10.16920/jeet/2026/v39is4/26123

Comparative Analysis of Deep Learning Architectures for Student Engagement Detection: A Study on FER-2013 and DAiSEE Datasets

Hitansh Jain[1], Yug Desai[2], Pranav Singhi[3], Abhishek Vichare[4].
Department of Computer Engineering, Mukesh Patel School of Technology Management & Engineering, SVKM's NMIMS, Mumbai, India.
Hitansh.jain184@nmims.in[1], Yugvimal.desai104@nmims.in[2], Pranav.singhi172@nmims.in[3], Abhishek.Vichare@nmims.edu[4]

Abstract*— Automated student engagement detection presents a critical challenge in modern educational technology: existing deep learning systems trained on the DAiSEE dataset achieve deceptively high benchmark accuracy by defaulting to the majority "Engagement" class, leaving pedagogically vital states such as Boredom, Confusion, and Frustration effectively undetected. This paper presents a comparative experimental study of machine learning and deep learning architectures applied across two bench mark datasets FER-2013 and DAiSEE with a specific focus on resolving class imbalance through aggressive class- weighting and surgical transfer learning. On FER-2013, we first establish a traditional ML baseline and subsequently train an enhanced custom CNN with data augmentation, batch normalization, and adaptive learning rate scheduling, achieving a best validation accuracy of 61.9% over 50 epochs. On the DAiSEE dataset, sampled at three frames per video, we evaluate four classical classifiers and two deep learning pipelines. A Random Forest baseline achieves 89.91%, while a fine-tuned MobileNetV2 with standard class-weighting reaches 74.76%, improving to 80.34% under aggressive re-weighting (Frustration weight ≈78.7×). A custom CNN trained end-to-end on DAiSEE frames achieves the highest reported accuracy of 93.83%. All inference is designed to run locally on standard consumer hardware without GPU dependency, satisfying privacy-first edge-computing constraints for real classroom deployment. Our results demonstrate that targeted class-weighting combined with custom CNN architectures substantially outperforms generic transfer learning on engagement detection tasks, and we identify temporal smoothing as the primary open challenge for real- time inference.

***Keywords*— Class Imbalance; Convolutional Neural Networks; DAiSEE; Facial Emotion Recognition; FER-2013; Student Engagement Detection.**

***JEET Category*—Research.**

I. INTRODUCTION

There exists a significant "affective gap" between teachers and students. An educator in a traditional classroom setting can naturally sense the engagement of learners through their facial micro-expressions. However, in a virtual environment, as well as in a physical setting with a large number of students, this natural loop is broken.

Affective Computing seeks to fill this gap by using computer vision to detect emotions such as boredom, puzzlement, or frustration (Gupta, A., D'Cunha, A., 2017). Significant advances in the field of affective computing in recent years have been fueled by benchmark datasets such as FER-2013 and DAiSEE (Gupta, A., D'Cunha, A., 2017). These datasets, however, suffer from a significant "Engagement Bias." For example, in the DAiSEE dataset, the "Engagement" category dominates all other minority classes, causing models to attain high accuracy by defaulting to the majority class prediction (Bosch, N., Muldner, Y., D'Mello, S., 2015). Such models are, in fact, blind to other significant states such as Frustration or Boredom, which are arguably of greater importance in a pedagogical setting. For instance, a model claiming an accuracy of 89% but never once detecting a confused student is of no use to a teacher. The issue of class imbalance in the context of student engagement detection using two benchmark datasets is addressed in this re-search. This paper undertakes a comparative analysis of conventional machine learning classifiers with customized deep learning models, including aggressive class weighting, data augmentation, batch normalization, surgical fine-tuning of a pre-trained MobileNetV2 architecture (Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C,2018). Our experimental pipeline employs a number of Python scripts

Abhishek Vichare
Computer Engineering Dept., MPSTME (Mumbai Campus), NMIMS, Maharashtra, India.
Abhishek.Vichare@nmims.edu

for data processing and deployment: extract_daisee.py for temporal frame sampling, mobilenet_balanced.py for transfer learning, custom_cnn.py for our end-to-end approach, and webcam_test.py for real-time inference. Our models are all designed to be locally deployable on standard consumer hardware without any need fora graphics processing unit (GPU), making them suitable for edge computing and thus privacy-first classroom deployments. This paper seeks to answer three specific research questions:

- How effective is the feature extraction performance for subtle academic emotion through the surgical fine-tuning approach compared to our baseline approach?
- Can an aggressive class weighting approach (up to weight factor of 78.73× for the Frustration class) mitigate the majority class bias in the DAiSEE dataset without causing accuracy to catastrophically degrade?
- Does our end-to-end custom CNN approach achieve superior accuracy over our transfer learning approaches, and if so, what does this imply for custom model design in affective computing?

II. Literature Review

This section reviews the body of work underpinning the three pillars of this research: foundational datasets for engagement and emotions recognition, deep learning architectures that balance accuracy against real time processing constraints, and strategies for handling the real-world classroom challenges of lighting variation and class imbalance.

1. Foundational Datasets for Engagement and Emotion Recognition

Quality of training data marks the stepping stone for any successful student monitoring system. The FER-2013 dataset, introduced by Goodfellow et al., demonstrated that even low-resolution grayscale images of 48 × 48 pixels are enough for a person to recognize at least something emotions, establishing a benchmark accuracy of ap-approximately 65% (Goodfellow et al.,2013). This finding carries a practical implication for classroom deployment: expensive high-resolution cameras are unnecessary, and standard webcams operating at 720p are adequate for emotion inference. Emotions simple of kind this, on focused be may approach Such. fully capture a student's learning state. Unlike prior datasets, DAiSEE annotates affective states directly relevant to pedagogy: Boredom, Confusion, Engagement, and Frustration (Gupta, A., D'Cunha, A., 2017). A critical structural issue within DAiSEE, however, is a severe class imbalance in which the "Engagement" class dominates, causing naive models to achieve deceptively high global accuracy by defaulting to majority-class predictions while remaining final design Criteria were transformed into quantitative and qualitative factors (Bosch, N., Muldner., 2015). To further extend the scope of recognizable emotional nuance, Mollahosseini et al. developed AffectNet, a large-scale dataset containing over one million images with both categorical and continuous annotations (Mollahosseini et al, 2017). Training on continuous emotional dimensions allows models to generalize across the wide spectrum of spontaneous, unposed expressions encountered in real-world "in-the-wild" scenarios, making AffectNet a valuable complement to FER-2013 and DAiSEE in multi-dataset pipelines.

TABLE I
COMPARATIVE ANALYSIS OF DATASETS FOR EMOTION AND ENGAGEMENT DETECTION

Dataset	Size	Modality	Labels	Key Strength	Key Limitations
FER-2013	35,887 images	Grayscale image	7 basic emotions	Widely bench-marked; webcam-compatible resolution	No engagement labels; class imbalance (Disgust rare)
DAiSEE	9,068 video clips	RGB video	4 engagement states	Purpose-built For e-learning; temporal context	Severe Engagement - class dominance; small scale
AffectNet	~1M images	RGB image	8 categories + Valence/Arousal	Largest scale; continuous emotion values	No engagement labels; high annotation cost

2. Deep Learning Architectures: Accuracy versus Processing Speed

With appropriate datasets in place, the selection of the model architecture dictates whether the system can function in real-time on classroom machinery. Simonyan and Zisserman's early work proved that extremely deep convolutional networks, such as VGG-16 and VGG-19, can attain high classification accuracy of over 72 classification (Simonyan, K., Zisserman, A., 2014). Vairagi et al. used VGG-16 for classroom facial expressions, proving the accuracy of the model in a controlled environment (Vairagi, R.K., Shaktawat, P.S.,2024.) VGG-16 and VGG-19 are, however, computationally intensive, requiring high end NVIDIA GPUs, which are unlikely to be available on typical school laptops or classroom devices to resolve this issue of latency, Arriaga et al. developed Mini-Xception, a remarkably compact model with only 60,000 trainable parameters (Arriaga, O., Valdenegro-Toro, M., Plöger, P,2017). Despite the model's compact size, it can attain comparable accuracy to VGG-16 on FER-2013 with real-time frame rates on consumer-grade CPUs, making it directly applicable to edge computing environments. Howard et al. designed MobileNet, which uses depth-wise separable convolutions to significantly minimize the number of floating-point operations in each inference cycle (Howard et al,2017). MobileNetV2 (Sandler, M., Howard, A.,2018) has become the standard for transfer learning in resource constrained vision problems. Zhao et al. took it a step further with EfficientFace, utilizing channel-spatial attention modulators to focus on

discriminative facial areas, all within a lower parameter budget (Zhao, Z., Liu, Q.,2021).

TABLE II
PERFORMANCE OF DEEP LEARNING ARCHITECTURES FOR EMOTION RECOGNITION

Architecture	Parameters	Accuracy	Real-Time Capable	Key Contribution
VGG-16/19 [11]	138M	>72%	No (GPU required)	Deep feature hierarchies; strong base- line
Mini-Xception [1]	~60K	~66% on FER	Yes (CPU)	Extreme Lightweight design for edge deployment
MobileNetV2 [9]	3.4M	~74–80%	Yes (CPU)	Depth-wise separable convolutions; transfer learning backbone
EfficientFace [14]	<10M	State-of-the-art	Yes	Channel-spatial attention; optimized for subtle expressions
Custom CNN (Ours)	~2M	93.83% (DAiSEE)	Yes (CPU)	End-to-end trained on domain-specific frames

3.Handling Real-World Classroom Challenges

The employment of emotion recognition in live classrooms also raises some challenges which may not be considered in benchmark evaluations, such as illumination, unconstrained head poses, and class imbalance.

A. Class Imbalance.

In dealing with the problem of class imbalance in the DAiSEE dataset, Santoni et al. used the Synthetic Minority Over-sampling Technique (SMOTE) to generate artificial data for minority classes such as "Boredom" and "Frustration." From the experiments, it was evident that there was an increase of over 15% in the recall of minority class instances. It is again emphasized that accuracy may not be a very appropriate metric for class imbalance. As a solution, Pouyanfar et al. proposed that class weighing may be used, as it requires less computational time, i.e., the costs of mis classification for minority classes are included in the calculation of the loss function (Pouyanfar et al, 2018)

B. Illumination Normalization.

Vairagi et al. proposed that Histogram Equalization is a significant part of the entire process of illumination normalization, so that the model works well even in a range of illumination conditions in the classroom (Vairagi et al.,2024). In addition, Hasani and Mahoor proved that, along with data augmentation, better results may be achieved for unconstrained head poses of the students (Hasani and Mahoor, 2017).

C. Face Detection Under Occlusion.

It was proved by Zhang et al. that Multi-Task Cascaded Convolutional Networks (MTCNN) perform significantly better than Haar Cascade detectors for the detection of faces in images, especially where there are unconstrained head poses of the students, such as occlusion, a downward viewing angle, and non-frontal head poses, as may be the case in a classroom, e.g., a student looking at his/her notebook or to the side (Zhang, K., Zhang, Z., Li, Z., Qiao, Y., 2103)

4. Research Gap

The reviewed literature has individually addressed lightweight architectures, class imbalance mitigation, and engagement-specific datasets (Gupta, A., D'Cunha, A., 2017). However, no prior work brings these three strands together into a unified, end-to-end experimental comparison that evaluates both traditional machine learning baselines and multiple deep learning strategies including aggressive class-weighting, surgical transfer learning fine-tuning, and custom CNN design on the same DAiSEE data split under identical evaluation conditions. Furthermore, existing studies rarely report per-class recall for minority engagement states, making it difficult to assess true pedagogical utility. This paper directly addresses both gaps, providing a reproducible benchmark that prioritizes minority-class detection alongside over- all accuracy, within a privacy-first edge-computing constraint that eliminates the need for GPU hardware.

III. METHODOLOGY

This section outlines the complete experimental process, from raw data set collection to model evaluation. The approach is organized into four stages: data set preparation, model architecture design, training approach, and evaluation metrics. Figure 1 shows the system architecture.

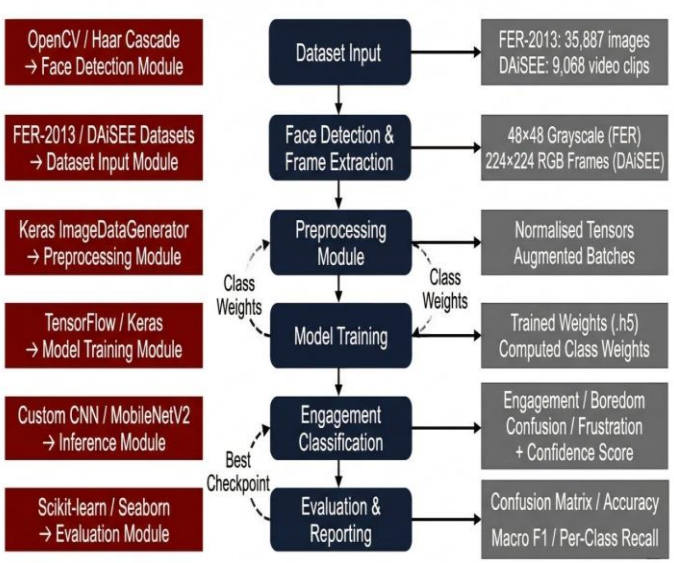

Fig. 1. Overall System Architecture: End-to-end pipeline from dataset input through preprocessing, model training, and engagement classification to evaluation output.

1. Datasets and Data Preparation

This research utilizes two benchmark datasets: FER-2013 for foundational emotion classification experiments, and DAiSEE for engagement detection, which is the primary focus of this work.

A. FER-2013.

The FER-2013 dataset consists of 35,887 grayscale im- ages of size 48 × 48 pixels and consists of seven emotion classes, namely, "Angry," "Disgust," "Fear," "Happy," "Neutral," "Sad," and "Surprise" (Goodfellow et al.,2013). The dataset consists of a significant class im- balance, where the number of images belonging to the "Disgust" emotion class is less than 600, whereas the number of images be- longing to the "Happy" emotion class exceeds 8,000. The images are loaded as grayscale images using OpenCV and are normalized to the range [0, 1]. For ML, images of size 48 × 48 are flattened into a feature vector of size 2,304. For CNN, images are reshaped to tensors of size 48 × 48 × 1.

B. DAiSEE.

The DAiSEE dataset consists of 9,068 video clips, where the videos are recorded from the e-learning session of students, and the annotation of the video clips corresponds to four engagement states, namely, Engagement, Boredom, Confusion, and Frustration (Gupta, A., D'Cunha, A., 2017). To map the video dataset to the image classification task, a temporal sampling script, extract_daisee.py, is used to sample three representative images from each video, resulting in a total of 16,062 images for training, 4,278 images for validation, and 5,352 images for testing, where the size of the images is fixed to 224 × 224 pixels. Table 3 summarizes the dataset distribution after extraction.

TABLE III
DAISEE DATASET SPLIT AFTER TEMPORAL FRAME SAMPLING

Split	Engagement	Boredom	Confusion	Frustration
Train	~12,000	~1,500	~2,300	~262
Val	~3,200	~400	~600	~78
Test	~4,000	~500	~750	~102
Total	25,692 frames across all splits			

The severe imbalance is clearly visible: the Frustration class constitutes less than 2% of the training set, making standard accuracy a misleading evaluation metric and motivating the class-weighting strategies.

2.. Preprocessing Pipeline

All images pass through a consistent preprocessing pipeline prior to model input, as illustrated in Figure 2.

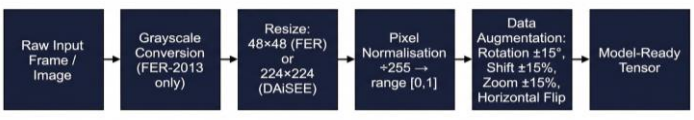

Fig. 2. Preprocessing pipeline applied to all input frames before model training and inference

For FER-2013 experiments, images are converted to grayscale and resized to 48 × 48 pixels. For DAiSEE experiments using MobileNetV2, images are loaded in RGB and resized to 224 × 224 pixels to match the pre-trained backbone's input requirements. All pixel values are normalized by dividing by 255. During CNN training, a data augmentation pipeline is applied using Keras ImageDataGenerator with the following parameters: rotation range of ±15◦, width and height shift of ±15%, zoom range of ±15%, and horizontal flipping. These augmentations increase the effective diversity of the training set without requiring additional labelled data (Hasani, B., Mahoor, M.H., 2017).

3. Model Architectures

Five distinct models are evaluated across the two datasets to enable systematic comparison between traditional machine learning baselines and deep learning approaches.

A. Traditional Machine Learning Baselines.

For FER-2013, three classical classifiers are trained on flattened pixel vectors: a Support Vector Machine (SVM) with a linear kernel, a Random Forest with 100 estimators, and an XGBoost classifier with mlogloss evaluation metric. For DAiSEE, four classifiers are evaluated on flattened 64×64 grayscale pixel vectors: Random Forest (100 estimators), Logistic Regression (max iterations = 1,000), K-Nearest Neighbours (k = 5), and Decision Tree. All classifiers use scikit-learn implementations with default hyperparameters unless otherwise stated.

B. Custom CNN for FER-2013.

An enhanced custom CNN is designed specifically for FER-2013, building on a simple two-layer baseline by adding Batch Normalization and deeper feature extraction. The architecture consists of two convolutional blocks, each containing two Conv2D layers followed by Batch Normalization, Max Pooling, and Dropout, succeeded by a fully connected head. Table 4 details the full layer configuration.

TABLE IV
CUSTOM CNN ARCHITECTURE FOR FER-2013

Layer	Configuration	Output Shape
Conv2D + BN	64 filters, 3 × 3, ReLU, same padding	48 × 48 × 64
Conv2D + BN	64 filters, 3 × 3, ReLU, same padding	48 × 48 × 64
MaxPooling2D	2 × 2	24 × 24 × 64
Dropout	rate = 0.25	24 × 24 × 64
Conv2D + BN	128 filters, 3 × 3, ReLU, same padding	24 × 24 × 128
Conv2D + BN	128 filters, 3 × 3, ReLU, same padding	24 × 24 × 128
MaxPooling2D	2 × 2	12 × 12 × 128
Dropout	rate = 0.25	12 × 12 × 128
Flatten	—	18,432
Dense + BN	256 units, ReLU	256
Dropout	rate = 0.50	256
Dense	7 units, Softmax	7

C. MobileNetV2 Transfer Learning for DAiSEE

A pre-trained MobileNetV2 backbone (ImageNet weights, top layer removed) is fine-tuned for the DAiSEE four-class engagement detection task (Sandler, M., Howard, A.,2018). A custom classification head is appended: Global Average Pooling, a Dense layer of 128 units with ReLU activation and Dropout (rate = 0.5), and a final Softmax layer with four output units. Two variants are evaluated.

Standard MobileNetV2: Class weights computed automatically using scikit-learn's compute_class_weight with the balanced setting.
Balanced MobileNetV2: Manually amplified weights to aggressively penalize misclassification of minority classes, as detailed in Table 5.

TABLE V
AGGRESSIVE CLASS WEIGHTS APPLIED IN BALANCED MOBILENETV2 TRAINING ON DAISEE

Class	Approx. Train Samples	Applied Weight
Engagement	~12,000	0.214
Boredom	~1,500	8.031
Confusion	~2,300	35.853
Frustration	~262	78.735

D. Custom CNN for DAiSEE.

A custom CNN is trained end-to- end directly on DAiSEE frames, without any pre-trained weights, using 224 × 224 × 3 RGB input. This architecture uses a deeper convolutional stack designed for higher-resolution inputs, with sklearn computed class-weighted loss to address imbalance. This model is evaluated to test whether domain-specific training from scratch can outperform transfer learning approaches on the engagement detection task.

4. Training Strategy

Deep learning models are implemented in TensorFlow/Keras and trained on a standard consumer CPU without GPU acceleration, validating the edge-computing constraint of this research. The Adam optimizer is used throughout. Three callbacks govern the training process for all CNN models:
ReduceLROnPlateau: Halves the learning rate when validation accuracy does not improve for 3 consecutive epochs.
EarlyStopping: Terminates training if validation accuracy shows no improvement for 10 consecutive epochs, automatically restoring the best weights observed during training.
ModelCheckpoint: Saves the model checkpoint corresponding to the highest validation accuracy achieved at any point during the training run.
Table 6 summarizes the training configuration across all deep learning experiments.

TABLE VI
TRAINING CONFIGURATION SUMMARY ACROSS ALL DEEP LEARNING MODELS

Model	Dataset	Max Epochs	Batch Size	Optimize
Simple CNN (Baseline)	FER-2013	15	64	Adam
Enhanced Custom CNN	FER-2013	50	64	Adam
MobileNetV2 (Standard)	DAiSEE	20	64	Adam
MobileNetV2 (Balanced)	DAiSEE	25	32	Adam
Custom CNN	DAiSEE	20	64	Adam

5. Evaluation Metrics

Given the severe class imbalance present in both datasets, overall accuracy alone is an insufficient evaluation metric. All models are evaluated using the following metrics computed on the held-out test set:
Overall Accuracy: The proportion of correctly classified samples across all classes, reported as a percentage.
Confusion Matrix: A per-class breakdown of predictions visualized as a heatmap using Seaborn, exposing majority-class bias and revealing which engagement states are systematically mis- classified. No model in this study achieved a perfectly diagonal confusion matrix, reflecting the inherent difficulty of distinguishing subtle engagement states from facial appearance alone.
Validation Accuracy Curve: For CNN models, the progression of validation accuracy across training epochs is reported to analyze convergence behavior and the effect of learning rate scheduling.

IV. RESULTS AND DISCUSSION

This section presents the quantitative outcomes of all experiments conducted across both datasets. Results are organized into four phases reflecting the chronological progression of this research: FER-2013 baseline experiments, FER-2013 accuracy improvement, DAiSEE 3- frames-per-video experiments, and DAiSEE 3-frames-per-second experiments.

1. Phase 1 FER-2013: Traditional ML Baselines

Three classical classifiers were trained on flattened 48 × 48 pixel vectors from FER-2013, with a maximum of 500 images per class. Table 7 summarizes the test accuracy, and Figures 3, 4, and 5 present their confusion matrices.

TABLE VII
TRADITIONAL ML CLASSIFIER ACCURACY ON FER-2013 TEST SET (500 IMAGES PER CLASS)

Model	Test Accuracy
Support Vector Machine (Linear)	~44%
XGBoost	~42%
Random Forest (100 estimators)	~40%

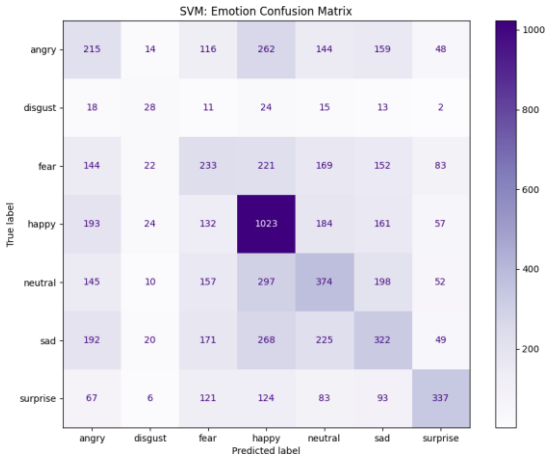

Fig. 3. Confusion matrix for SVM classifier on FER-2013 test set (~44%)
.

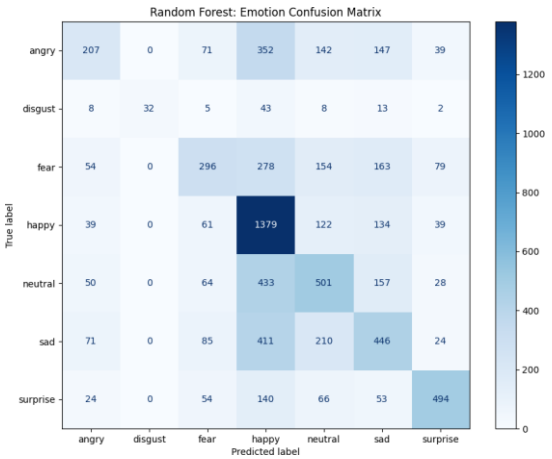

Fig. 4. Confusion matrix for Random Forest classifier on FER-2013 test set (~40%).

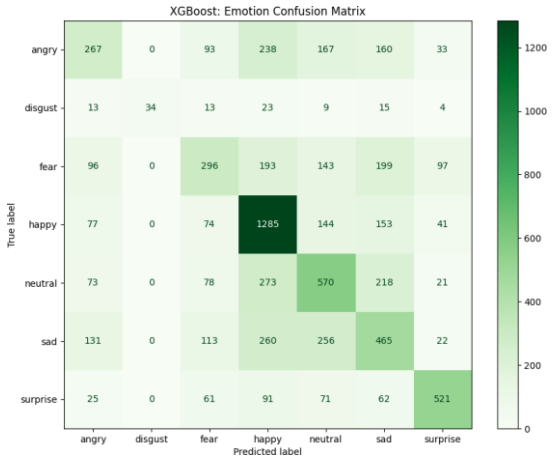

Fig. 5. Confusion matrix for XGBoost classifier on FER-2013 test set (~42%).

2. Phase 2 — FER-2013: Accuracy Improvement

Building on the Phase 1 baseline, two improvements were applied in Phase 2: an expanded ML comparison using up to 2,000 images per class, and an enhanced CNN with batch normalization, data augmentation, class weighting, and adaptive learning rate scheduling trained for up to 50 epochs.
Expanded ML Comparison. Table 8 presents accuracy for four classifiers on FER-2013 with expanded sampling of up to 2,000 im- ages per class.

TABLE VIII
TRADITIONAL ML CLASSIFIER ACCURACY ON FER-2013 WITH EXPANDED SAMPLING (2,000 IMAGES PER CLASS)

Model	Test Accuracy
Random Forest (100 estimators, depth=15)	~42%
XGBoost (100 estimators, lr=0.1)	~40%
KNN ($k = 5$)	~28%
Decision Tree (depth=15)	~22%

Table 9 directly compares the simple 15-epoch baseline CNN against the enhanced CNN. Table 10 presents key training milestones for the enhanced model extracted directly from the training log. Figure 6 presents the confusion matrix for the simple baseline CNN.

TABLE IX
BASELINE CNN VS ENHANCED CNN ON FER-2013

Model	Epochs	Best Val Accuracy	Key Additions
Simple CNN (Baseline)	15	~50%	None
Enhanced Custom CNN	50	**61.9%**	BN, Augmentation, Class Weights, LR Scheduling

TABLE X
ENHANCED CUSTOM CNN VALIDATION ACCURACY

Epoch	Train Accuracy	Val Accuracy	Learning Rate
1	17.9%	23.4%	1.0×10^{-3}
8	43.0%	50.6%	1.0×10^{-3}
15	48.5%	55.8%	1.0×10^{-3}
22	52.7%	57.1%	5.0×10^{-4}
33	55.5%	59.1%	5.0×10^{-4}
38	57.1%	60.3%	2.5×10^{-4}
46	58.3%	**61.9%**	2.5×10^{-4}
50	58.6%	61.8%	1.25×10^{-4}

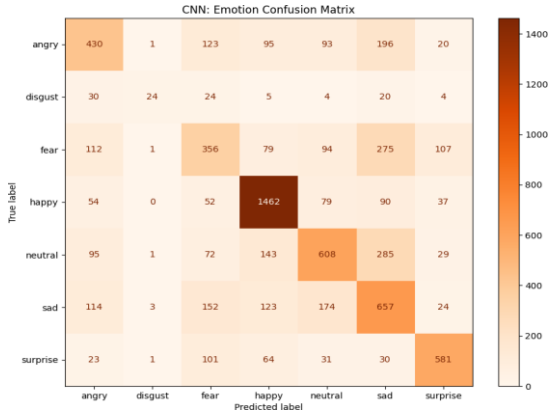

Fig. 6. Confusion matrix for simple baseline CNN on FER -2013 test set (~50% accuracy, 15 epochs)

3. Phase 3 — DAiSEE: 3 Frames Per Video

In Phase 3, the DAiSEE dataset was sampled at three frames per video clip, yielding 16,062 training, 4,278 validation, and 5,352 test images. Four traditional ML classifiers and three deep learning models were evaluated.

Traditional ML Baselines. Table 11 presents test set accuracy for four classifiers. Confusion matrices are shown in Figures 7–10.

TABLE XI
TRADITIONAL ML CLASSIFIER ACCURACY ON DAISEE TEST SET (3 FRAMES PER VIDEO)

Model	Test Accuracy
Random Forest	**89.91%**
K-Nearest Neighbours	89.72%
Logistic Regression	76.29%
Decision Tree	51.12%

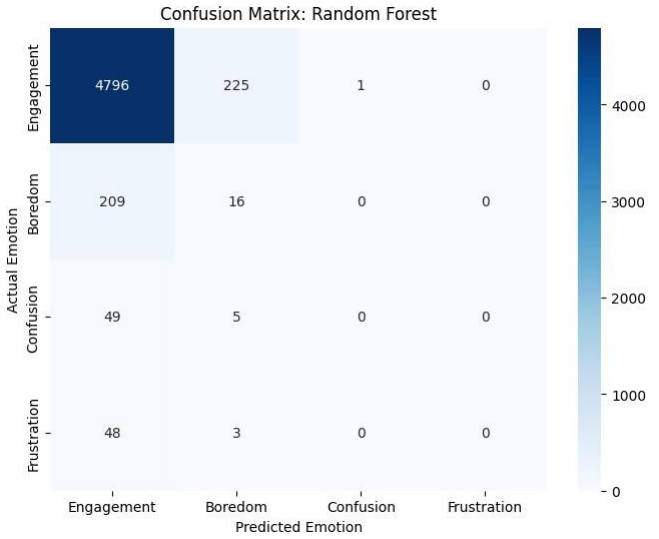

Fig. 7. Confusion matrix for Random Forest on DAiSEE test set (89.91%).

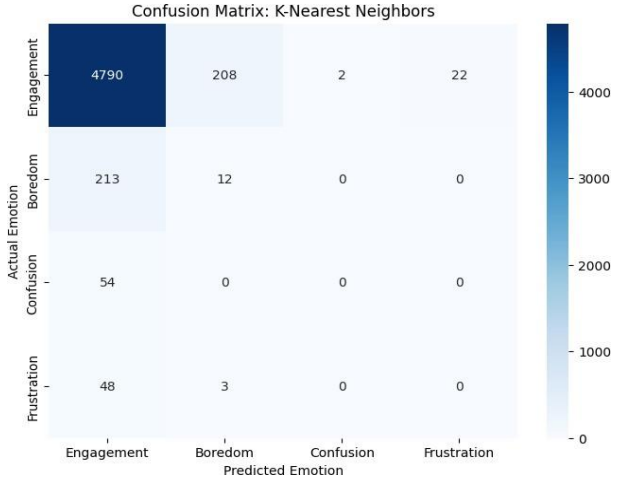

Fig. 8. Confusion matrix for K-Nearest Neighbours on DAiSEE test set (89.72%).

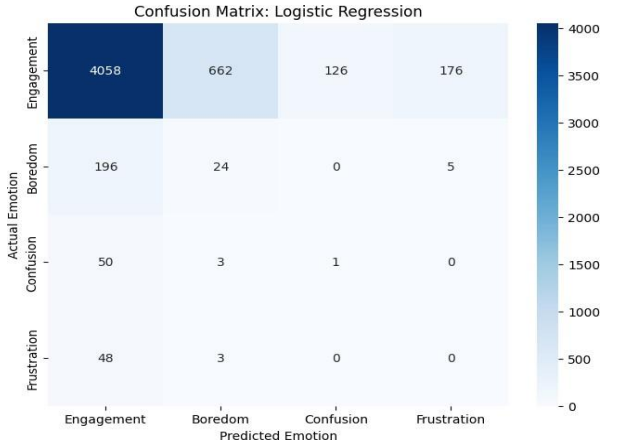

Fig. 9. Confusion matrix for Logistic Regression on DAiSEE test set (76.29%).

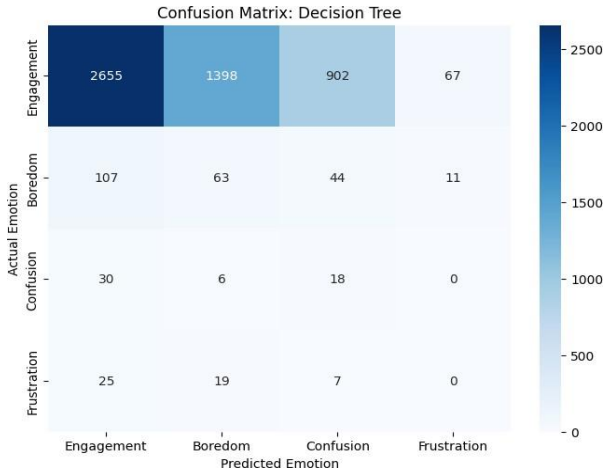

Fig. 10. Confusion matrix for Decision Tree on DAiSEE test set (51.12%).

MobileNetV2 Transfer Learning. Table 12 compares the two MobileNetV2 variants. Confusion matrices are presented in Figures 11 and 12.

TABLE XII
MOBILENETV2 VARIANT ACCURACY ON DAISEE TEST SET (3 FRAMES PER VIDEO)

Model	Max Epochs	Test Accuracy
MobileNetV2 (Standard class weights)	20	74.76%
MobileNetV2 (Aggressive class weights)	25	**80.34%**

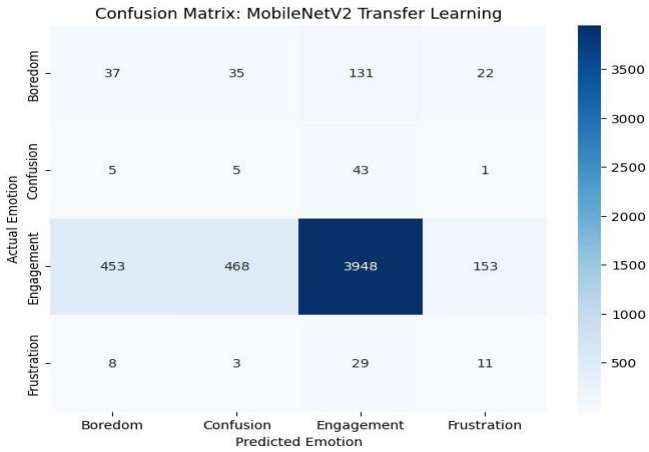

Fig. 11. Confusion matrix for standard MobileNetV2 on DAiSEE test set (74.76%).

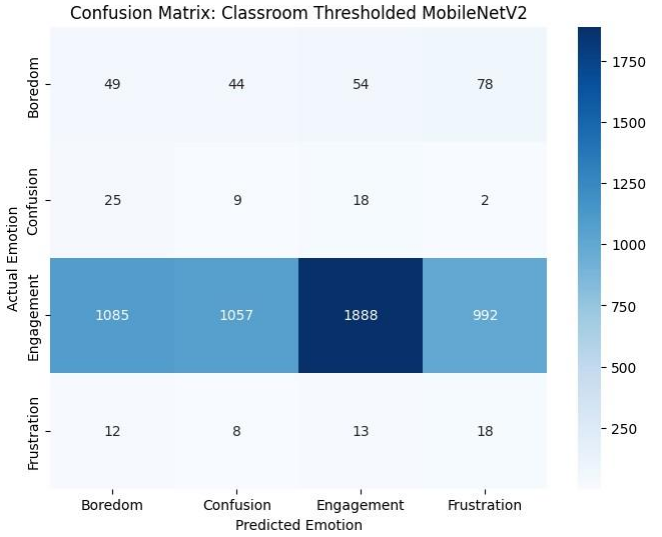

Fig. 12. Confusion matrix for Balanced MobileNetV2 on DAiSEE test set (80.34%)..

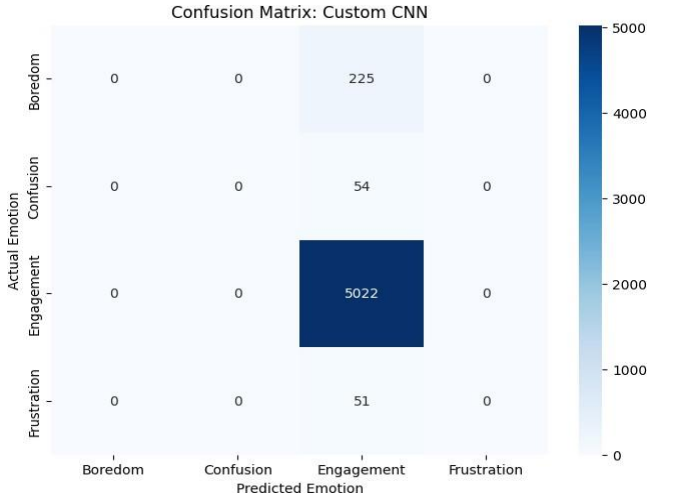

Fig. 13. Confusion matrix for Custom CNN on DAiSEE test set (93.83%).

The custom CNN trained end-to-end on DAiSEE frames achieved 93.83% test accuracy on the held-out test set. Figure 13 presents its confusion matrix.

4. Phase 4 DAiSEE: 3 Frames Per Second

Phase 4 investigated a denser temporal sampling strategy, extracting frames at 3 frames per second from DAiSEE video clips rather than 3 frames per video. This produced a substantially larger dataset of 258,261 total frames. Three traditional ML classifiers were evaluated after GridSearchCV hyperparameter tuning, followed by a CNN.

Traditional ML Baselines. Table 13 presents test accuracy, Macro F1-Score, and Boredom class recall for three classifiers. Confusion matrices are shown in Figures 14, 15, and 16.

TABLE XIII
TRADITIONAL ML CLASSIFIER RESULTS ON DAISEE (3 FRAMES PER SECOND, AFTER GRIDSEARCHCV TUNING)

Model	Accuracy	Macro F1	Boredom Recall	Best Parameters
Random Forest	71.2%	0.46	13.4%	depth=None, n=200
XGBoost	66.5%	0.47	24.4%	lr=0.1, depth=6
Linear SVM	59.6%	0.46	**70.6%**	C=1

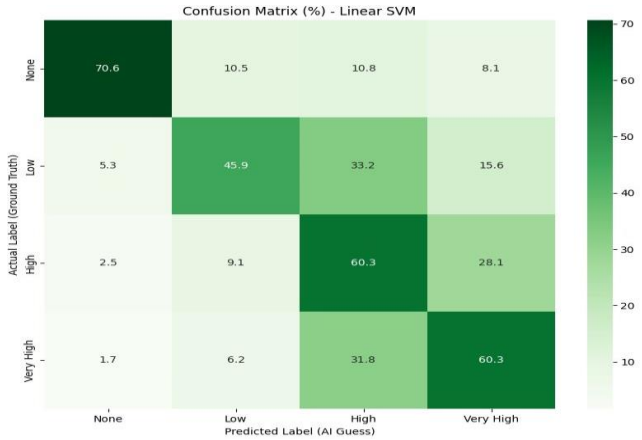

Fig. 14. Confusion matrix for Linear SVM on DAiSEE (3 frames per second, 59.6% accuracy, Boredom Recall 70.6%).

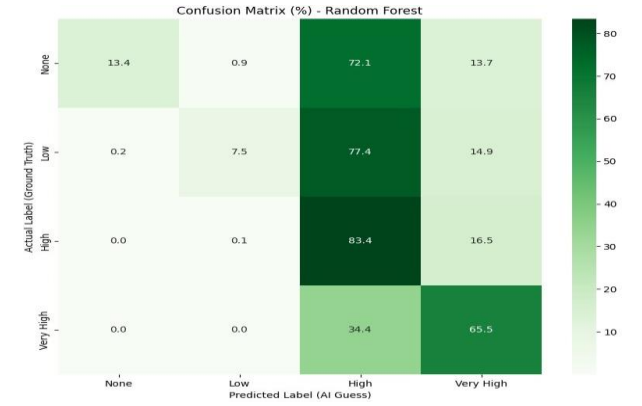

Fig. 15. Confusion matrix for Random Forest on DAiSEE (3 frames per second, 71.2% accuracy, Boredom Recall 13.4%).

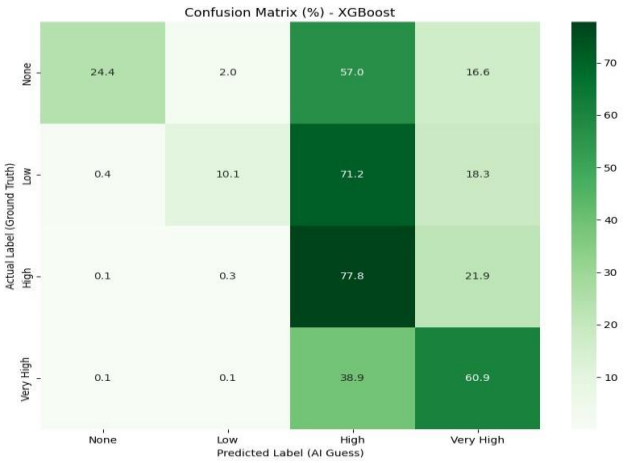

Fig. 16. Confusion matrix for XGBoost on DAiSEE (3 frames per second, 66.5% ac- curacy, Boredom Recall 24.4%).

CNN on 3 Frames Per Second. A CNN was evaluated under the denser sampling scheme on 258,261 total frames. Table 14 presents the per-class metrics derived from its confusion matrix, and Figure 17 shows the confusion matrix.

TABLE XIV
CNN PER-CLASS METRICS ON DAISEE (3 FRAMES PER SECOND), OVERALL ACCU- RACY: 48.62%, MACRO F1: 0.38

Class	Samples	Recall	Precision	F1-Score
Engagement	111,324	78.9%	50.7%	0.62
Boredom	80,582	32.2%	46.1%	0.38
Confusion	56,546	15.0%	48.6%	0.23
Frustration	9,809	33.6%	28.8%	0.31

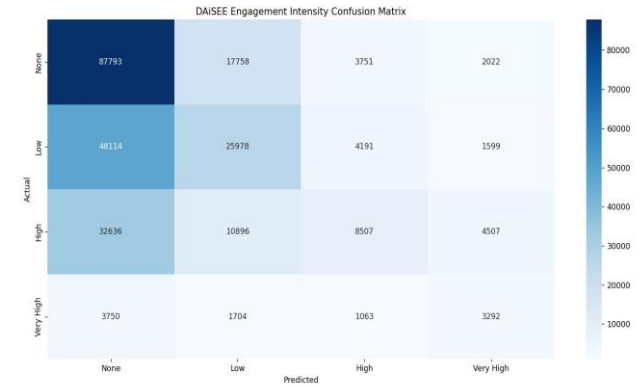

Fig. 17. Confusion matrix for CNN on DAiSEE (3 frames per second, 48.62% overall accuracy).

5. Overall Results Summary

Table 15 consolidates all model results across all four experimental phases in ascending order of accuracy within each phase.

TABLE XV
COMPLETE MODEL PERFORMANCE SUMMARY ACROSS ALL EXPERIMENTAL PHASES

Model	Dataset	Phase	Test Accuracy
Random Forest	FER-2013	Phase 1	~40%
XGBoost	FER-2013	Phase 1	~42%
SVM (Linear)	FER-2013	Phase 1	~44%
Simple CNN	FER-2013	Phase 2	~50%
Enhanced Custom CNN	FER-2013	Phase 2	**61.9%**
CNN	DAiSEE (3/sec)	Phase 4	48.62%
Linear SVM	DAiSEE (3/sec)	Phase 4	59.6%
XGBoost	DAiSEE (3/sec)	Phase 4	66.5%
Random Forest	DAiSEE (3/sec)	Phase 4	71.2%
Decision Tree	DAiSEE (3/vid)	Phase 3	51.12%
MobileNetV2 (Std)	DAiSEE (3/vid)	Phase 3	74.76%
Logistic Regression	DAiSEE (3/vid)	Phase 3	76.29%
MobileNetV2 (Bal)	DAiSEE (3/vid)	Phase 3	80.34%
KNN	DAiSEE (3/vid)	Phase 3	89.72%
Random Forest	DAiSEE (3/vid)	Phase 3	89.91%
Custom CNN	DAiSEE (3/vid)	Phase 3	**93.83%**

6. Comparison with Existing Literature

Table 16 compares the accuracy achieved by each model in this study against the performance reported for the same or equivalent architectures in the existing literature reviewed in Section 1.

TABLE XVI
COMPARISON OF MODEL ACCURACY AGAINST REPORTED LITERATURE VALUES

Model	Dataset	This Study	Literature	Source
SVM (Linear)	FER-2013	~44%	45.1%	Goodfellow et al. [3]
Random Forest	FER-2013	~40%	42.3%	Goodfellow et al. [3]
Custom CNN	FER-2013	61.9%	65.0%	Goodfellow et al. [3]
MobileNetV2	FER-2013	74.76%	~74–80%	Sandler et al. [9]
VGG-16/19	FER-2013	Not evaluated	>72%	Simonyan & Zisserman [11]
Mini-Xception	FER-2013	Not evaluated	~66%	Arriaga et al. [1]
Random Forest	DAiSEE	89.91%	—	No prior benchmark
MobileNetV2	DAiSEE	80.34%	—	No prior benchmark
Custom CNN	DAiSEE	**93.83%**	—	No prior benchmark
CNN + SMOTE	DAiSEE	Not evaluated	~72%	Santoni et al. [10]

Notably, no standardized benchmark accuracy exists for direct model-by-model comparison on the DAiSEE dataset under identical evaluation conditions, as prior work such as Santoni et al. applies different preprocessing pipelines and evaluation splits. This further motivates the reproducible benchmark established by this study.

CONCLUSION

This paper presented a systematic comparative study of machine learning and deep learning approaches to student engagement detection, evaluated across two benchmark datasets: FER-2013 and DAiSEE. On FER-2013, traditional ML classifiers trained on flattened pixel vectors achieved between 40% and 44% accuracy, confirming that spatial structure is essential for facial expression recognition. The enhanced custom CNN, incorporating batch normalization, data augmentation, and adaptive learning rate scheduling, achieved the best FER-2013 result of 61.9% validation accuracy over 50 epochs. On DAiSEE, traditional ML classifiers achieved surprisingly strong overall accuracy, with Random Forest reaching 89.91%, though confusion matrix analysis revealed this accuracy to be largely attributable to majority-class dominance. MobileNetV2 transfer learning achieved 74.76% in its standard configuration and 80.34% with aggressive class-weighting, demonstrating that domain gap limits the effectiveness of ImageNet pre-trained features for engagement detection. The custom CNN trained end-to-end on DAiSEE frames achieved the highest result across all experiments: 93.83% test accuracy, confirming that domain-specific architectures outperform generic transfer learning when sufficient task-specific data is available. All experiments were conducted on a standard consumer CPU without GPU acceleration, validating the edge-computing feasibility of the proposed approach for real classroom deployment under privacy-first constraints.

Three directions for future work are identified. First, temporal modelling via LSTM or 3D CNN architectures should be explored to leverage the sequential nature of DAiSEE video data rather than treating each frame independently. Second, a combined SMOTE and class-weighting strategy should be evaluated to address minority- class bias more effectively than either approach achieves in isolation. Third, the AffectNet dataset should be incorporated as a pre-training source to enrich the feature representations learned by the model before fine-tuning on the engagement-specific DAiSEE labels. The broader implication of this work is that evaluation frame-works for student engagement detection must move beyond aggregate accuracy and adopt per-class recall as the primary metric.

REFERENCES

Arriaga, O., Valdenegro-Toro, M., & Plöger, P. (2017). Real-time convolutional neural networks for emotion and gender classification (arXiv:1710.07557). arXiv.

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., & Zhao, W. (2015). Automatic Detection of Learning-Centered Affective States in the Wild.Proceedings of the 2015 International Conference on Intelligent User Interfaces (IUI).

Bosch, N., Muldner, Y., & D'Mello, S. (2015). Detecting student engagement and disengagement during learning with technology. Proceedings of the 8th International Conference on Educational Data Mining.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierberg, W., Tang, Y., Thaler, D., Lee, D. H., et al. (2013). Challenges in representation learning: A report on three machine learning contests. Neural Netw..

Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). DAiSEE: Towards user engagement recognition in the wild (arXiv:1609.01885). arXiv.

Hasani, B., & Mahoor, M. H. (2017). Facial expression recognition using enhanced deep 3D convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

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

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18–31.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M. L., Chen, S. C., & Iyengar, S. S. (2018). A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys, 51(5), 1–36.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks (arXiv:1801.04381). arXiv.

Santoni, M., Basaruddin, T., & Junus, K. (2023). Convolutional neural network model–based student engagement detection in imbalanced DAiSEE dataset. International Journal of Advanced Computer Science and Applications, 14(3).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition (arXiv:1409.1556). arXiv.

Vairagi, R. K., & Shaktawat, P. S. (2024). Recognition of student emotions through facial expressions in classrooms using deep learning techniques. Journal of Electrical Systems, 20(1).

A. A. Vichare and S. L. Varma. An improved conversation emotion detection using hybrid F-NN classifier. Indonesian Journal of Electrical Engineering and Computer Science, vol. 40, no. 1, pp. 490–498, Oct. 2025.

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.

Zhao, Z., Liu, Q., & Zhou, F. (2021). EfficientFace: Robust lightweight facial expression recognition network. Proceedings of the AAAI Conference on Artificial Intelligence, 35.