

Recognizing Student Facial Expressions with Compact Convolutional and Transformer Networks

Tuba Elmas Alkhan¹, Erdal Alimovski¹ ✉

¹Department of Computer Engineering, Istanbul Sabahattin Zaim University, Istanbul, Turkey

masa.karabala@izu.edu.tr; erdal.alimovski@izu.edu.tr

Corresponding Author: *erdal.alimovski@izu.edu.tr*

Abstract

Facial expression recognition has been proposed as a way to give teachers feedback on how students respond to online and blended lessons, but many educational studies evaluate their models with a single run, no held-out test set, and results that are difficult to reproduce. This study re-evaluates a compact eight-layer convolutional neural network (CNN) for recognizing students' facial expressions and compares it with a ResNet-18 and a Vision Transformer, all trained from scratch under the same protocol on the official training, validation, and test partitions of the FER-2013 dataset. The main task distinguishes happy, sad, and neutral faces (17,010 training and 2,099 test images); the standard seven-class task serves as a benchmark. Each three-class model was trained with three random seeds, and results are reported as accuracy, macro-averaged F1-score, per-class metrics, verified confusion matrices, Grad-CAM attention maps, and CPU inference time. On the held-out test set, the proposed CNN reached 79.4% accuracy and 77.7% macro-F1, compared with 77.8% and 76.2% for ResNet-18 and 58.2% and 54.0% for the Vision Transformer, whose performance was limited by the absence of pretraining. On seven classes, the CNN and ResNet-18 both reached about 64% accuracy. Happy faces were recognized with an F1-score above 90%, while sad and neutral faces were frequently confused with each other. Adam outperformed stochastic gradient descent and Adadelta, and all models classified a face in under 6 ms on a two-core CPU. A compact CNN is therefore a strong and efficient baseline when pretraining is unavailable, although its uneven accuracy across expressions, the limited validity of inferring emotions from faces, and recent legal restrictions argue for using such systems only for aggregate, consent-based feedback.

Article History

Manuscript Received
March 13, 2025

First Revision
May 8, 2025

Final Acceptance
June 2, 2025

Keywords: facial expression recognition; convolutional neural network; vision transformer; FER-2013; e-learning

1 Introduction

Online and blended teaching expanded quickly during the COVID-19 pandemic and has remained part of higher education since. It solved problems of access and scheduling, but it also removed a source of information that teachers use constantly in a classroom: the faces of their students. A lecturer who sees puzzled or bored expressions can slow down, repeat an example, or change activity. In a video lecture

with dozens of small thumbnails, or in recorded material watched alone, that feedback largely disappears. Automatic facial expression recognition (FER) has therefore been proposed as a way to give instructors an aggregate signal of how a class is responding [1, 2, 3, 4].

Research on FER has itself changed substantially. Early systems combined a face detector such as the Viola–Jones cascade [5] with hand-crafted features and shallow classifiers. Convolutional neural networks (CNNs) then became the dominant approach, and the FER-2013 benchmark, released for the ICML 2013 representation-learning challenge, became a standard test bed for them [6, 7]. Since 2020 the field has moved again, towards attention mechanisms and Vision Transformers (ViTs) that model relationships between distant facial regions [8, 9, 10, 11], towards larger in-the-wild datasets such as RAF-DB and AffectNet [12, 13], and towards explainability methods that show which parts of the face drive a prediction [14].

Two gaps remain for educational use. First, many studies that apply FER to students report a single accuracy figure from one training run, often without a held-out test set, per-class results, or a check that the reported confusion matrix is consistent with the headline accuracy. Results produced this way are difficult to compare or reproduce. Second, comparisons between compact CNNs and newer architectures are usually made with large pretrained models and GPU budgets that a university deploying a teaching tool may not have. It is not obvious how a small CNN, a residual network, and a Vision Transformer compare when all are trained from scratch on the same data, under the same protocol, and at a size that can run on an ordinary computer.

The question also has an ethical and legal dimension that earlier work in this area rarely addressed. The link between facial movements and inner emotional states is weaker and more context-dependent than basic-emotion labels suggest [15], FER systems can perform unequally across demographic groups [16], and the European Union’s Artificial Intelligence Act now prohibits AI systems that infer the emotions of people in education institutions, except for medical or safety reasons [17]. Any work in this area therefore needs to be explicit about what a model can and cannot support.

This paper revisits an earlier CNN-based system for recognizing students’ facial expressions and re-evaluates it under a rigorous protocol. It addresses the following question: *how accurately, and how consistently, can compact CNN, residual, and transformer models trained from scratch recognize the facial expressions most relevant to student feedback, and what do their errors and attention patterns reveal about their suitability for educational use?* The main contributions are as follows:

- A reproducible re-evaluation of the proposed eight-layer CNN on FER-2013 using the official training, validation, and held-out test partitions, with results averaged over three random seeds and verified confusion matrices.
- A controlled comparison with a ResNet-18 and a compact Vision Transformer trained from scratch under the same data, augmentation, and stopping rules, on both a three-class (happy, sad, neutral) and the standard seven-class task.
- A comparison of the Adam, stochastic gradient descent (SGD), and Adadelta optimizers for the proposed CNN.
- An analysis of per-class errors, Grad-CAM attention maps, and CPU inference latency, which together indicate where the models succeed, where they fail, and what they would cost to run.
- A discussion of the data-quality, validity, fairness, and regulatory constraints that should govern any educational deployment.

The remainder of the paper is organized as follows. Section 2 reviews research on face detection, facial expression recognition, and emotion and engagement recognition in education. Section 3 describes the dataset, the three models, and the training and evaluation protocol. Section 4 presents the quantitative results, the error analysis, and the visual explanations. Section 5 discusses the findings and their limitations, and Section 6 concludes.

2 Literature Review

2.1 Face Detection and Facial Expression Recognition

An FER system first has to find the face. The Viola–Jones detector combined Haar-like features, an integral image, and a cascade of boosted classifiers to achieve real-time frontal face detection on the hardware of the early 2000s [5], and it remains widely used in educational prototypes because it is fast and available in standard libraries [3]. It is, however, sensitive to pose, occlusion, and lighting. Multitask cascaded CNNs (MTCNN) addressed these weaknesses by detecting faces and facial landmarks jointly in three stages [18], and single-shot detectors such as RetinaFace localize faces, landmarks, and dense 3D shape in one pass and perform well on crowded, in-the-wild images [19]. For classroom video, where students sit at different distances and angles, these detectors are a better match than a frontal cascade.

Expression recognition proper has followed the broader evolution of computer vision. Li and Deng’s survey documents the move from hand-crafted appearance and geometric features to deep networks and identifies the main difficulties of the task: small and noisy datasets, strong variation in pose and illumination, and identity cues that are often more salient than expression cues [7]. FER-2013, with 35,887 low-resolution grayscale faces collected by web search and labeled into seven categories, has been the most widely used benchmark for CNNs; its organizers estimated human accuracy at about 65% [6], a reminder that many labels are ambiguous. A carefully tuned VGG-style network reached 73.28% on its test set without extra data [20], and Gabor-filtered convolutional networks have offered a more efficient alternative [21]. Larger in-the-wild datasets changed the picture. RAF-DB used crowdsourced labels with reliability estimation [12], and AffectNet collected about a million web images, a large share of them manually annotated with categories and continuous valence and arousal [13]. Both showed that label noise is intrinsic to the task. Methods such as self-cure networks suppress uncertain samples during training [22], and data-centric work has shown that relabeling FER-2013 and similar datasets can raise accuracy more than changing the architecture [23].

Attention and transformer models are the most recent development. The Vision Transformer splits an image into patches and processes them with self-attention, which lets every region interact with every other, but it needs large-scale pretraining to outperform CNNs [8]. FER-specific designs combine the two. TransFER learns relation-aware representations from local CNN features with a transformer encoder [9], the Distract Your Attention Network uses several attention heads to capture complementary facial regions [10], and POSTER++ fuses landmark and image features in a simplified two-stream architecture that reached leading results on RAF-DB and AffectNet [11]. Lightweight hybrids built on MobileViT have reported about 70% on FER-2013 and 86.5% on RAF-DB with a few megabytes of parameters [24]. Most of these gains, however, depend on ImageNet or face-recognition pretraining and GPU training. Much less is known about how the architectures compare when trained from scratch at a modest scale, which is the situation of many teaching institutions.

2.2 Emotion and Engagement Recognition in Education

Interest in reading students’ faces predates deep learning. Whitehill and colleagues showed that engagement judgments made by human observers from brief video clips could be approximated by machine classifiers applied to facial features, and related the resulting engagement estimates to students’ learning outcomes [25]. The DAiSEE dataset later provided 9,068 video snippets from 112 users labeled for boredom, confusion, engagement, and frustration at four intensity levels [26]. A review of engagement detection in online learning distinguished computer-vision-based, sensor-based, and log-file-based methods and examined the evidence for each [27].

Applied systems for e-learning have mostly combined a face detector with an emotion classifier and a dashboard for the instructor. Krithika and Lakshmi Priya estimated a concentration metric from eye and head movements and facial cues to flag boredom or disinterest [1]. Ayvaz, Gürüler, and Devrim compared several classical classifiers on facial features and fed the recognized emotions back to the

instructor in real time [2]. Lasri, Solh, and El Belkacemi used a Haar cascade and a CNN trained on FER-2013 and reported about 70% accuracy for seven expressions [3], and Begum, Neelima, and Valan proposed a CNN-based emotion recognizer for e-learning environments [28]. Savchenko and colleagues went a step further and used the facial features of a single expression-recognition network to classify both students' emotions and their engagement in online classes [4]. Across this literature, three weaknesses recur. Evaluation protocols are often underspecified, with no held-out test split or repeated runs. The expression categories used are rarely justified by what a teacher would act on. And validity, fairness, and consent receive little attention.

Those questions have become central. Barrett and colleagues reviewed the evidence on facial expressions and concluded that people do not reliably display a unique facial configuration for each emotion, so a smile or frown is a cue whose meaning depends on context and culture [15]. Xu and colleagues found systematic differences in FER performance across gender, age, and race groups and showed that common mitigation strategies only partly remove them [16]. In the European Union, Article 5(1)(f) of the Artificial Intelligence Act prohibits placing on the market or using AI systems that infer the emotions of natural persons in education institutions, except for medical or safety reasons [17]. Work in this area therefore has to distinguish between technical benchmarking, which remains legitimate, and deployment decisions, which now face legal limits in some jurisdictions and ethical scrutiny everywhere.

Taken together, the literature shows rapid architectural progress on FER benchmarks and a steady stream of educational prototypes, but few studies that compare CNNs and transformers under a single, fully specified, from-scratch protocol, report variability across runs, and examine errors and attention patterns alongside accuracy. The present study is designed to fill that gap for the expressions most relevant to classroom feedback.

3 Materials and Methodology

3.1 Dataset and Tasks

All experiments use FER-2013, the dataset of the ICML 2013 facial expression recognition challenge [6]. It contains 35,887 grayscale face images of 48×48 pixels, each labeled with one of seven expressions: angry, disgust, fear, happy, sad, surprise, or neutral. The faces were found by web image search, automatically registered so that each is roughly centered and occupies a similar area, and then labeled. We used the official file and its official partition: 28,709 training images, 3,589 validation images (the challenge's "public test" set), and 3,589 test images (the "private test" set). The validation set was used only for early stopping and learning-rate scheduling; the test set was used once, after training, for every reported result. An audit of the file found 3,369 images that occur more than once and 13 nearly blank images. They were left in place so that the results remain comparable with other work on the official split, and they are discussed as a limitation.

Two tasks were defined. The *three-class task* keeps the happy, sad, and neutral images, which correspond to the states a teacher is most likely to act on in lectures: satisfaction, discouragement, and a neutral baseline. It has 17,010 training, 2,155 validation, and 2,099 test images (Table 1). The *seven-class task* uses all categories and serves as a benchmark against published results. Because FER-2013 faces are already detected and cropped, face detection was not part of the experiments. In a live system, a detector such as MTCNN or RetinaFace would precede the classifier [18, 19] (Figure 1).

Table 1: FER-2013 images per class in the official partition

Expression	Training	Validation	Test	Total
Angry	3,995	467	491	4,953
Disgust	436	56	55	547
Fear	4,097	496	528	5,121
Happy	7,215	895	879	8,989
Sad	4,830	653	594	6,077
Surprise	3,171	415	416	4,002
Neutral	4,965	607	626	6,198
All seven classes	28,709	3,589	3,589	35,887
Three-class task	17,010	2,155	2,099	21,264

Note: Rows in bold form the three-class task. Validation = official “public test” partition; Test = official “private test” partition.

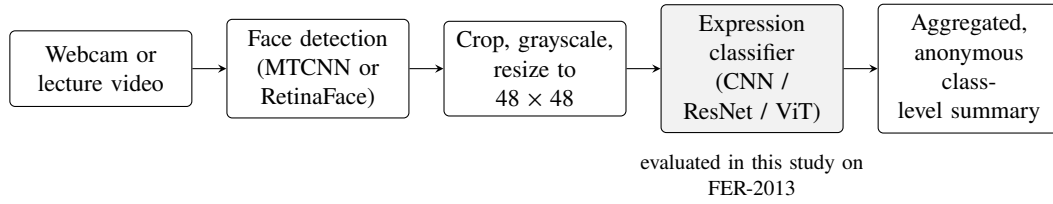


Figure 1: Processing pipeline of a facial expression feedback system. This study evaluates the shaded classification stage; face detection and aggregation are shown for context.

3.2 Models

Three architectures were compared (Table 2). The *proposed CNN* re-implements the network of the original manuscript: eight 3×3 convolutional layers with ReLU activations and He-normal initialization, arranged in three blocks with 32, 64, 128, and 256 filters, batch normalization, and 2×2 max pooling, followed by two fully connected layers of 64 units with batch normalization and dropout of 0.5, and a softmax output layer. The *ResNet-18* follows the residual design of He and colleagues [29], with four stages of two basic blocks each. To suit 48×48 inputs and a CPU budget, it uses a 3×3 stride-2 stem and half the standard channel widths (32, 64, 128, and 256). The *Vision Transformer* follows Dosovitskiy and colleagues [8]: each image is divided into 6×6 patches (64 tokens), which are projected to 128 dimensions, prepended with a class token, combined with learned position embeddings, and processed by six pre-normalized encoder blocks with four attention heads and a 256-unit feed-forward layer. All three were trained from randomly initialized weights. Pretrained ImageNet or face-recognition weights were deliberately not used, so that differences reflect the architectures and not the pretraining data.

Table 2: Architectures and training settings

	Proposed CNN	ResNet-18	Vision Transformer
Building blocks	8 conv. (3×3), BN, max pooling	4 stages $\times$ 2 residual blocks	6 encoder blocks, 4 heads
Feature widths	32–256 filters	32–256 channels	128-d tokens, 256-d MLP
Classifier head	2 dense (64) + dropout 0.5	global average pooling	class token
Parameters	1.33 M	2.80 M	0.81 M
Optimizer (initial LR)	Adam (10^{-3})	Adam (10^{-3})	AdamW (5×10^{-4} , wd 0.05)
CPU latency per image	2.5 ms	3.1 ms	5.4 ms

Note: Parameter counts refer to the three-class models. All models take 48×48 grayscale inputs scaled to $[0, 1]$ and were trained on the same augmented data with a batch size of 64, early stopping on validation loss (patience 6, best weights restored), and halving of the learning rate after 3 epochs without improvement. CPU latency is the median of 200 single-image forward passes.

3.3 Training and Evaluation Protocol

Training images were augmented on the fly with horizontal flips, rotations of up to $\pm 10^\circ$, translations of up to 10% of the image size, and zooms of up to 10%, which corresponds to the transformations used in the original study. The CNN and ResNet-18 were optimized with Adam at an initial learning rate of 10^{-3} [30]. The Vision Transformer used AdamW with a learning rate of 5×10^{-4} and a weight decay of 0.05, the usual choice for transformers. To reproduce the optimizer comparison of the original study, the proposed CNN was also trained with plain SGD and with Adadelta, both at a learning rate of 0.1. Training ran for at most 30 epochs with the stopping and scheduling rules given in Table 2. Because the computing environment could be suspended, the model and optimizer state were checkpointed after every epoch and training resumed from the last checkpoint, with shuffling and augmentation seeded per epoch so that an interrupted run follows the same sequence as an uninterrupted one. Each three-class configuration with Adam or AdamW was trained with three random seeds (0, 1, and 2), which control weight initialization, data shuffling, and augmentation. The optimizer comparison and the seven-class benchmark used seed 0. All experiments ran on a two-core CPU without a GPU, using TensorFlow and Keras.

Performance is reported on the held-out test set as overall accuracy, macro-averaged F1-score (the unweighted mean of per-class F1, which is not dominated by the largest class), per-class precision, recall, and F1, and confusion matrices. For each class c , precision $P_c = TP_c / (TP_c + FP_c)$, recall $R_c = TP_c / (TP_c + FN_c)$, and

$$F1_c = \frac{2P_cR_c}{P_c + R_c}, \quad \text{Macro-F1} = \frac{1}{K} \sum_{c=1}^K F1_c, \quad (1)$$

where K is the number of classes. Test predictions were produced in a fixed order with shuffling disabled, and the evaluation code checks that each confusion matrix sums to the number of test images and that its diagonal equals the number of correct predictions. This check matters because computing a confusion matrix from a shuffled data generator misaligns predictions and labels, a common error in Keras-based studies that produces matrices near chance level even for accurate models.

To see which facial regions drive predictions, Grad-CAM maps [14] were computed for the CNN (last pooling layer) and ResNet-18 (last residual stage). Grad-CAM weights each feature map of a layer by the average gradient of the class score with respect to that map and keeps the positive part of the weighted sum:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right), \quad \alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \quad (2)$$

where A^k is the k -th feature map, y^c the score for class c , and Z the number of spatial positions. Finally, CPU inference latency was measured as the median time of 200 single-image forward passes after warm-up, to indicate whether each model could run on an ordinary computer.

4 Results

4.1 Overall Performance

Table 3 summarizes test performance. On the three-class task, the proposed CNN reached a mean accuracy of 79.4% (± 0.6 across seeds) and a macro-F1 of 77.7%. ResNet-18 followed closely at 77.8% accuracy and 76.2% macro-F1, even though it has twice as many parameters. The Vision Transformer trained from scratch lagged far behind, at 58.2% accuracy and 54.0% macro-F1. Across seeds, test accuracy varied by at most about 2 percentage points for any model (standard deviations of 0.6 to 1.2 points), so the ranking does not depend on a favorable initialization. The CNN also generalized slightly better than its validation scores suggested: its test accuracy exceeded its validation accuracy in every run.

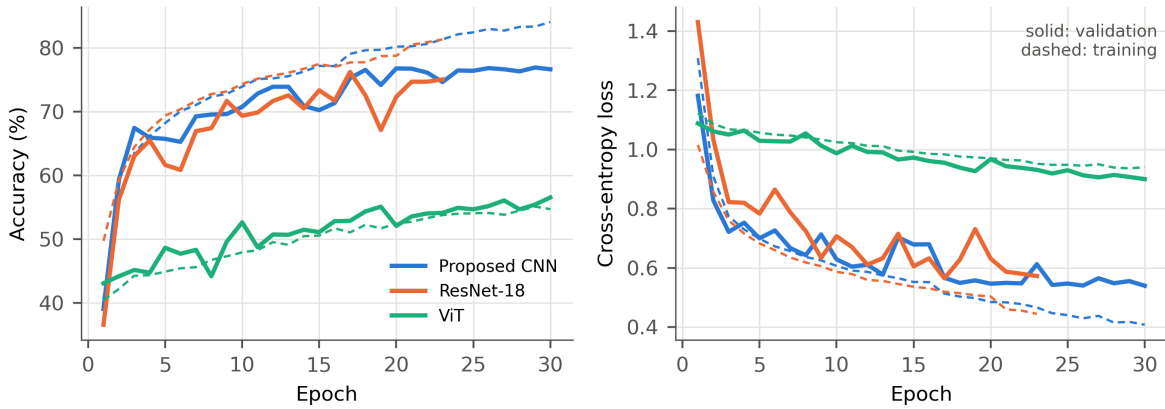
Table 3: Test-set performance on FER-2013

Model	Parameters	Accuracy (%)	Macro-F1 (%)	Best epoch	Training (min)
<i>Three-class task (happy, sad, neutral); mean $\pm$ SD over 3 seeds</i>					
Proposed CNN	1.33 M	79.4 $\pm$ 0.6	77.7 $\pm$ 0.5	27	39
ResNet-18	2.80 M	77.8 $\pm$ 0.9	76.2 $\pm$ 0.8	21	30
Vision Transformer	0.81 M	58.2 $\pm$ 1.2	54.0 $\pm$ 1.5	30	56
<i>Seven-class task; seed 0</i>					
Proposed CNN	1.33 M	64.4	56.0	24	63
ResNet-18	2.80 M	64.3	61.3	24	62
Vision Transformer	0.81 M	40.8	30.6	28	75

Note: Official private-test partition (2,099 images for three classes, 3,589 for seven). Best epoch = epoch with the lowest validation loss (mean over seeds for the three-class task). Training time on a two-core CPU without a GPU. SD = standard deviation.

On the seven-class benchmark, the CNN and ResNet-18 reached almost identical accuracy (64.4% and 64.3%), close to the estimated human agreement of about 65% [6] but below the 73.3% reported for a heavily tuned VGG-style network [20]. Their macro-F1 scores differed more (56.0% versus 61.3%) because of a single class. The CNN almost never predicted *disgust*, which has only 436 training images, whereas ResNet-18 recognized about half of the disgust test images (Figure 6). The Vision Transformer reached 40.8% accuracy, and its macro-F1 of 30.6% reflects a strong bias toward the frequent happy class.

The training curves in Figure 2 explain part of the transformer’s weakness. The CNN and ResNet-18 exceeded 70% validation accuracy within about ten epochs, while the transformer’s validation accuracy rose slowly and was still improving when training stopped at 30 epochs. Its training and validation curves stayed close together, which points to under-fitting rather than over-fitting. Without pretraining, the transformer has to learn from 17,010 small images what convolutional layers encode by design, namely that nearby pixels belong together and that the same pattern can appear anywhere in the image.

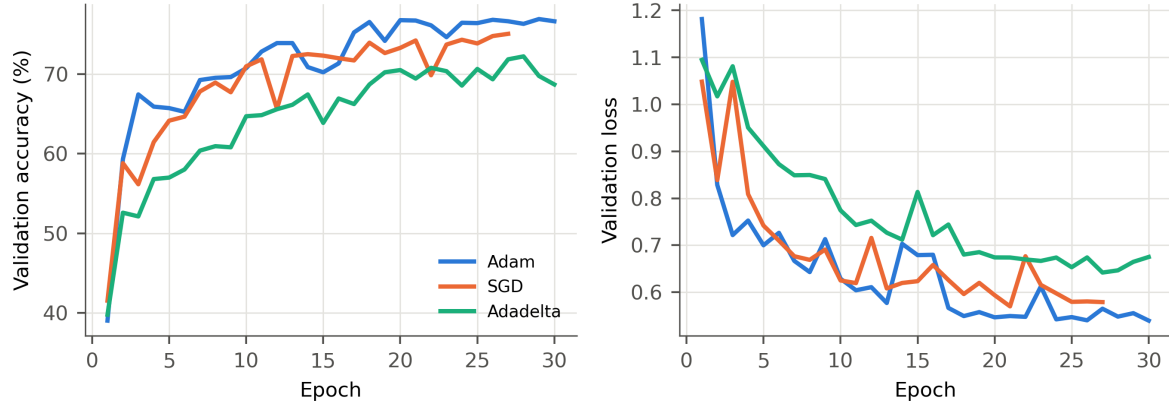
**Figure 2:** Training (dashed) and validation (solid) accuracy and loss of the three models on the three-class task (seed 0).

4.2 Optimizer Comparison

Table 4 and Figure 3 compare the optimizers used in the original study for the proposed CNN. Adam gave the best result (80.1% test accuracy), followed by SGD (78.4%) and Adadelata (72.6%). SGD reached its best validation loss earlier, at epoch 21, and stopped early, but its validation curve was noisier than Adam’s. Adadelata improved steadily but slowly and ended about 7 points below Adam. The ranking agrees with the conclusion of the original manuscript, but the corrected evaluation places all three optimizers at a higher level and shows that the gap between Adam and SGD is small.

Table 4: Optimizer comparison for the proposed CNN (three-class task, seed 0)

Optimizer (learning rate)	Validation acc. (%)	Test acc. (%)	Test macro-F1 (%)	Best / run epochs
Adam (10^{-3})	76.6	80.1	78.3	30 / 30
SGD (0.1)	74.2	78.4	76.6	21 / 27
Adadelata (0.1)	71.8	72.6	70.5	27 / 30

**Figure 3:** Validation accuracy and loss of the proposed CNN trained with Adam, SGD, and Adadelata (three-class task, seed 0).

4.3 Per-Class Errors

The three expressions were not equally easy. Averaged over seeds, the CNN recognized happy faces with an F1 of 91.2%, but neutral and sad faces with 72.1% and 69.8% (Table 5, Figure 5). ResNet-18 showed the same pattern. The confusion matrices in Figure 4 show where the errors lie: most mistakes were between sad and neutral. For the CNN, 23% of sad faces were labeled neutral and 19% of neutral faces were labeled sad, while fewer than 10% of either class were mistaken for happy. The transformer's errors were different in kind. It labeled 39% of sad faces and 38% of neutral faces as happy, which indicates that it had learned little beyond the dominant class.

Table 5: Per-class precision, recall, and F1-score on the three-class task (% , mean over 3 seeds)

Class (test n)	Proposed CNN			ResNet-18			Vision Transformer		
	P	R	F1	P	R	F1	P	R	F1
Happy (879)	91.3	91.1	91.2	91.0	88.2	89.6	60.2	84.3	70.2
Sad (594)	70.2	69.6	69.8	67.0	70.8	68.8	50.7	35.0	41.4
Neutral (626)	71.7	72.4	72.1	70.6	69.8	70.1	59.5	43.6	50.3

Note: P = precision; R = recall. Test images per class in parentheses.

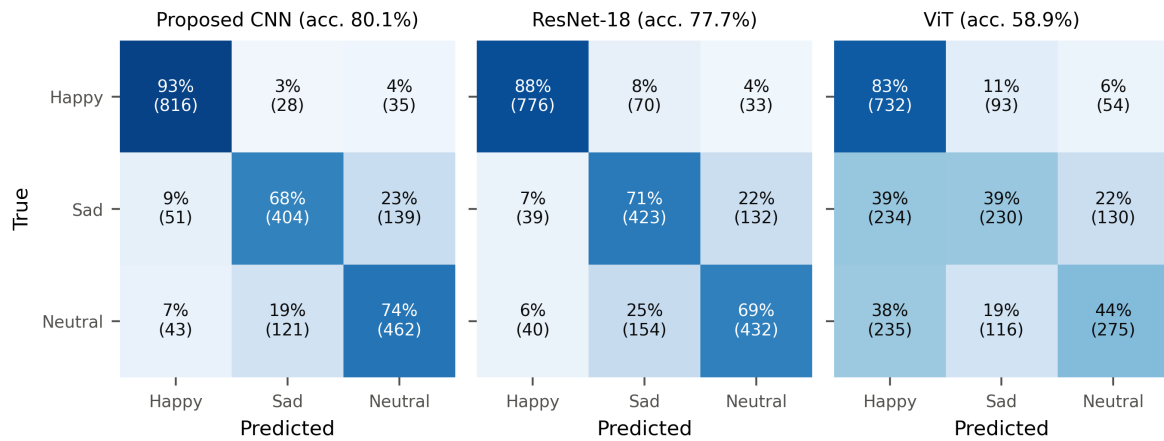


Figure 4: Row-normalized confusion matrices on the three-class test set (seed 0). Cells show the percentage of each true class and, in parentheses, the number of images.

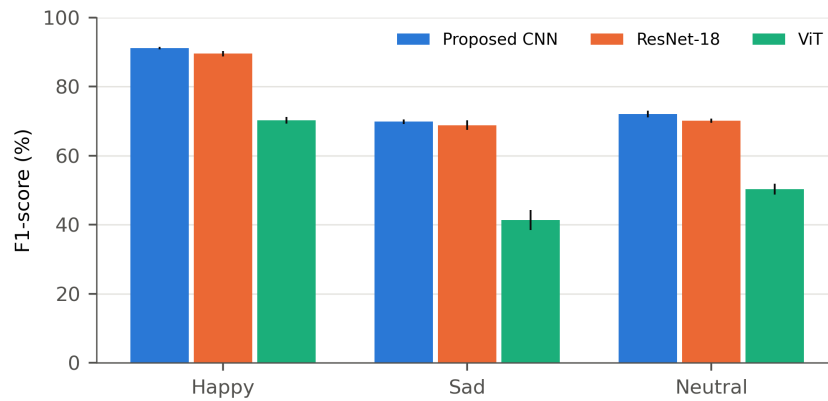


Figure 5: Per-class F1-scores on the three-class task; bars show the mean over three seeds and whiskers one standard deviation.

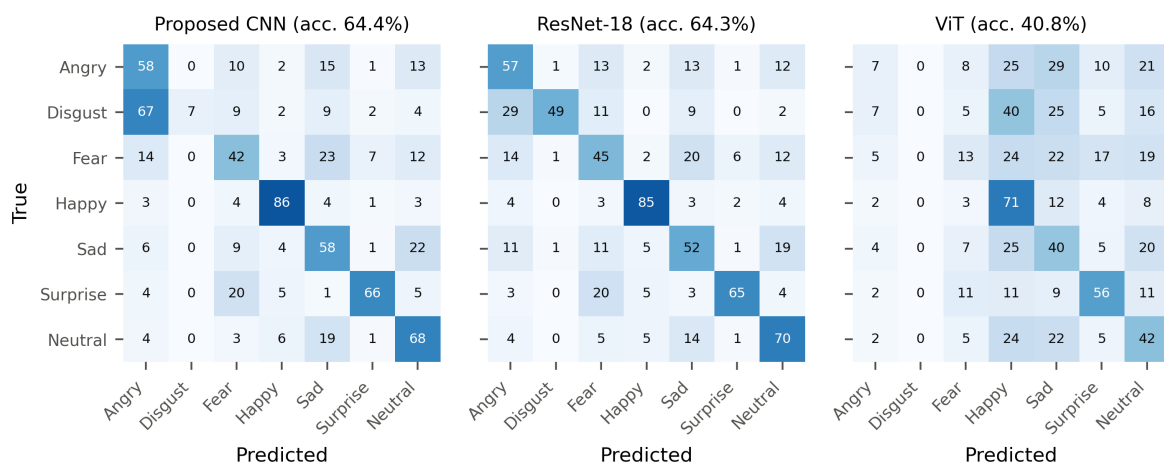


Figure 6: Row-normalized confusion matrices on the seven-class test set (percent of each true class; seed 0).

Figure 7 shows typical correct predictions and the CNN’s most confident errors. Several of these errors are not clearly model failures. Faces labeled sad but showing an open-mouthed, laugh-like expression were classified as happy with over 95% confidence, and several “neutral” faces carry a visible smile.

This agrees with the known label noise of FER-2013 [6, 23] and suggests that part of the remaining error lies in the data rather than in the model.

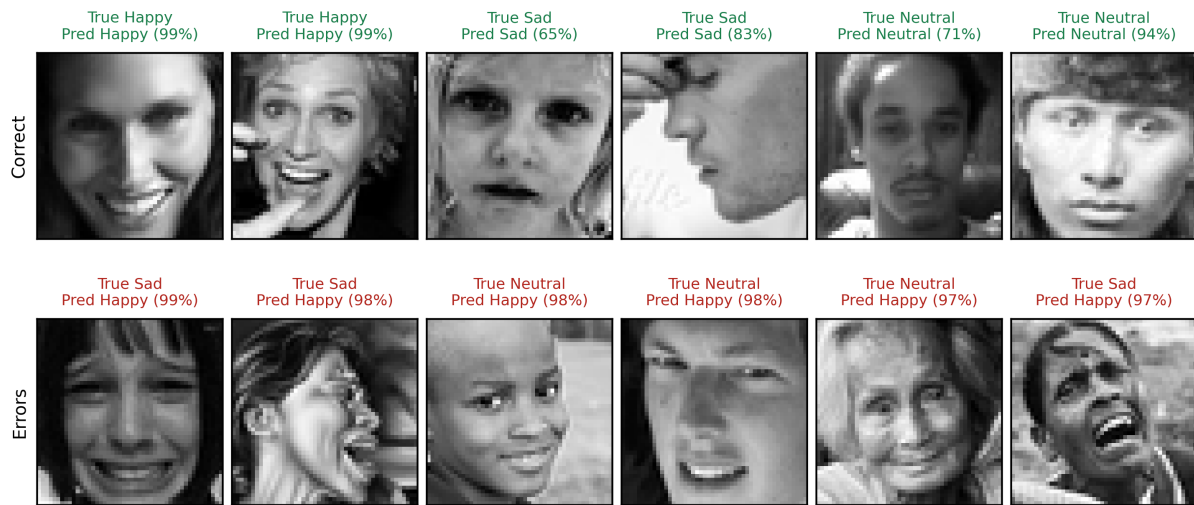


Figure 7: Examples of correct predictions (top) and the most confident errors (bottom) of the proposed CNN on the three-class test set, with the predicted probability in parentheses.

4.4 Attention Maps and Efficiency

Grad-CAM maps (Figure 8) show where the two convolutional models looked. For happy faces, both concentrated on the mouth and the cheeks, where smiles are expressed. For sad and neutral faces, the CNN focused on the central region around the nose, eyes, and upper lip. ResNet-18 attended more to the lower face and chin for neutral faces and, in some sad examples, to regions at the edge of the face or to a hand. These maps indicate that the models rely mainly on expression-related facial regions, but also that ResNet-18 sometimes uses contextual cues that would not transfer reliably to classroom images.

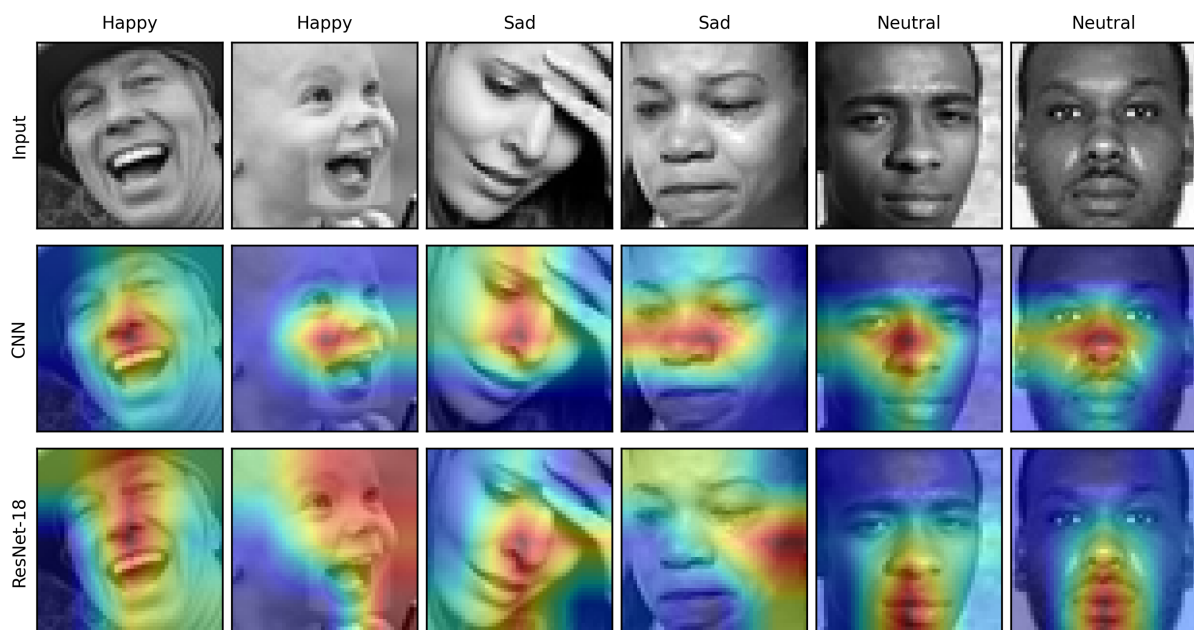


Figure 8: Grad-CAM attention maps of the proposed CNN and ResNet-18 for correctly classified test images (red = high relevance).

All three models are small enough for real-time use on ordinary hardware (Table 2). On the same two-core CPU used for training, classifying a single face took a median of 2.5 ms for the CNN, 3.1 ms for ResNet-18, and 5.4 ms for the Vision Transformer. Even the slowest model could therefore process several faces per video frame at typical frame rates, and the cost of face detection would dominate the pipeline.

5 Discussion

Under a controlled, from-scratch protocol, the compact eight-layer CNN was the most accurate of the three models on the three-class task and matched ResNet-18 on the seven-class benchmark, while being the fastest and among the smallest. This result runs against the general trend in which deeper residual networks and transformers outperform plain CNNs [7], but it is consistent with what those architectures need. Residual networks gain most from depth and large datasets, and transformers from pretraining at scale [29, 8]. With 17,010 grayscale images of 48×48 pixels and no pretraining, the inductive bias of convolutions is an advantage, and the Vision Transformer’s weak performance is what one would expect rather than evidence that attention is unsuited to expression recognition. The published transformer results that exceed CNNs on FER benchmarks rely on pretrained backbones, larger inputs, or hybrid designs [9, 11, 24]. A fair conclusion is that, for institutions without GPUs or pretraining pipelines, a small CNN remains a strong baseline, while transformer-based models should be adopted only with suitable pretraining.

The corrected evaluation also changes the substance of the original study. Under the new protocol, the proposed network reaches about 79% accuracy on the held-out test set for happy, sad, and neutral faces, and the confusion matrices are consistent with that figure. In the original manuscript they implied accuracy near chance. The per-class analysis shows that the headline accuracy hides an important asymmetry. Happy faces are recognized reliably, but sad and neutral faces are confused with each other in roughly one case in five. This asymmetry matters for education, because a system meant to alert a teacher to discouraged students would be least reliable for exactly the expression of interest. Part of the confusion reflects genuinely ambiguous faces and noisy labels [6, 23], and part reflects the fact that sadness and neutrality are expressed more subtly than a smile. Label-noise-robust training [22], cleaner datasets such as RAF-DB [12], and temporal information from video would all be expected to help.

The attention maps give some reassurance that the convolutional models rely on the mouth and central facial regions, but they also show occasional reliance on context, such as a hand near the face, which would not generalize to webcam images of students. More fundamentally, a correctly classified facial configuration is not the same as a correctly inferred emotional state. The mapping from facial movements to emotions varies across people, cultures, and situations [15], FER performance can differ across demographic groups [16], and students may express confusion or boredom with faces that FER-2013 would label neutral. For these reasons, and because Article 5(1)(f) of the EU Artificial Intelligence Act prohibits inferring the emotions of people in education institutions except for medical or safety reasons [17], the models evaluated here should not be deployed to monitor individual students. Where the law permits their use at all, it should be limited to opt-in, aggregate, and anonymous signals that inform rather than replace a teacher’s judgment, and students should be told clearly how such signals are collected and used.

The study has several limitations. All models were trained from scratch on a CPU, which limited their size and the number of epochs; the proposed CNN and the Vision Transformer were still improving at the 30-epoch limit, so their absolute accuracy is likely somewhat underestimated. The seven-class benchmark and the optimizer comparison were run with a single seed. FER-2013 contains duplicate and nearly blank images and noisy labels, and it consists of posed and spontaneous web images rather than classroom recordings, so the reported accuracies may not transfer to students in real learning settings. Face detection was not evaluated, although in practice its errors would add to those of the classifier. Finally, no demographic attributes are available for FER-2013, so the fairness of the models could not be assessed.

6 Conclusion

This study re-evaluated a CNN for recognizing students' facial expressions under a reproducible protocol and compared it with a ResNet-18 and a Vision Transformer trained from scratch on the official FER-2013 partitions. The compact CNN recognized happy, sad, and neutral faces with about 79% accuracy on held-out data, slightly ahead of ResNet-18 and far ahead of the transformer, and it matched ResNet-18 on the seven-class benchmark at about 64%, with inference times of a few milliseconds on an ordinary CPU. Happy faces were recognized reliably, whereas sad and neutral faces were often confused, and Grad-CAM showed that predictions relied mostly on the mouth and central face. The findings show that careful evaluation, including held-out testing, repeated runs, and verified confusion matrices, matters more than architectural novelty for small-scale expression recognition. They also show that a compact CNN is a sound starting point when pretraining and GPUs are unavailable. Any educational use should nonetheless be limited to aggregate, consent-based feedback, because recognized expressions are an uncertain guide to students' emotions and emotion inference in education is now legally restricted in some jurisdictions.

Data and Code Availability

FER-2013 is publicly available through the ICML 2013 Kaggle challenge. The training, evaluation, and figure-generation code used in this study is available from the corresponding author.

References

- [1] L. B. Krithika and G. G. Lakshmi Priya, "Student emotion recognition system (SERS) for e-learning improvement based on learner concentration metric," *Procedia Computer Science*, vol. 85, pp. 767–776, 2016.
- [2] U. Ayvaz, H. Gürüler, and M. O. Devrim, "Use of facial emotion recognition in e-learning systems," *Information Technologies and Learning Tools*, vol. 60, no. 4, pp. 95–104, 2017.
- [3] I. Lasri, A. R. Solh, and M. El Belkacemi, "Facial emotion recognition of students using convolutional neural network," in *Proceedings of the Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, 2019, pp. 1–6.
- [4] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2132–2143, 2022.
- [5] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
- [6] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2015.
- [7] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2022.
- [8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: <https://arxiv.org/abs/2010.11929>

- [9] F. Xue, Q. Wang, and G. Guo, "TransFER: Learning relation-aware facial expression representations with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3601–3610.
- [10] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *Biomimetics*, vol. 8, no. 2, p. 199, 2023.
- [11] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, A. Huang, and Y. Wang, "POSTER++: A simpler and stronger facial expression recognition network," *Pattern Recognition*, vol. 157, p. 110951, 2025.
- [12] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2584–2593.
- [13] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019.
- [14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.
- [15] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, 2019.
- [16] T. Xu, J. White, S. Kalkan, and H. Gunes, "Investigating bias and fairness in facial expression recognition," in *Computer Vision – ECCV 2020 Workshops*, ser. Lecture Notes in Computer Science, vol. 12540. Springer, 2020, pp. 506–523.
- [17] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of 13 june 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, L series, 2024. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- [18] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
- [19] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5202–5211.
- [20] Y. Khairuddin and Z. Chen, "Facial emotion recognition: State of the art performance on FER2013," arXiv preprint arXiv:2105.03588, 2021. [Online]. Available: <https://arxiv.org/abs/2105.03588>
- [21] P. Jiang, B. Wan, Q. Wang, and J. Wu, "Fast and efficient facial expression recognition using a Gabor convolutional network," *IEEE Signal Processing Letters*, vol. 27, pp. 1954–1958, 2020.
- [22] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6896–6905.
- [23] C. Mejia-Escobar, M. Cazorla, and E. Martínez-Martín, "Towards a better performance in facial expression recognition: A data-centric approach," *Computational Intelligence and Neuroscience*, vol. 2023, p. 1394882, 2023.
- [24] N. Li, Y. Huang, Z. Wang, Z. Fan, X. Li, and Z. Xiao, "Enhanced hybrid vision transformer with multi-scale feature integration and patch dropping for facial expression recognition," *Sensors*, vol. 24, no. 13, p. 4153, 2024.

- [25] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagement from facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.
- [26] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards user engagement recognition in the wild," arXiv preprint arXiv:1609.01885, 2016. [Online]. Available: <https://arxiv.org/abs/1609.01885>
- [27] M. A. A. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: A review," *Smart Learning Environments*, vol. 6, no. 1, pp. 1–20, 2019.
- [28] F. Begum, A. Neelima, and J. A. Valan, "Emotion recognition system for e-learning environment based on facial expressions," *Soft Computing*, vol. 27, no. 22, pp. 17 257–17 265, 2023.
- [29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: <https://arxiv.org/abs/1412.6980>