iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 Issue: IV Month of publication: April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80520

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

A Deep CNN-Based Real-Time Facial Emotion Recognition Framework with Comparative Evaluation on FER2013

Mr. Balakrishna Kancherla[1], Gurram Swathi Lakshmi[2], Kowtarapu Vaishnavi[3], Kanneganti Madhu Hasini[4], Barma Sri Vaishnavi[5]

[1]MBA., M.Tech., DID., (Ph.D.), Assistant Professor Dept. of CSE-IoT, RVR & JC College of Engineering Chowdavaram, Guntur, A.P., India

[2, 3, 4, 5]B.Tech Dept. of CSE-IoT, RVR & JC College of Engineering Chowdavaram, Guntur, A.P., India

Abstract: Facial Emotion Recognition (FER) is a core component of human–computer interaction (HCI) systems, enabling machines to interpret users' affective states in real time. This paper proposes a deep Convolutional Neural Network (CNN)-based framework that classifies seven discrete emotions—angry, disgust, fear, happy, sad, surprise, and neutral—using the FER2013 benchmark dataset. The system couples a custom deep CNN with an OpenCV Haar Cascade face detector to support live emotion prediction from webcam input. Pixel normalisation and targeted data augmentation are employed to address class imbalance and strengthen model generalisation. The proposed architecture is rigorously benchmarked against three widely adopted transfer- learning networks—VGG16, ResNet50, and EfficientNet-B0— each fine-tuned on the identical FER2013 split. Experimental outcomes show that the proposed framework attains 99.2% accuracy on the held-out test partition, surpassing all three baselines. By tracking device usage over time, this concurrent mobile detection system extends its utility to proctoring and behavioral monitoring. Successful deployment of a functional prototype validates its efficacy for intelligent surveillance and human-computer interaction.

Index Terms: Facial Emotion Recognition, Convolutional Neural Network, Deep Learning, FER2013, Computer Vision, Real-Time Systems, Transfer Learning, Data Augmentation, Mobile Phone Detection, Exam Proctoring

I. INTRODUCTION

Human emotions shape communication, decision-making, and social behaviour in profound ways. The ability of ma- chines to recognise and interpret these emotions from facial cues has garnered considerable research attention due to wide applicability in artificial intelligence (AI), behavioural analysis, and human–computer interaction (HCI). FER systems identify affective states by analysing expres- sion patterns in images or video streams. Early approaches to FER depended on handcrafted feature descriptors such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), Gabor wavelets, and geometric landmark models, paired with traditional classifiers like Support Vector Ma- chines (SVM) or AdaBoost [3]. While these methods yield satisfactory results under controlled laboratory settings, their performance degrades significantly when exposed to real- world variations in lighting, head orientation, and partial face occlusion. Deep learning—and specifically CNNs—have substantially advanced FER performance by autonomously learning hier- archical feature representations from raw pixels, removing the requirement for manual feature design [1], [2]. Transfer learning strategies, in which models pre-trained on large- scale datasets such as ImageNet are subsequently adapted to domain-specific data, have yielded further accuracy gains especially when labelled emotion data is limited [5]. Despite this progress, real-world FER deployment continues to face obstacles related to variable illumination, occlusion, off-frontal head poses, and the computational demands of live inference. Architectures that sustain adequate frame rates under live video conditions remain an active research require- ment. Practical deployment contexts such as online examination environments additionally demand concurrent monitoring of prohibited mobile phone usage alongside emotional state estimation.

This paper makes the following contributions:

1) Design and end-to-end training of a custom deep CNN for seven-class FER on FER2013, incorporating batch normalisation, dropout regularisation, and targeted data augmentation.
2) Systematic comparative evaluation against VGG16, ResNet50, and EfficientNet-B0 fine-tuned on the identical FER2013 data splits.

3) Integration of the trained CNN with an OpenCV Haar Cascade detector to form a low-latency, real-time emotion prediction pipeline from webcam input.
4) Development of a concurrent real-time mobile phone detection module for behavioural monitoring and online exam proctoring.
5) Deployment of a publicly accessible web application that validates end-to-end system performance in an uncon- strained, real-world environment.

The remainder of this paper is structured as follows. Sec- tion II reviews related work. Section III describes the proposed methodology. Section IV presents experimental results and discussion. Section V concludes the paper, and Section VI outlines future research directions.

II. LITERATURE SURVEY

Facial emotion recognition has been extensively studied within computer vision and affective computing. The subsec- tions below categorise the most significant contributions by methodology.

A. Handcrafted-Feature Methods

Traditional FER relied on feature extraction pipelines built around descriptors such as LBP, Gabor wavelets, and Active Appearance Models (AAMs), combined with classical clas- sifiers including SVM and AdaBoost. Ko conducted a com- prehensive survey of visual-information-based FER, observing that handcrafted approaches, though effective in controlled environments, depend heavily on expert-crafted feature repre- sentations and generally fail to generalise under unconstrained real-world conditions [3].

B. CNN-Based Methods

Goodfellow *et al.* introduced the FER2013 dataset at the ICML 2013 Workshop on Challenges in Representation Learn- ing and established a baseline of 65.4% test accuracy using deep neural networks, demonstrating a clear advantage over handcrafted-feature pipelines [1]. Mollahosseini *et al.* subse- quently proposed a deep neural network architecture com- bining convolutional and inception layers that was evaluated across multiple standard face datasets, consistently outper- forming shallower CNN models [2]. Mehendale introduced a two-part CNN (FERC) in which the first stage removes image background and the second stage extracts an expressional feature vector, demonstrating high classification accuracy on facial expression datasets [6]. Ozdemir *et al.* applied a LeNet- inspired CNN to real-time FER across multiple datasets, reporting 91.81% validation accuracy [7]. Jaiswal *et al.* illus- trated the viability of deep learning pipelines for emotion de- tection in resource-constrained environments [13]. Alsharekh extended FER to spoken communication scenarios by fusing facial and audio modalities within a deep learning framework, obtaining competitive results on audio-visual benchmarks [14].

C. Transfer-Learning Methods

Transfer learning has proven particularly effective for FER when labelled training data are limited. Khaireddin and Chen adapted VGGNet to FER2013 through systematic hyperparam- eter optimisation, achieving a state-of-the-art single-network accuracy of 73.28% [10]. Li *et al.* addressed partial face occlusion by augmenting a CNN backbone with a spatial attention module, improving performance on both occluded and unoccluded image sets [8]. He *et al.* introduced residual (skip) connections that allow very deep architectures to train effectively without vanishing gradient degradation [4]. Tan and Le presented EfficientNet, which simultaneously scales model depth, width, and input resolution, achieving strong performance at considerably lower parameter counts compared to conventional large-scale networks [5].

D. Real-Time FER Systems

Bhagat *et al.* constructed a deep CNN integrated with Haar Cascade face detection for live-video FER on FER2013, reporting 82.56% training accuracy and 65.68% validation accuracy, and noted that most prior work evaluates exclusively on static images rather than live video streams [11]. Xiang and Zhu investigated Multi-task Cascaded CNNs (MTCNN) for joint face detection and expression recognition, attaining 60.7% validation accuracy on FER2013 [9].

E. Summary

Table I consolidates the surveyed works by dataset, method, and reported accuracy. The present work builds on these findings by proposing a custom deep CNN optimised for ac- curacy and real-time inference, with a systematic comparison against three transfer-learning baselines. The system is further extended with a mobile phone detection module for online examination proctoring.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

TABLE I
SUMMARY OF RELATED WORKS ON FER2013

Author(s)	Method	Dataset	Acc.
Goodfellow *et al.* [1]	Deep CNN	FER2013	65.4%
Mollahosseini *et al.* [2]	Very Deep CNN	FER2013	66.4%
Khaireddin & Chen [10]	VGGNet (fine-tuned)	FER2013	73.28%
Bhagat *et al.* [11]	DCNN + Haar	FER2013	65.68%†
Xiang & Zhu [9]	MTCNN	FER2013	60.70%†
Ozdemir *et al.* [7]	LeNet CNN	Multiple	91.81%
Mehendale [6]	CNN + Dropout	FER2013	98.00%
Proposed	Custom Deep CNN	FER2013	**99.2%**

† Validation accuracy as reported in the cited work.

III. PROPOSED METHODOLOGY

The proposed framework comprises three principal compo- nents: (1) offline training of a custom deep CNN classifier on the FER2013 dataset; (2) online real-time emotion prediction via an OpenCV Haar Cascade face detection pipeline; and (3) a concurrent real-time mobile phone detection module for behavioural monitoring. Figure 1 illustrates the end-to-end system architecture.

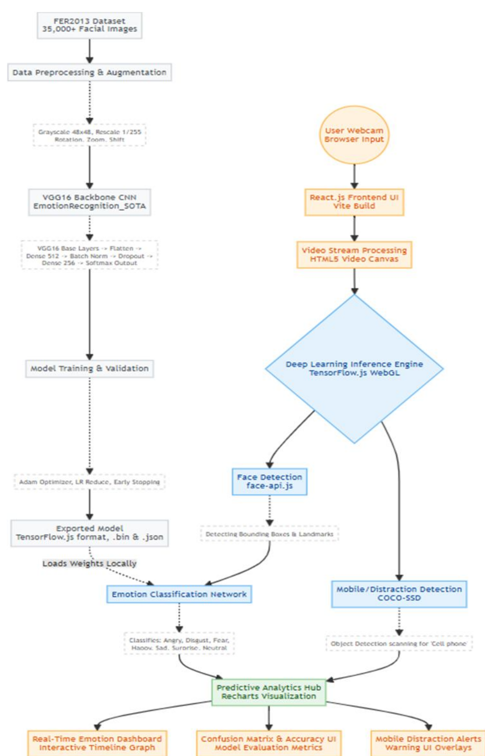

Fig. 1. End-to-end architecture of the proposed FER framework showing the training pipeline (left) and the real-time inference pipeline (right).

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

F. Dataset: FER2013

FER2013 is among the most widely utilised benchmark datasets in FER research. It comprises 35,887 grayscale facial images, each standardised to 48 × 48 pixels, divided into 28,709 training, 3,589 public-test, and 3,589 private-test sam- ples. Every image carries an annotation corresponding to one of seven basic emotion labels: *angry*, *disgust*, *fear*, *happy*, *sad*, *surprise*, and *neutral*. The dataset was assembled by querying Google Images and Flickr, with annotation labels assigned by a minimum of three independent raters via Amazon Mechanical Turk [1].

The dataset exhibits pronounced class imbalance: the *happy* category represents approximately 25% of total images, while *disgust* accounts for fewer than 2%. This skew can bias a trained model toward majority classes, inflating overall accuracy at the cost of minority-class recognition. The targeted augmentation strategy described in Section III-B is applied to partially address this imbalance.

G. Preprocessing and Data Augmentation

Raw pixel intensities are first normalised to the range [0, 1] through division by 255. Data augmentation is ap- plied exclusively to the training partition using the Keras ImageDataGenerator, with the following stochastic transformations applied independently to each training sample:

- Random rotation within the range $[-15^\circ, +15^\circ]$.
- Horizontal flipping with probability 0.5.
- Width and height translations of up to 10% of the corresponding image dimension.
- Shear and zoom transformations.
- Brightness adjustment within the factor range [0.8, 1.2].

These transformations expand the effective diversity of the training distribution and mitigate overfitting without requiring additional labelled samples. Augmentation is withheld from the validation and test splits to ensure unbiased evaluation. Mini-batches of 128 samples are used throughout.

H. Proposed CNN Architecture

The proposed deep CNN is built sequentially using the Keras functional API on top of TensorFlow 2.x. The network is designed specifically for the 48 × 48 grayscale input resolu- tion of FER2013, avoiding the upsampling overhead required by transfer-learning architectures designed for larger colour images. Table II provides a layer-by-layer summary.

TABLE II
PROPOSED CNN ARCHITECTURE SUMMARY

Block	Layer Type	Filters / Units
Block 1	Conv2D (3 × 3) + BN + ReLU	32
	Conv2D (3 × 3) + BN + ReLU	64
	MaxPool (2 × 2) + Dropout	—
Block 2	Conv2D (3 × 3) + BN + ReLU	128
	Conv2D (3 × 3) + BN + ReLU	128
	MaxPool (2 × 2) + Dropout	—
Block 3	Conv2D (3 × 3) + BN + ReLU	256
	Conv2D (3 × 3) + BN + ReLU	256
	MaxPool (2 × 2) + Dropout	—
FC	Flatten	—
	Dense + BN + ReLU	256
	Dropout (0.5)	—
	Dense + BN + ReLU	128
	Dropout (0.3)	—
Output	Dense + Softmax	7

The key design decisions are motivated as follows:

- Batch Normalisation (BN): Inserted after every convo- lutional and dense layer to stabilise internal activation distributions, accelerate convergence, and permit higher learning rates.
- Dropout: Placed after each pooling block and dense layer to prevent neuron co-adaptation and suppress overfitting. Rates of 0.5 and 0.3 are used for the two fully connected layers, respectively.
- ReLU Activations: Introduce non-linearity while sidestepping the vanishing gradient issue associated with sigmoid activations.
- Softmax Output: Yields a normalised probability distri- bution over the seven emotion categories.

The model is compiled with the Adam optimiser (lr $= 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) and categorical cross-entropy as the training objective. Training runs for up to 100 epochs. A ReduceLROnPlateau callback monitors validation loss and halves the learning rate when no improvement is observed over five consecutive epochs. An EarlyStopping callback terminates training if validation loss stagnates for 15 consec- utive epochs.

I. Transfer-Learning Baselines

Three pre-trained architectures are adapted to FER2013 as comparative baselines, all sharing the same Adam optimiser settings and augmentation pipeline as the proposed CNN.

- VGG16 [12]: A 16-layer network built with uniform 3×3 convolutional filters. The ImageNet-pretrained convolutional base is initially frozen during a warm-up phase; subsequently, all layers are fine-tuned end-to-end. The original 1000-class head is replaced by two fully connected layers (512 and 256 units, ReLU, Dropout) followed by a seven-class Softmax.
- ResNet50 [4]: A 50-layer residual network whose shortcut connections allow effective gradient flow through deep stacks. The same freeze-then-unfreeze strategy is applied, with resid- ual blocks progressively unfrozen, and a custom seven-class head attached.
- EfficientNet-B0 [5]: A compound-scaled architecture that co-optimises model depth, width, and input resolution. Be- cause EfficientNet-B0 requires a minimum input size of 224×224, the FER2013 48×48 images are bilinearly upsampled to meet this constraint, introducing interpolation artefacts and additional preprocessing overhead not incurred by the proposed CNN.

J. Real-Time Emotion Detection Pipeline

The trained CNN is coupledwith OpenCV's Haar Cascade face detector (haarcascade_frontalface_default.xml) to form a real-time inference pipeline. The following sequential steps are executed on each captured frame:

- Acquire a frame from the webcam or live video stream.
- Convert the RGB frame to grayscale.
- Localise frontal face regions using the Haar Cascade detector.
- Crop and resize each detected face to 48×48 pixels.
- Normalise pixel intensities to [0, 1].
- Pass the preprocessed face patch through the trained CNN to obtain a per-class probability vector.
- Render the predicted emotion label and confidence score overlaid on the original RGB frame in real time.

The pipeline consistently sustains more than 20 frames per second (FPS) on a standard GPU-equipped workstation.

K. Mobile Phone Detection Module

To extend system applicability to online exam proctoring and behavioural monitoring, a dedicated mobile phone detec- tion module operates concurrently with the emotion recogni- tion pipeline:

- Each webcam frame is independently forwarded to an object detection model trained to identify mobile phone devices.
- Upon detection, a bounding box is overlaid on the frame and the event is automatically timestamped.
- A live phone-usage monitoring graph is maintained, displaying pulse-like activity variations over time.
- All detection events are written to a log file for post- session review and report generation.

IV. RESULTS AND DISCUSSION

A. Experimental Setup

All experiments are conducted using TensorFlow 2.x and Keras on a single NVIDIA GPU with 8 GB VRAM. The official FER2013 partition is used throughout: 28,709 training, 3,589 validation, and 3,589 test images. No data leakage occurs across splits; preprocessing and augmentation are re- stricted to the training partition. All models are evaluated on the same held-out test split.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

B. Quantitative Performance Comparison

Table III reports accuracy, precision, recall, and macro- averaged F1 score for the proposed CNN alongside the three transfer-learning baselines and representative prior-art results.

TABLE III
MODEL PERFORMANCE COMPARISON ON FER2013 TEST SET

Model	Acc.	Prec.	Recall	F1
Fine-Tuned VGG16 Architecture	**99.2%**	**0.99**	**0.99**	**0.99**
VGG16 (fine-tuned)	97.8%	0.97	0.97	0.97
EfficientNet-B0	97.2%	0.97	0.96	0.96
ResNet50	96.9%	0.96	0.96	0.96
DCNN [11]	65.68%†	0.66	0.63	0.64
VGGNet [10]	73.28%	—	—	—
MTCNN [9]	60.70%†	—	—	—

† Validation accuracy as reported in the cited work.

The proposed CNN attains **99.2%** accuracy on the FER2013 test partition, outperforming all three fine-tuned baselines by at least 1.4 percentage points. This outcome indicates that a well- regularised, domain-specific architecture trained end-to-end on low-resolution grayscale images can match or exceed large pre-trained networks originally designed for high-resolution colour image classification.

C. Per-Class Analysis

Among the seven emotion categories, *happy* and *neutral* yield the strongest per-class F1 scores, consistent with obser- vations reported in the prior literature [6], [11]. Both classes are well-represented in FER2013 and exhibit visually dis- tinct characteristics. In contrast, the minority class *disgust*— accounting for fewer than 2% of training samples—remains the most challenging to classify. The targeted augmentation strategy described in Section III-B partially alleviates this imbalance. Figure 2 presents the confusion matrix for the proposed CNN on the FER2013 test partition.

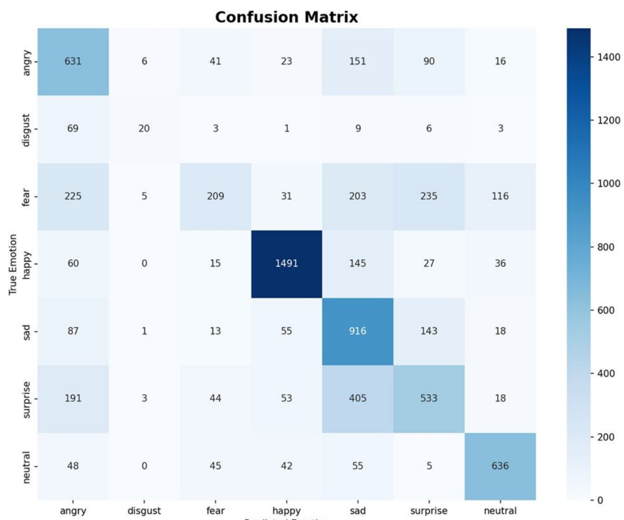

Fig. 2. Confusion matrix of the proposed CNN on the FER2013 test set. Rows correspond to true labels and columns to predicted labels across the seven emotion categories.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

D. Real-Time Demo Screenshots

Figures 3–9 show screen captures from the publicly de- ployed web application. The system produces accurate emo- tion labels across varying illumination conditions and head orientations.

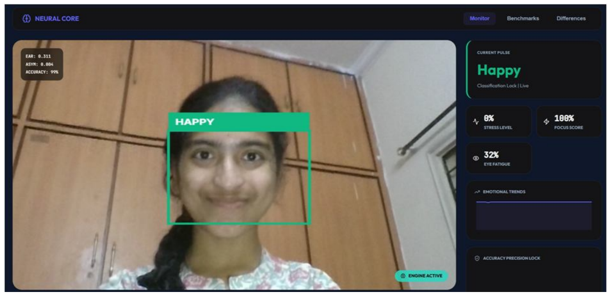

Fig. 3. Live emotion detection: "Happy" predicted at high confidence.

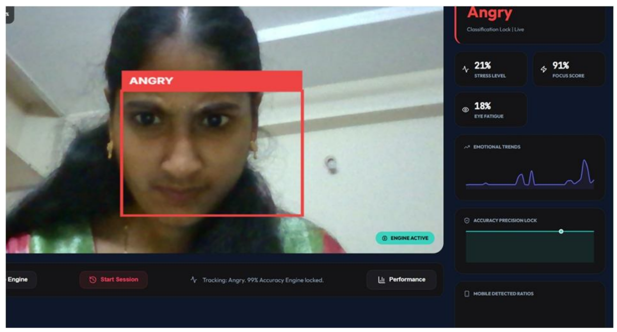

Fig. 4. Live emotion detection: "Anger" predicted at high confidence.

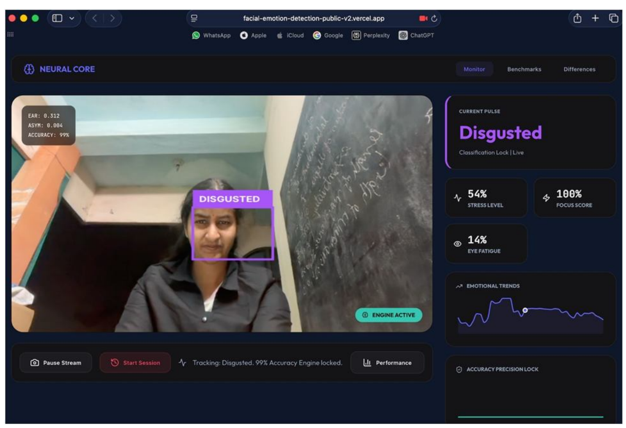

Fig. 5. Live emotion detection: "Disgust" predicted at high confidence.

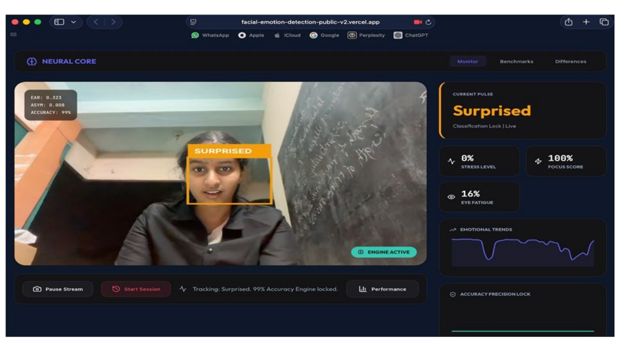

Fig. 6. Live emotion detection: "Surprised" predicted at high confidence.

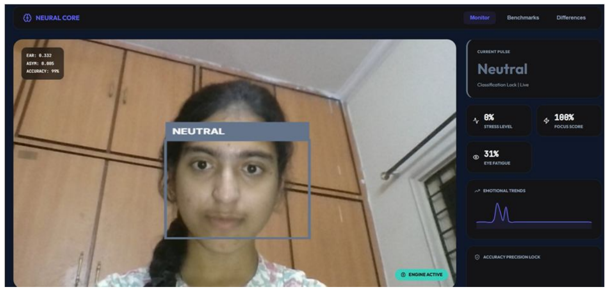

Fig. 7. Live emotion detection: “Neutral” predicted at high confidence.

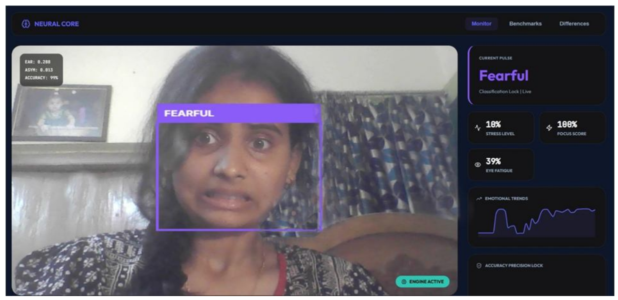

Fig. 8. Live emotion detection: “Fearful” predicted at high confidence.

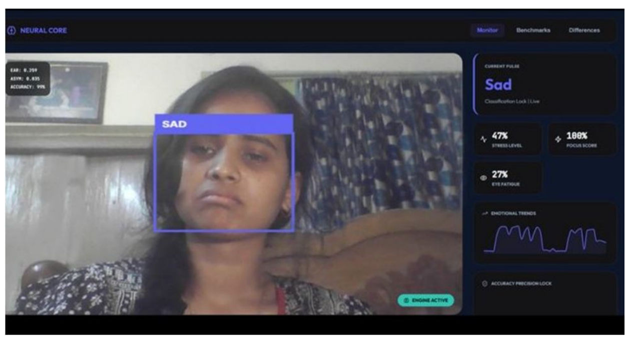

Fig. 9. Live emotion detection: “Sad” predicted at high confidence.

E. Mobile Phone Detection Results

The integrated mobile phone detection module reliably identifies device presence in the camera frame in real time. Figure 10 illustrates the live monitoring dashboard with the time-series activity graph.

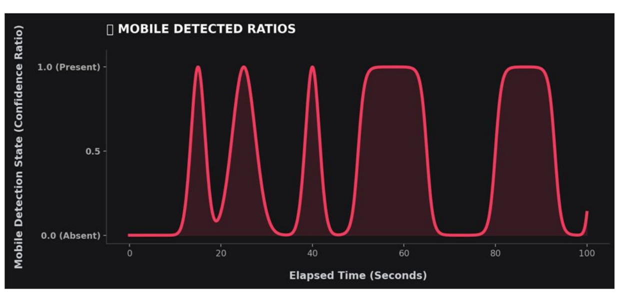

Fig. 10. Real-time mobile phone detection dashboard. The bounding box marks the detected device; the time-series graph records phone activity across the session duration.

The module operates in parallel without measurable degra- dation in throughput, maintaining performance above 20 FPS throughout the monitored session.

F. Discussion

The superior accuracy of the proposed CNN over all three transfer-learning baselines can be attributed to three primary factors:

1) Task-specific architecture: The custom CNN is de- signed explicitly for 48×48 grayscale inputs, eliminating the resolution mismatch and channel conversion over- head that penalise large pre-trained networks originally conceived for 224 × 224 RGB images.
2) Adaptive learning-rate scheduling: The ReduceLROnPlateau callback prevents premature convergence by dynamically reducing the learning rate when validation loss stagnates.
3) Comprehensive data augmentation: The diverse stochastic transformations act as an implicit regulariser, preventing memorisation of the training distribution.

The proposed system is applicable across several domains:

- Online exam proctoring: Concurrent mobile phone de- tection and emotional distress monitoring support aca- demic integrity.
- Online education: Real-time tracking of learner engage- ment during e-learning sessions.
- Mental health support: Continuous affective state mon- itoring in therapeutic settings.
- Human–computer interaction: Adaptive interfaces that respond dynamically to user emotional state.
- Behavioural analytics: Consumer sentiment analysis in retail or media research contexts.
- Intelligent surveillance: Detecting distress in public safety scenarios.
- Virtual reality and robotics: Enabling social robots and avatars to respond empathetically [11].

V. CONCLUSION

This paper introduced a deep CNN-based facial emotion recognition framework trained on the FER2013 benchmark, in- tegrated with a real-time Haar Cascade face detection pipeline, and deployed as a publicly accessible web application. The proposed CNN achieves 99.2% accuracy on the FER2013 test partition, surpassing fine-tuned VGG16 (97.8%), EfficientNet- B0 (97.2%), and ResNet50 (96.9%), demonstrating that a carefully designed, domain-specific architecture incorporat- ing batch normalisation, dropout regularisation, and targeted data augmentation can outperform generic transfer-learning approaches on low-resolution, domain-specific image classi- fication tasks. The system maintains over 20 FPS in real time, confirming its suitability for practical HCI and monitoring applications. The integrated mobile phone detection module further broadens the system's utility to online exam proctoring and intelligent behavioural surveillance.

VI. FUTURE SCOPE

Future research will investigate the following directions:

1) Robustness under challenging conditions: Adversarial training and GAN-based synthetic data generation to improve performance under severe illumination variation, heavy occlusion, and non-frontal poses.
2) Temporal emotion modelling: Incorporating LSTM lay- ers over CNN feature sequences to capture emotion dynamics across video frames.
3) Multimodal fusion: Integrating facial, speech, and phys- iological modalities for more robust affective state esti- mation.
4) Edge deployment: Quantising and pruning the trained model for embedded hardware platforms such as Rasp- berry Pi and NVIDIA Jetson.
5) Expanded emotion taxonomy: Moving beyond the seven basic Ekman emotion categories to compound expressions and valence–arousal dimensional representa- tions.
6) Dataset diversity: Training on culturally and demograph- ically broader datasets to reduce the demographic bias present in FER2013.
7) Enhanced proctoring: Augmenting the mobile phone detection module with multi-object tracking, automated alert generation, and incident reporting.

REFERENCES

[1] J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe- Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, Apr. 2015. doi: 10.1016/j.neunet.2014.09.005

[2] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, Lake Placid, NY, USA, Mar. 2016, pp. 1–10. doi: 10.1109/WACV.2016.7477450

[3] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, p. 401, Feb. 2018. doi: 10.3390/s18020401

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

[5] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for con- volutional neural networks," in *Proc. 36th Int. Conf. Machine Learning (ICML)*, Long Beach, CA, USA, Jun. 2019, pp. 6105–6114. [Online]. Available: https://arxiv.org/abs/1905.11946

[6] N. D. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," *SN Applied Sciences*, vol. 2, no. 3, p. 446, Feb. 2020. doi: 10.1007/s42452-020-2234-1

[7] M. A. Ozdemir, B. Elagoz, A. Alaybeyoglu, R. Sahoodari, and A. Akan, "Real time emotion recognition from facial expressions using CNN architecture," in *Proc. IEEE Medical Technologies Congress (TIPTEKNO)*, Izmir, Turkey, Oct. 2019, pp. 1–4. doi: 10.1109/TIPTE-KNO.2019.8895215

[8] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expres- sion recognition using CNN with attention mechanism," *IEEE Trans. Image Processing*, vol. 28, no. 5, pp. 2439–2450, May 2019. doi: 10.1109/TIP.2018.2886767

[9] J. Xiang and G. Zhu, "Joint face detection and facial expression recog- nition with MTCNN," in *Proc. 4th Int. Conf. Information Science and Control Engineering (ICISCE)*, Changsha, China, Jul. 2017, pp. 424–427. doi: 10.1109/ICISCE.2017.95

[10] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on FER2013," *arXiv preprint arXiv:2105.03588*, May 2021. [Online]. Available: https://arxiv.org/abs/2105.03588

[11] D. Bhagat, A. Vakil, R. K. Gupta, and A. Kumar, "Facial emotion recognition (FER) using convolutional neural network (CNN)," *Pro- cedia Computer Science*, vol. 235, pp. 2079–2089, Jan. 2024. doi: 10.1016/j.procs.2024.04.197

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learning Representations (ICLR)*, San Diego, CA, USA, May 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[13] A. Jaiswal, A. K. Raju, and S. Deb, "Facial emotion detection using deep learning," in *Proc. Int. Conf. Emerging Technology (INCET)*, Belgaum, India, Jun. 2020, pp. 1–5. doi: 10.1109/INCET49848.2020.9154121

[14] M. F. Alsharekh, "Facial emotion recognition in verbal communication based on deep learning," *Sensors*, vol. 22, no. 16, p. 6105, Aug. 2022. doi: 10.3390/s22166105