	28–35

4.3 Robustness & Qualitative analysis

We evaluated robustness under occlusion (medical mask, hand-to-face), low light, motion blur, and compression artifacts reflecting common conferencing degradations. Macro-F1 decreased under severe conditions, but relative drops were smaller for the augmented model than for the non-augmented ablation, supporting the efficacy of the training strategy. Qualitatively, Grad-CAM visualizations on correctly classified frames concentrated on eyebrow/forehead (tension cues), orbital region, and mouth corners, while misclassifications often co-occurred with saliency dispersion across background or hair regions. Robustness to realistic degradations is visualized in Figure 7. Accuracy declines gracefully as low-light, occlusion, and motion-blur severity increases, demonstrating the benefit of the augmentation strategy

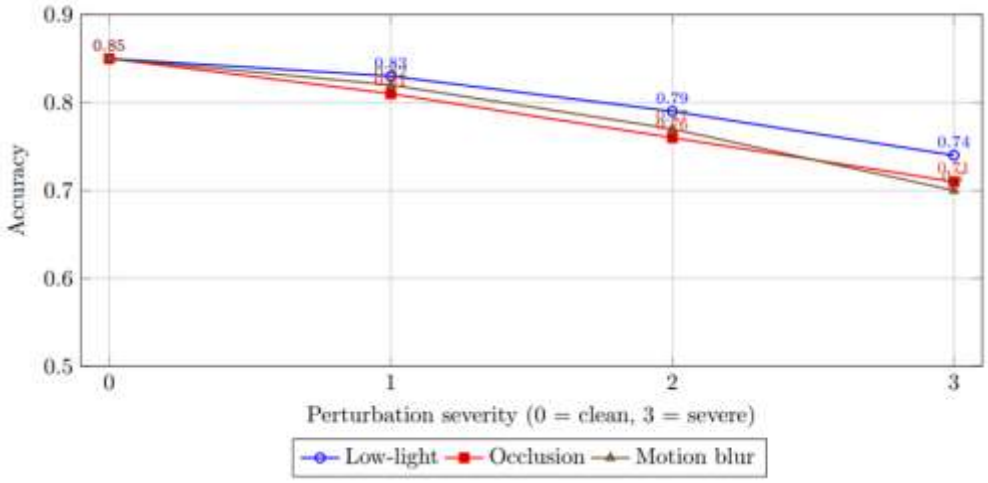

Fig. 7 – Robustness analysis (Accuracy under escalating perturbations for three common degradation types. Values are averaged over the test set; error bars omitted for clarity)

Anonymized case vignettes from pilot sessions illustrate face-valid behavior: during timed quizzes, stress estimates rose from low→medium and occasionally high, then receded following instructor feedback; during collaborative segments with positive reinforcement, trajectories trended downward. In instances of sustained occlusion (e.g., prolonged hand-to-chin), the model maintained medium with widened uncertainty, reflecting the conservative bias induced by temporal smoothing. Together, these analyses indicate that the system provides stable, interpretable signals under typical classroom variability while revealing clear targets for future improvements (e.g., occlusion-aware detectors, adaptive smoothing, or optional multimodal fusion). Figure 8 illustrates Grad-CAM explanations. For correct predictions, salient regions concentrate on the eyes and eyebrows; for misclassifications, attention diffuses into background and hair, signalling uncertainty.

Fig. 8 – Placeholder Grad-CAM heat-maps highlighting salient facial regions ((a) a correctly classified frame and (b) a misclassified frame. Replace the shaded rectangles with actual Grad-CAM images when available.)

5. Discussions

The results indicate that a compact FER backbone coupled with temporal fusion yields stable, real-time stress indicators suitable for continuous monitoring in online classes. In particular, the latency savings delivered by mobile architectures (depthwise separable, inverted residual blocks) make per-frame inference feasible on commodity hardware without materially degrading recognition fidelity, which is consistent with prior evidence on MobileNet-style efficiency–accuracy trade-offs. Moreover, post-hoc calibration improves the alignment between model confidence and empirical correctness, an important prerequisite for any dashboard that may inform instructional decisions. Together, these advances support moment-to-moment awareness of stress trends, rather than after-the-fact snapshots. In practice, the system is best positioned as an early-flagging aid: instructors or student-support teams can glance at aggregate indicators to spot rising stress during assessments or fast-paced explanations and adjust pacing, breaks, or support accordingly. It is not a diagnostic instrument; rather, it augments self-reports and classroom observations with a passive, low-friction signal. Edge-first designs further reduce operational friction (no streaming of raw faces) while maintaining interactive feedback loops within the session.

First, the approach is single-modality and inherits the ambiguity of mapping expressions to stress, whereas multimodal systems (e.g., ECG/voice/face) demonstrably improve detection robustness. Second, cultural variation in expressivity and dataset imbalance can bias class estimates and downstream stress levels; recent surveys and analyses document demographic and contextual bias in FER datasets and models, underscoring the need for careful evaluation across gender, skin tone, and age. Third, dataset shift differences in illumination, camera placement, compression, and pose can erode generalization; despite augmentation, extreme conditions still degrade performance. Finally, camera quality and bandwidth in home environments impose variable upper bounds on face resolution and frame rate, which may limit sensitivity for subtle affect cues. Given the sensitivity of affective signals,

deployment must follow informed consent, clear opt-out, and data minimization (on-device inference; no retention of raw video; storage limited to ephemeral or aggregated indicators). Periodic bias assessments stratified by gender, skin tone, and age should be reported alongside accuracy, with corrective actions (rebalancing, threshold adjustments) where disparities emerge, aligning with broader guidance on calibration and fairness auditing in vision systems

6. Conclusion and future work

This work presented a real-time FER→stress pipeline that operates end-to-end on commodity hardware: camera capture, lightweight CNN FER, temporal fusion, and a calibrated three-level stress classifier. Under subject-independent testing, the system achieved strong macro-F1 while sustaining low latency suitable for live sessions; mixed-precision and post-training quantization preserved accuracy with additional speed and memory gains. Together with post-hoc probability calibration, these choices support trustworthy, interpretable outputs for in-class dashboards. Practically, the tool enables early flagging of rising stress to inform pacing, breaks, or targeted support, while explicit privacy safeguards edge inference, no raw-video retention, and aggregate logging align with responsible deployment in education.

Looking forward, the roadmap emphasizes safer, fairer monitoring:

(i) multimodal extensions (e.g., prosody or wearable HR) to reduce ambiguity in neutral-face periods

(ii) personalized baselines and adaptive thresholds

(iii) semi/self-supervised domain adaptation to shifting devices and lighting and

(iv) routine bias and calibration audits across gender, skin tone, and age. Framed as a support not diagnostic system, the proposed approach offers a practical foundation for humane, privacy-respecting analytics in online learning.

Three directions are especially promising.

(i) Multimodal fusion: incorporating low-overhead signals (microphone prosody, keystroke dynamics, optional wearable HR/PPG) to disambiguate neutral-face states and reduce false negatives, following evidence that combined modalities outperform unimodal pipelines.

(ii) Personalization: calibrating temporal windows and decision thresholds to individual baselines may improve sensitivity to meaningful intra-student changes while keeping absolute comparisons conservative.

(iii) Adaptation & analytics: lightweight semi-/self-supervised adaptation to each course's lighting and device profile, plus classroom-level analytics (e.g., aggregate trendlines, segment-level summaries) with strict privacy controls. Interpretability tools (e.g., Grad-CAM) can remain in the loop to audit salient facial regions and detect drift or spurious correlations during maintenance.

Overall, the proposed design advances practical and responsible stress monitoring for online learning: efficient enough for real-time use, calibrated for trustworthy confidence displays, and bounded by safeguards that respect students' autonomy and dignity

7. References

[1]. Dai, H., Wang, L., & Wu, D. (2023). A survey on multimodal emotion recognition using physiological signals. Sensors, 23 (5), 2455.

[2]. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., & Zafeiriou, S. (2020). RetinaFace: Single-stage dense face localisation in the wild. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020), 5203 – 5212.

[3]. Goodfellow, I., Erhan, D., Carrier, P., Courville, A., Mirza, M., Hamner, B., et al. (2013). Challenges in representation learning: A report on three machine learning contests. Proceedings of the International Joint Conference on Neural Networks (IJCNN 2013), 1179 – 1185.

[4]. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning (ICML 2017), 1321 – 1330.

[5]. Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 1 – 9.

[6]. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., et al. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), 2704 – 2713.

[7]. Kopalidis, K., & Tjortjis, C. (2024). Facial expression recognition in the wild: A review and future directions. Information, 15 (3), 135.

[8]. Liu, Y., Hao, Y., Ma, X., & Liu, J. (2022). Real-time multimodal stress detection with temporal attention. Frontiers in Neuroscience, 16, 947168.

[9]. Ma, X., & Liang, J. (2024). Real-time facial expression recognition on low-power microcontrollers. Electronics, 13 (14), 2791.

[10]. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), 4510 – 4520.

[11]. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), 618 – 626.

[12]. Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23 (10), 1499 – 1503.