A Real-Time Deep Learning Framework for Classroom Facial Expression Recognition: Performance Optimization and Model Evaluation

Shardha Nand
Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Kuala Lumpur, Malaysia
nand.shardha@s.unikl.edu.my

Siti Haryani Shaikh Ali
Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Kuala Lumpur, Malaysia
sharyani@unikl.edu.my

Shahrulniza Musa
Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Kuala Lumpur, Malaysia
shahrulniza@unikl.edu.my (corresponding author)

Mazliham Mohd Su'ud
Multimedia University, Persiaran Multimedia, Cyberjaya, Selangor, Malaysia
mazliham@mmu.edu.my

Received: 21 December 2025 | Revised: 9 January 2026, 20 January 2026, and 24 January 2026 | Accepted: 28 January 2026

 DOI: https://doi.org/10.48084/etasr.17088

ABSTRACT

Facial expressions indicate a person's affective state and can be a significant determinant of cognitive performance. This study proposes a Facial Expression Recognition System (FERS) to detect and analyze students' real-time emotions, thereby providing teachers with insights. The proposed system was trained and evaluated using eight pretrained models on the CK+ dataset. A comparative analysis indicates that the Xception model achieved the highest accuracy in emotion classification. To improve model performance, the grayscale images in the CK+ dataset were enhanced and used as input to an Xception-based Convolutional Neural Network (CNN), employing 3×3 Conv2D filters with ReLU activation and same-padded layer feature extraction. The model demonstrated excellent performance, achieving an accuracy of 99.34%, a precision of 90%, a recall of 87%, and an F1-score of 88%, confirming the system's reliability and efficiency, applying macro averaging. In conclusion, the proposed Xception-based CNN FERS can accurately recognize students' emotions, allowing teachers to monitor students' moods in the classroom.

Keywords-facial emotion recognition; pretrained models; deep learning; Xception; transfer learning

I. INTRODUCTION

Nowadays, Deep Learning (DL) plays a significant role in the integration of technology and education [1]. The human face is dynamic and expresses a wide range of emotional states, mental activity, and social messages through facial expressions [2]. The recognition of these expressions by the computer is called Facial Expression Recognition (FER) [3], and it plays a significant role in computer vision, affective computing, and Human-Computer Interaction (HCI). FER systems are designed to mimic human facial expression perception and understanding via a face detection pipeline, feature extraction, and expression classification [4].

The transformation of the classroom into a dynamic, technology-driven learning environment is facing significant challenges, including teachers' difficulty in understanding and interpreting real-time non-verbal cues of students [5]. Facial expressions are indicators of the affective state, which can often go unnoticed, particularly in large or virtual classes, resulting in missed pedagogical intervention opportunities [6]. These challenges can be addressed by the effective use of FER,

which can analyze and recognize various facial expressions using still images or videos [7].

Human emotions are assumed to be inborn and are not influenced by extraneous factors, such as culture, race, and education; that is why they can be interpreted in the purest emotional sense [8]. With the advancement in video surveillance, a facial picture or video can be effectively used to detect facial expressions [9, 16]. However, processing facial images and videos requires considerable computing resources [10]. Compared to the offline (conventional) classroom, in the online classroom, teachers cannot keep track of each student's activities or even their activity level during the class [11]. In a conventional classroom, instructors can observe facial expressions, body language, and students' reactions due to small classroom sizes [12]. These challenges can be addressed by modern FER systems coupled with a DL-transfer-driven system [13]. FER systems can process video streams and recognize basic emotions, thereby identifying educationally relevant states with high accuracy [14]. In addition, FER technology offers real-time analytics dashboards, providing anonymized, class-level comprehension and engagement data to enable prompt instructional changes [15]. With the popularity of video surveillance, using facial images or videos to recognize facial emotions has been proven to be more accurate [16].

However, existing literature lacks an optimized FER system that can integrate transfer learning, effective preprocessing, and custom CNN refinement to ensure credible performance in real-time classrooms [17]. Moreover, research mainly focuses on individual-model structures or curated data, with limited comparative analysis on multi-pretrained DL models in a realistic learning environment [18]. FER in education remains experimental, as facial features are complicated and algorithms are not yet developed to the level of practical use. The performance of models examining students' expressions in classrooms is still not at the desired level. To address these challenges, the present study proposes a real-time Facial Expression Recognition System (FERS) that can accurately identify students' emotional state in the classroom. The study also compares the performance of different pretrained DL models on the CK+ dataset and identifies the most effective FER architecture in a learning environment. In addition, the highest performing model is optimized by employing image preprocessing techniques and custom CNN layers to boost its capability in accuracy, precision, recall, and overall robustness for real-time classroom deployment.

II. METHODOLOGY

A. Data Collection

The current study employed eight CNN models pretrained on the CK+ dataset [19] for facial emotion recognition. The dataset contains 981 images belonging to seven emotion categories collected using a surveillance camera. The class imbalance in the dataset was addressed by performing data augmentation to the total of 1,670 images. The augmented dataset was further divided into training and validation sets using a 70:30 split, with each emotion class being proportionately represented in each subset. Authors in [20] employed the CK+ dataset to identify various facial expressions with excellent performance in both geometric and DL methods.

Surveillance cameras are often used to collect data, reducing costs and restricting the disturbance of/caused by the students during classroom activities. The present study employed DL techniques for pre-processing the original data to extract the students' facial features. The eight pretrained models were first evaluated on the original dataset, followed by further testing on cropped Regions of Interest (ROIs) to focus on emotion-relevant facial features. Since CK+ is a relatively small dataset, transformations including rotation, scaling, and flipping were applied to 30% of the data to enhance generalization and reduce overfitting. This combination of transfer learning, ROI-based preprocessing, and augmentation enabled a more effective performance evaluation. Figure 1 illustrates the model development and FER process.

The CK+ dataset was used to perform various experiments, as depicted in Table I. The results indicate that the Xception model achieved the highest validation accuracy of 91.10%, with a very small train-validation gap of 8.24%. MobileNet achieved the lowest validation loss of 0.3723% with stable learning performance. In contrast, EfficientNetB4 and MobileNetV2 demonstrated a sign of overfitting, with a performance difference of about 44%. This low performance can be attributed to poor hyperparameter selection.

B. Xception Pretrained Model

Xception is a deep model, completely composed of depthwise separable convolutions, thus separating the spatial and channel correlations and increasing effectiveness. It has stacked separable modules and residual connections, which enhance representational power, while the computational cost is also low [21].

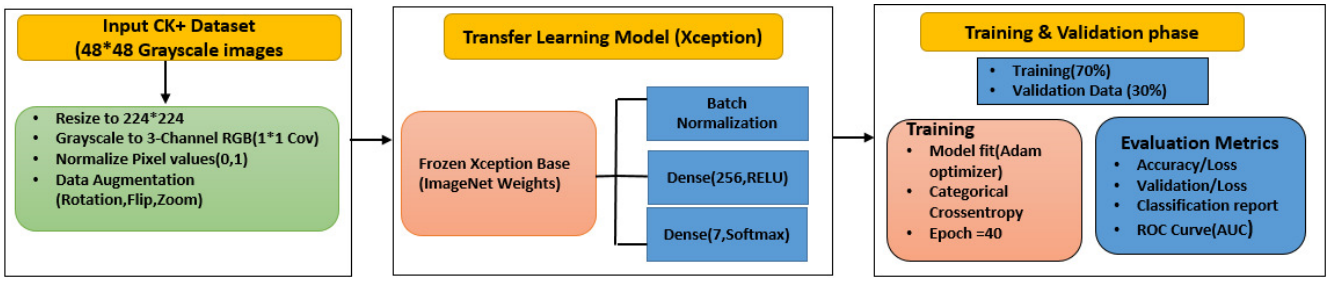

Fig. 1. Model development and evaluation framework for FER.

TABLE I. TRAIN-VALIDATION ACCURACY GAP AND LOSS FOR DIFFERENT CNN ARCHITECTURES

Model	Best training accuracy (%)	Best validation accuracy (%)	Training–validation gap (%)	Lowest validation loss (%)
AlexNet	32.80	20.54	12.26	1.9446
DenseNet121	95.79	69.86	25.93	0.8059
EfficientNetB4	62.41	18.15	44.26	2.0682
InceptionV3	90.57	81.16	9.41	0.5282
MobileNet	96.95	86.99	9.96	0.3723
MobileNetV2	81.86	34.93	46.93	1.7290
ResNet50	50.80	25.34	25.46	2.1219
VGG16	25.54	25.34	0.20	1.8153
Proposed Xception	99.34	91.10	8.24	0.3685

The model has standardized input processing, effective augmentation, and an Xception-based feature extractor to enhance robustness. A dense layer with dropout facilitates regularization, while softmax classification and the Adam optimizer help maintain a stable learning process and consistent performance.

C. Model Implementation

The model is provided with a single-channel input image and is run through a convolutional layer with three 3×3 filters applied with the same padding, followed by ReLU activation for non-linearity. The model transforms the single-channel input of original data into three channels that comprise the feature map needed by the Xception backbone. The transformed output is then fed into the pre-trained Xception model, which is utilized without its top layers to obtain high-level features of the input image:

$$F = X(C) = X\left(ReLU(\sum_{i=1}^{3} K_i * I + b_i)\right) \quad (1)$$

where I is the input tensor of shape $(H, W, 1)$, where H = IMG_SIZE [0] and W = IMG_SIZE [1]; K_i is the i^{th} convolution kernel of shape 3×3×1; and b_i is the bias associated with the i^{th} kernel. ReLU(·) is the activation function; $X(\cdot)$ is the transformation performed by the pretrained Xception, and F is the final feature representation produced by Xception.

D. Feature-Extraction

The input tensor x is subjected to global average pooling, which is a process of converting each feature map into a single value by averaging all the spatial locations. The resulting feature vector is then batch-normalized to stabilize and accelerate the training process. The normalized feature vector is sent through a dense layer containing 256 units and a ReLU activation to add non-linearity. To minimize overfitting, a dropout layer with a probability of 0.5 is used, as:

$$y = Dropout(ReLU(BN(GAP(x)) \cdot W + b), 0.5) \quad (2)$$

where x is the Input tensor, shape (H, W, C), $GAP(x)$ is the Global average pooling, $BN(v)$ is the batch normalization applied to vector v, and $Dropout(v, p)$ is the dropout function with probability p.

E. Model Compilation

To prepare the model for training, the model compile step provides the optimizer and loss function. In multi-class classification, categorical cross-entropy is used to calculate the difference between the model-predicted probabilities and the true one-hot labels. The Adam optimizer is set at a 1×10^{-4} learning rate, which is an effective way to update the model weights. To minimize loss, this configuration sets up the model for training using gradient-based learning, as:

$$\theta^{(t+1)} = Adam\left(\theta^{(t)}, \nabla_\theta \mathcal{L}(y, \hat{y})\right)$$

$$\hat{y} = f_\theta(x), \mathcal{L}(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (3)$$

where θ is the set of all trainable parameters (weights and biases) of the model. $\mathcal{L}(y, \hat{y})$ is the categorical cross-entropy loss function, where y is the true one-hot encoded label vector, and $\hat{y}$ is the predicted probability vector, - the learning rate $\alpha = 1 \times 10^{-4}$, $Adam(\alpha)$ is the Adam optimizer with learning rate, $\theta^{(t)}$ is the model parameters at training step t, x is the input model, $f_\theta(x)$ is the model's predicted probability vector, C is the number of classes, $\theta^{(t+1)}$ is updated according to the Adam optimizer using the gradient of the categorical cross-entropy loss. Table II presents the hyperparameters used for the Xception model.

TABLE II. XCEPTION HYPERPARAMETERS

Category	Hyperparameter
Input size	224 × 224 pixels
Batch size	32
Epochs	40
Train/validation split	70:30
Dense layer	256 neurons, ReLU activation
Dropout	0.5
Output layer	Softmax, neurons = number of classes
Loss function	Categorical cross-entropy
Optimizer	Adam, learning rate = 1×10^{-4}

III. RESULTS

The model exhibits a steady and efficient learning process with training and validation accuracies of 0.99 and 0.90, respectively, as illustrated in Figure 2. The loss during training approaches zero, and validation loss declines to 0.35-0.55, which is a strong indication of network optimization with first-order overfit. In general, the learning rate confirms that the network gains strong, discriminative features, necessary to make reliable facial-emotion recognition. The results indicate that the model attains promising overall performance with a validation accuracy of 0.85-0.90, with good recognition of happy and surprise, and moderate recognition of disgust and anger. Misclassifications occur mostly in subtle emotions, such as fear and sadness, but high true negative values suggest sound generalization, as displayed in Figure 3. Overall, the model is stable but can be improved to better reflect ambiguous expressions.

The Xception emotion classification model demonstrated excellent discriminatory abilities, evidenced by the ROC curves of the seven emotion classes shown in Figure 4. The loss during training approaches zero and validation loss declines 0.35-0.55, which can be seen as a strong indication of network optimization with a first-order overfit. In general, the curves confirm that the network gains strong, discriminative features necessary to make reliable facial-emotion recognition. The model achieved AUC scores of 0.95 and 1.00, significantly higher than the 0.50 mark of random performance. The curves close to the upper-left corner indicate that there is an optimal trade-off between false positives and true positives, which demonstrates that the model is both highly accurate and has low error rates in recognizing emotions. Overall, these findings suggest that the Xception model provides a promising visual emotion recognition solution.

Table III presents the classification performance of Xception for various emotions. As observed, the Xception model outperformed eight models (Table I), with an overall accuracy of 0.90 on 292 samples. The model achieved high F1-scores for the 'happy' (0.90), 'surprise' (0.97), and 'disgust' (0.94) classes, indicating a strong balance between precision and recall for these categories. Lower recall values for fear and sadness indicate that the model faced difficulty in identifying these emotions. In addition, the model achieved a macro F1-score of 0.88 and a weighted F1-score of 0.90, indicating a reliable and well-generalized emotion classification when class imbalance is present.

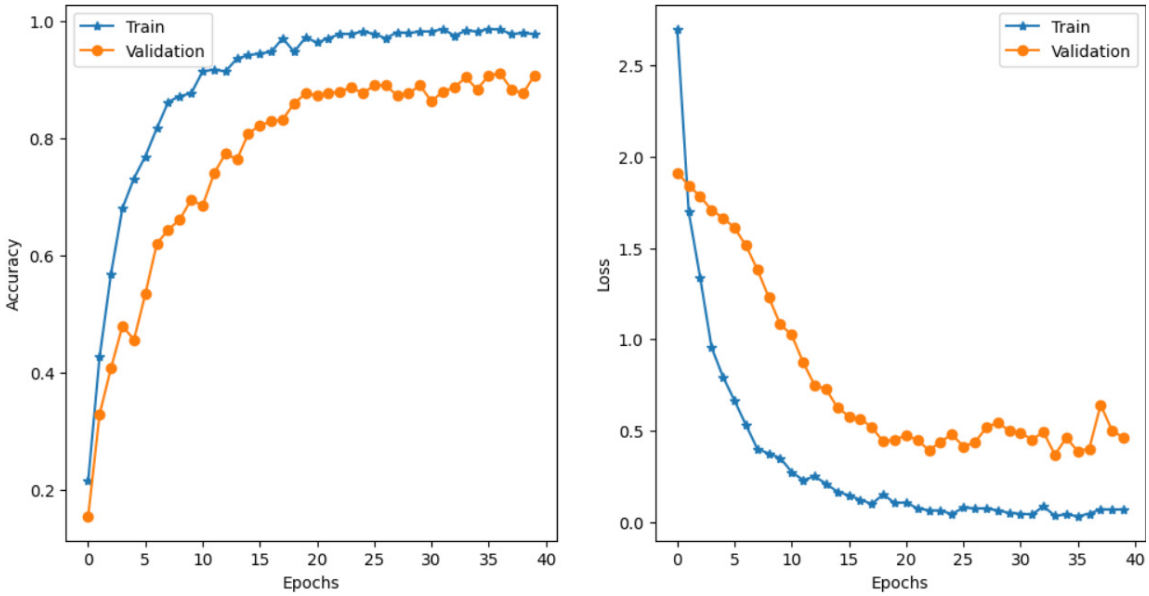

Fig. 2. Performance of the Xception model during training and validation.

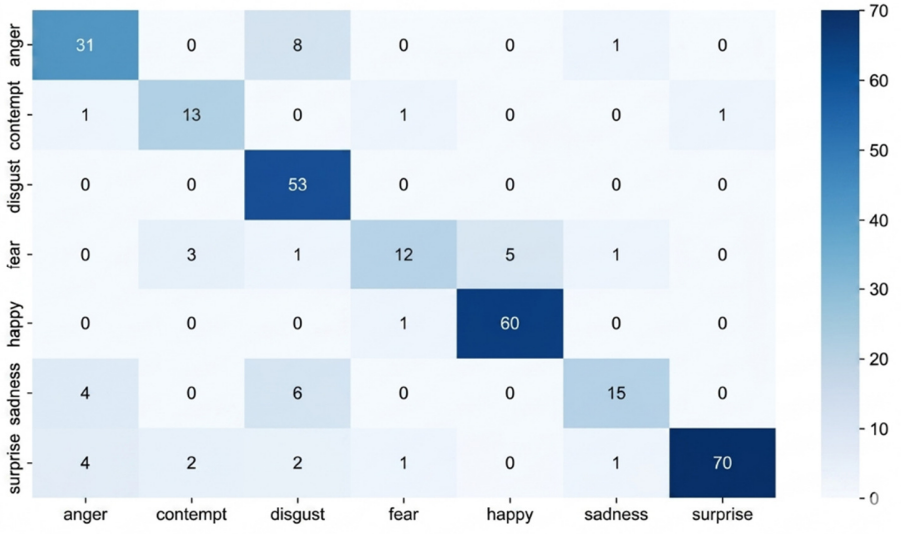

Fig. 3. Confusion matrix of the Xception model.

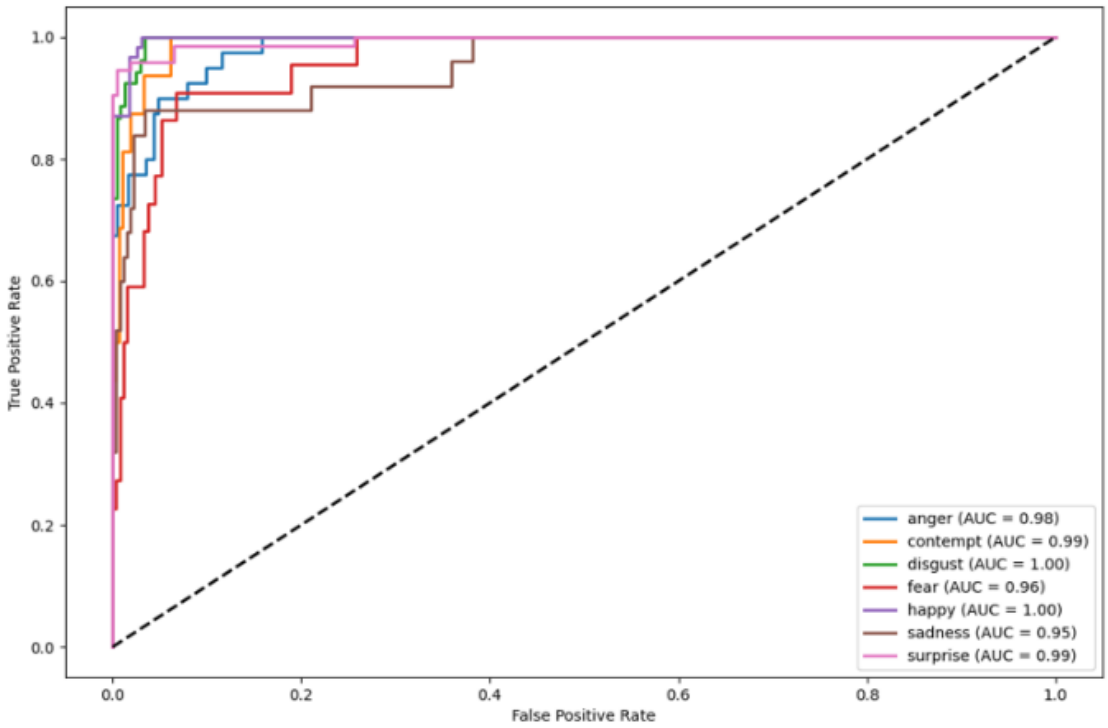

Fig. 4. ROC for multi-class emotion classification.

TABLE III. CLASSIFICATION PERFORMANCE OF THE XCEPTION MODEL

Class	Precision	Recall	F1-Score	Support
Anger	0.94	0.82	0.88	40
Contempt	0.84	1.00	0.91	16
Disgust	0.91	0.96	0.94	53
Fear	0.88	0.68	0.77	22
Happy	0.83	0.97	0.90	62
Sadness	0.89	0.68	0.77	25
Surprise	0.97	0.97	0.97	74
Accuracy	—	—	0.90	292
Macro Avg	0.90	0.87	0.88	292
Weighted Avg	0.91	0.90	0.90	292

IV. CONCLUSION

This study proposed a Facial Expression Recognition System (FERS) for real-time identification of the facial emotions of students in classrooms. The study employed Deep Learning (DL) to reliably classify seven emotions of students in order to provide academic support and early intervention. The results indicate that the Xception model outperformed the other eight pretrained models with a training accuracy of 99.34%, and excellent class-wise precision and recalls.

The novelty of this work lies in the systematic assessment and optimization of various trained models for educational emotion recognition. The study also provides an application-oriented design of practical use in classrooms. Despite the favorable results, the study has some limitations, including the smaller sample size of the CK+ dataset. For better generalization and stability, future studies should use bigger datasets such as FER-2013. In addition, the employed dataset contains surveillance camera images, which may not reflect real-world conditions. Future studies should collect facial expressions in an actual classroom.

COMPETING INTERESTS

The authors declare no competing interests.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to Universiti Kuala Lumpur Malaysia (UNIKL) for providing funding for the publication of this study.

DATA AVAILABILITY

The CK+ dataset used in the study is collected from [19].

AI USE AND DECLARATION OF GENERATIVE AI USE

The authors used generative AI tools, such as ChatGPT and Gemini AI, for language editing and improving the readability of the manuscript. All scientific content, analyses, and conclusions were independently reviewed and verified by the authors.

REFERENCES

[1] W. Strielkowski, V. Grebennikova, A. Lisovskiy, G. Rakhimova, and T. Vasileva, "AI-Driven Adaptive Learning for Sustainable Educational Transformation," *Sustainable Development*, vol. 33, no. 2, pp. 1921–1947, Apr. 2025, https://doi.org/10.1002/sd.3221.

[2] A. A. Salah and A. El Ali, "Affective User Interfaces," in *Handbook of Human Computer Interaction*, J. Vanderdonckt, P. Palanque, and M. Winckler, Eds. Cham: Springer Nature Switzerland, 2025, pp. 1–32.

[3] M. Carrasco, C. González-Martín, S. Navajas-Torrente, and R. Dastres, "Level of Agreement between Emotions Generated by Artificial

Intelligence and Human Evaluation: A Methodological Proposal," *Electronics*, vol. 13, no. 20, Oct. 2024, Art. no. 4014, https://doi.org/10.3390/electronics13204014.

[4] M. J. A. Dujaili, "Survey on Facial Expressions Recognition: Databases, Features and Classification Schemes," *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 7457–7478, Jan. 2024, https://doi.org/10.1007/s11042-023-15139-w.

[5] S. D. Qasim, *Beyond the Classroom: Emerging Technologies to Enhance Learning*. Book Bazooka Publication, 2024.

[6] A. McIntosh, "Teachers' Perception of Student Engagement Using Face-to-Face and Remote Instructional Methods," PhD Dissertation, St. John's University, New York City, NY, USA, 2025.

[7] M. Kaur and M. Kumar, "Facial Emotion recognition: A comprehensive review," *Expert Systems*, vol. 41, no. 10, Oct. 2024, Art. no. e13670, https://doi.org/10.1111/exsy.13670.

[8] K. Kusano, J. L. Napier, and J. Jost, "The Mismeasure of Culture: When Measurement Invariance Requirements Hinder Social-Psychological Research." PsyArXiv, Mar. 28, 2024, https://doi.org/10.31234/osf.io/9qe2k.

[9] H. L. Gururaj, B. C. Soundarya, S. Priya, J. Shreyas, and F. Flammini, "A Comprehensive Review of Face Recognition Techniques, Trends, and Challenges," *IEEE Access*, vol. 12, pp. 107903–107926, 2024, https://doi.org/10.1109/ACCESS.2024.3424933.

[10] A. Caruso *et al.*, "Adaptive 360° Video Streaming over a Federated 6G Network: Experimenting In-Network Computing for Enhanced User Experience," in *2024 20th International Conference on Network and Service Management*, Prague, Czech Republic, Oct. 2024, pp. 1–7, https://doi.org/10.23919/CNSM62983.2024.10814393.

[11] N. M. Alruwais and M. Zakariah, "Student Recognition and Activity Monitoring in E-Classes Using Deep Learning in Higher Education," *IEEE Access*, vol. 12, pp. 66110–66128, 2024, https://doi.org/10.1109/ACCESS.2024.3354981.

[12] K. Malta, C. Glickman, K. Hunter, and A. McBride, "Comparing the Impact of Online and In-Person Active Learning in Preclinical Medical Education," *BMC Medical Education*, vol. 25, no. 1, Mar. 2025, Art. no. 329, https://doi.org/10.1186/s12909-025-06846-z.

[13] B. Fang, X. Li, G. Han, and J. He, "Facial Expression Recognition in Educational Research from the Perspective of Machine Learning: A Systematic Review," *IEEE Access*, vol. 11, pp. 112060–112074, 2023, https://doi.org/10.1109/ACCESS.2023.3322454.

[14] M. B. Govind, A. S. Humaid, and G. Malu, "Revolutionizing Student Engagement: Real-Time Emotion Detection and Interest Identification in Live Video Streams," in *Fifth International Conference on Computing and Network Communications*, vol. 1221, S. M. Thampi, V. Chaudhary, A.-S. K. Pathan, K. Ching Li, and D. Krishnaswamy, Eds. Singapore: Springer Nature Singapore, 2025, pp. 409–422.

[15] C. R. Tirkey, "AI-Driven Real-Time Student Monitoring System for Enhancing Engagement, Learning, and Identifying Support Needs in K-12 Classrooms Through Visual Programming," M. S. Thesis, Kent State University, Kent, Ohio, 2025.

[16] A. Imran, R. Ahmed, M. M. Hasan, M. H. U. Ahmed, A. K. M. Azad, and S. A. Alyami, "FaceEngine: A Tracking-Based Framework for Real-Time Face Recognition in Video Surveillance System," *SN Computer Science*, vol. 5, no. 5, May 2024, Art. no. 609, https://doi.org/10.1007/s42979-024-02922-1.

[17] R. Grover and S. Bansal, "Optimizing Facial Expression Recognition in Challenging Environment: A Streamlined CNN with Pre-Processing Techniques," *Journal of The Institution of Engineers (India): Series B*, vol. 106, no. 4, pp. 1329–1348, Aug. 2025, https://doi.org/10.1007/s40031-024-01184-y.

[18] I. D. Mienye and T. G. Swart, "A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications," *Information*, vol. 15, no. 12, Nov. 2024, Art. no. 755, https://doi.org/10.3390/info15120755.

[19] S. Alok, "CK+ Dataset." Kaggle, Apr. 2025, [Online]. Available: https://www.kaggle.com/datasets/shuvoalok/ck-dataset.

[20] T. E. Köksal, "Deep Learning Based Real-Time Sequential Facial Expression Analysis Using Geometric Features," M. S. Thesis, İzmir Institute of Technology, Urla, Türkiye, 2023.

[21] J. A. Ballesteros, G. M. Ramírez, F. Moreira, A. Solano, and C. A. Pelaez, "Facial Emotion Recognition Through Artificial Intelligence," *Frontiers in Computer Science*, vol. 6, Jan. 2024, Art. no. 1359471, https://doi.org/10.3389/fcomp.2024.1359471.