A Deep Visual Approach to Student Engagement Analysis Using Affective and Behavioral Cues

Fatima Zahra Jobbid
Smart Systems Laboratory, ENSIAS, Mohammed V University in Rabat, Morocco
fatimazahra_jobbid@um5.ac.ma (corresponding author)

Aissam Berrahou
ENSIAS, Mohammed V University in Rabat, Morocco
aissam.berrahou@gmail.com

Hassan Berbia
Smart Systems Laboratory, ENSIAS, Mohammed V University in Rabat, Morocco
hassan.berbia@ensias.um5.ac.ma

Received: 4 August 2025 | Revised: 16 September 2025, 8 October 2025, 10 October 2025, and 26 October 2025 | Accepted: 29 October 2025

 DOI: https://doi.org/10.48084/etasr.13816

ABSTRACT

Assessing student engagement in educational environments is essential to support adaptive teaching strategies and enhance learning outcomes. This study presents a deep learning-based approach for automatically predicting student engagement, leveraging both behavioral and emotional cues. The proposed method integrates features derived from facial emotion recognition and head pose estimation, capturing a comprehensive representation of student affect and attention. Using the Student Engagement Dataset, a multi-layer neural network was trained to classify engagement states based on these multimodal inputs. The proposed framework achieves an accuracy of 88% on unseen validation data, demonstrating strong effectiveness in distinguishing between engaged and disengaged students. In addition, explainability analysis highlights the importance of neutral facial expressions and head orientation as key indicators of engagement, supporting the interpretability and practical relevance of the proposed approach for real-world educational environments.

Keywords-component; student engagement; emotion recognition; head pose estimation; deep learning; artificial intelligence in education

I. INTRODUCTION

Timely and objective assessment of student engagement is essential for improving learning outcomes in modern education. Engagement is a multidimensional construct that encompasses behavioral, emotional, and cognitive components and is strongly correlated with academic achievement, retention, and motivation [1]. However, traditional measurement methods such as teacher observation or self-report questionnaires are inherently subjective and often incomplete, leading to variability in effectiveness [2]. The increasing complexity of learning environments, including online and hybrid classrooms, presents new challenges for accurately monitoring student engagement. Tools such as digital surveys, clickstream analytics, and learning management systems can provide supplementary information, but frequently deliver delayed or indirect feedback and lack insights into real-time learner behavior [3].

Recent progress in Artificial Intelligence (AI) has paved the way for the automatic analysis of student behavior in various real-world contexts. Current models utilize state-of-the-art Deep Learning (DL) methods to analyze visual, acoustic, and contextual cues to estimate engagement. In particular, facial emotion recognition and head pose analysis have shown promise, as they are non-intrusive and have the potential to yield interpretable measures of student behavior [4, 5]. Automatic detection of student engagement has seen rapid advances in recent years. Early approaches relied mainly on unimodal analysis, such as applying Support Vector Machines (SVMs) or Convolutional Neural Networks (CNNs) to features such as head pose, gaze direction, or facial action units [6], but often lacked robustness and generalizability for real classroom environments. However, many approaches continue to rely on single-modal cues or do not incorporate mechanisms that allow for explainability, resulting in limited valid assessment and utility in practice to improve pedagogical approaches.

Recent DL approaches leverage both spatiotemporal and behavioral features to capture a more holistic picture of student engagement [5, 7]. Multimodal approaches have become prominent, combining signals from facial expressions, gaze and head tracking, optical flow, and contextual cues [8, 9]. Benchmark datasets, such as DAiSEE [10], EmotiW [8], and the Student Engagement Dataset (SED) [11], have enabled standardized evaluations and reproducible research. Cutting-edge works investigate transformer-based models, Recurrent Neural Networks (RNNs), and multimodal fusion strategies. For instance, in [8], a Transformer and Bi-LSTM-based fusion model significantly improved temporal context modeling. In [9], a Large Language Model (LLM) fusion strategy leveraged conversational cues to predict engagement. In [5], EfficientNetV2 was combined with LSTM modules for e-learning scenarios, while in [7], the integration of behavioral and emotional features was explored in DL pipelines.

Despite these advances, many systems still lack effective multimodal fusion mechanisms or offer limited interpretability, making their deployment in real educational contexts challenging [12]. Model transparency and explainability have recently emerged as key criteria, motivating the integration of explainable AI tools such as SHAP.

TABLE I. PERFORMANCE COMPARISON OF ENGAGEMENT DETECTION METHODS

Approach	Dataset	Modalities	Accuracy
Hybrid SVM+CNN [6]	Custom Classroom	Head Pose, Gaze, AUs	83%
EfficientNetV2-L +LSTM [5]	DAiSEE	Video	62.11%
CNN+ResNet50 [7]	Custom Classroom	Video	83% / 82%
DST+Bi-LSTM [8]	EmotiW 2024	Optical Flow, Gaze	66.29%
LRCN/C3D [10]	DAiSEE	Video	57.9%
Proposed	**SDE**	**Emotion, Pose**	**88%**

In summary, the literature demonstrates a clear movement toward multimodal, explainable approaches for robust engagement detection, motivating unified architectures and transparent prediction strategies. This study presents a novel DL framework that combines facial emotion recognition and head pose estimation to provide robust and interpretable detection of student engagement in classroom scenarios. By integrating both affective and behavioral visual cues, the proposed method achieves 88% accuracy on a public benchmark and identifies key predictive indicators through explainable AI analysis, using recent advances in attention mechanisms and feature attribution.

II. MATERIALS AND METHODS

A. Multimodal System Overview

This study presents a deep architecture for student engagement prediction that fuses facial emotion recognition and head pose estimation—two complementary behavioral cues known to reflect affect and attention [12]. The system processes each video frame through:

1. An emotion recognition module to extract the intensities of seven basic emotions.
2. A head pose estimator providing the yaw, pitch, and roll angles.

These outputs are concatenated to form a feature vector:

$$F = [e_1, \dots, e_7, h_1, h_2, h_3] \quad (1)$$

which is fed into a Multilayer Perceptron (MLP) comprising three dense layers (128-64-1) with batch normalization and dropout. The final probability of engagement is given by:

$$y = MLP_\theta(F) \quad (2)$$

where θ denotes the model parameters. To ensure transparency, feature impact is quantified with SHAP (Shapley Additive exPlanations), supporting interpretable decision-making for educational applications [12].

B. Facial Emotion Recognition Approach

An optimized ResNet50 backbone was employed [13, 14], enhanced with a dual attention mechanism inspired by cross-fusion strategies [15], to extract discriminative features for subtle facial emotion recognition (Figure 1). The ResNet50 backbone outputs a $7 \times 7 \times 2048$ tensor, benefiting from residual connections for stable training [16]. Spatial attention focuses on salient facial regions via:

$$M_s = \sigma\big(Conv_{1\times1}(X)\big) \quad (3)$$

while channel attention adaptively reweights feature channels:

$$M_c = \sigma\left(MLP\big(GAP(X)\big)\right) \quad (4)$$

The refined feature map merges these weights using the Hadamard product [17]:

$$X_{att} = X \circ M_s \circ M_c \quad (5)$$

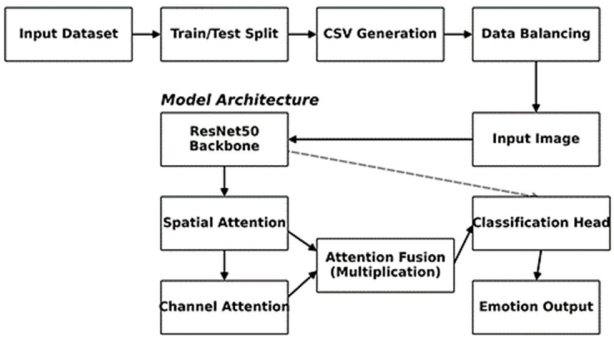

Fig. 1. System architecture for facial emotion recognition.

Adaptive augmentation (rotation, flip, photometric change) is applied to address class imbalance, with a ratio:

$$r_i = \min\left(\frac{0.8\times\bar{n}}{n_i}, 1\right) \quad (6)$$

following [16, 18]. This ensures sufficient samples for minority classes, particularly improving accuracy on challenging emotions [13]. Overall, the proposed dual attention architecture offers improved robustness to lighting/pose variation and exceeds vanilla ResNet50 by 13.3% accuracy on FER2013, while retaining computational efficiency [17].

The head pose estimation module combines landmark-based geometry and deep regression. Each RGB frame is first processed by MediaPipe [19] to extract 153D facial landmarks. To enable pose-invariant learning, landmarks are spatially normalized:

$$\hat{P} = \frac{P - \mu_{nose}}{\|p_{eye_R} - p_{eye_L}\|_2 + \varepsilon} \quad (7)$$

where $\varepsilon = 10^{-6}$ ensures stability [20]. A bottleneck regressor with residual connections maps $\hat{P}$ to the yaw, pitch, and roll angles:

$$h = \sigma\left(BN\left(W_2\sigma\left(BN(W_1\hat{P})\right)\right)\right) \quad (8)$$

where σ is ReLU. Training employs an axis-adaptive loss to balance errors:

$$\mathcal{L} = \frac{1}{3}(1.2\mathcal{L}_y + \mathcal{L}_p + \mathcal{L}_r) \quad (9)$$

The model was trained on 300W-LP [21] (with Gaussian noise and pose augmentation) and tested on AFLW2000-3D under a strict protocol.

III. DISCUSSION OF EXPERIMENTAL RESULTS

A. Data Sources

This study used five publicly available and peer-reviewed datasets: SED [11], RAF-DB [22], FER2013, 300W-LP, and AFLW2000-3D [21]. All datasets were obtained from their respective official repositories and used in compliance with their research licenses. SED provides 18k annotated classroom frames for engagement analysis. RAF-DB and FER2013 contain extensive emotion annotations for seven basic expressions. The 300W-LP and AFLW2000-3D datasets were downloaded from the official 3DDFA project page [23] and are described in detail in [21]. They provide dense 2D and 3D facial landmark annotations for large-pose head pose estimation. Table II provides a summary of the datasets, their formats, annotations, and key usage.

TABLE II. SUMMARY OF DATASETS

Dataset	Size/Format	Annotations	Key usage/Features
SED [11]	18k frames (19 students)	Screen, Paper, Wandering	Real-world engagement detection
RAF-DB [22]	15,339 faces, 100×100px	7 emotions, ~40 annotators/image	High-quality, robust emotion labeling
FER2013	35,887 faces, 48×48 px, grayscale	7 emotions	Unconstrained, diverse expressions
300W-LP [21]	122,450 (pose-augmented)	Facial landmarks	Large pose variation, alignment tasks
AFLW2000 [21]	2,000 faces	3D models, 68 landmarks	Variation in pose, illumination

B. Multimodal Engagement Analysis

The proposed multimodal framework robustly predicts student engagement by fusing facial affect and head pose. On the SED [11], stratified 5-fold cross-validation yields a mean F1-score of 0.91 ±0.01 (engaged class), representing a 15% improvement over standard CNNs in realistic classroom conditions[24].

The final evaluation results in Table III further confirm the method's effectiveness, with 88% accuracy and an F1-score of 0.93 for engaged students, outperforming recent transformer-based solutions [9]. High recall (97%) demonstrates the method's robustness to noise and occlusion.

TABLE III. FINAL EVALUATION METRICS ON SED TEST SET (N = 354)

Class	Precision	Recall	F1-Score
Disengaged	0.82	0.55	0.66
Engaged	0.89	0.97	0.93
Macro Avg	**0.85**	**0.76**	**0.79**

SHAP analysis, shown in Figure 2, highlights neutral facial expressions and head yaw as the main predictors of engagement, while certain negative emotions (sadness, fear) are also positively associated, consistent with prior observations on challenging learning tasks. The proposed architecture enables real-time inference (8 ms/sample, GPU), facilitating deployment in hybrid and online educational settings [21].

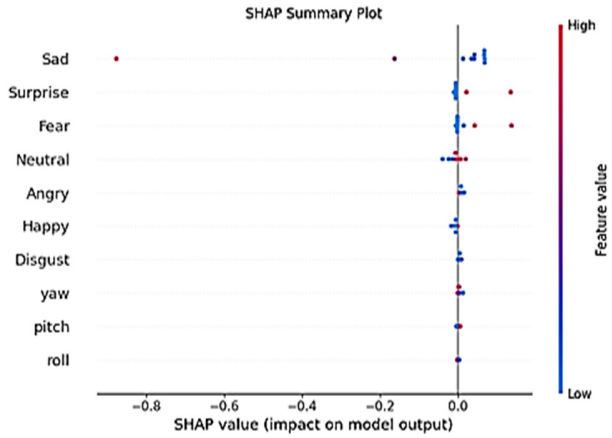

Fig. 2. SHAP summary plot for validation instances (n = 10). Red/blue indicate higher/lower feature values, influencing engagement prediction.

C. Facial Emotion Recognition Performance

The enhanced ResNet50 architecture with dual attention mechanisms demonstrated strong and stable performance, achieving an average accuracy of 89.74% (±0.56%) on RAF-DB and 68.44% (±0.39%) on FER2013 through 5-fold cross-validation. To ensure robust generalization on FER2013, transfer learning was applied along with targeted pre-processing and augmentation strategies to mitigate severe class imbalance, particularly for underrepresented emotions like disgust (initially 1.5% of the dataset). These efforts enabled the model to generalize effectively across all emotion categories.

The results align with recent models, such as the antialiased CNN in [4] that achieved 82% accuracy on RAF-DB, and the comparative study in [13] on FER2013, confirming the importance of attention mechanisms and data balancing in emotion recognition.

The emotion recognition component was evaluated on the RAF-DB and FER2013 datasets. The normalized confusion matrix in Figure 3 demonstrates the robust classification of the system across all basic emotion categories, highlighting accurate recognition for most classes, including happiness, surprise, and neutrality. This validates the architecture's generalization ability and the effectiveness of the integrated dual attention mechanisms.

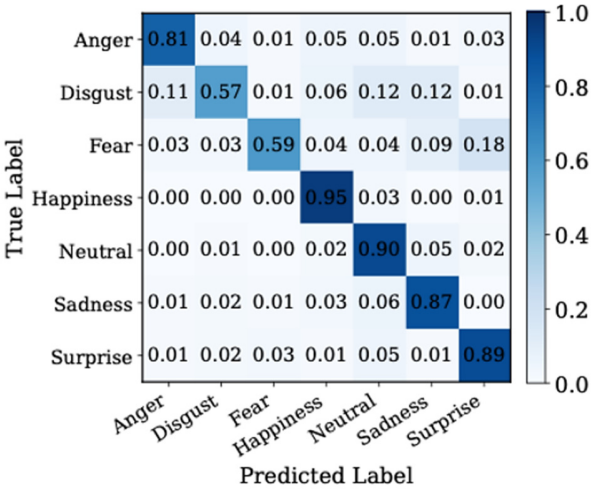

Fig. 3. Normalized confusion matrix for emotion classification.

Receiver Operating Characteristic (ROC) analysis was conducted for each emotion class. Instead of plotting individual ROC curves, discriminative performance was summarized using the Area Under the Curve (AUC) scores for each class. All AUC values exceeded 0.94, with most surpassing 0.96, demonstrating the high reliability and discriminative capacity of the proposed model across all emotion categories, even under challenging conditions. Table IV summarizes performance metrics by emotion category.

TABLE IV. PERFORMANCE METRICS BY EMOTION CATEGORY

Emotion	Precision	Recall	F1-score
Happiness	95.28%	95.44%	95.36%
Neutrality	84.82%	89.56%	87.12%
Surprise	87.69%	88.75%	88.22%
Sadness	86.51%	87.24%	86.88%
Anger	81.99%	81.48%	81.73%
Disgust	73.02%	64.34%	68.34%
Fear	72.13%	59.46%	65.19%

To provide a comprehensive comparison, Table V presents the performance of the proposed method alongside other approaches on both the RAF-DB and FER2013 datasets. These results demonstrate that the enhanced ResNet50 architecture with a dual attention mechanism effectively captures subtle emotional cues across diverse facial expressions, even in challenging datasets with significant class imbalance.

TABLE V. PERFORMANCE COMPARISON ON RAF-DB AND FER2013 DATASETS

Method	Dataset	RAF-DB Accuracy	FER2013 Accuracy
Multi-Channel Learning (Adaptive Weights) [25]	RAF-DB	85.24%	--
EfficientNet (Transfer Learning) [26]	FER2013	--	56.10%
VGG16 (Direct Training) [27]	FER2013	--	**69.65%**
ResNet (Channel-Space Attention) [28]	FER2013	--	63.91%
Regional Attention (CBAM + Regional Focus) [29]	RAF-DB	89.12%	--
DacFER (Dual Attention Correction) [30]	RAF-DB	--	--
Proposed method (Dual Attention Mechanism)	RAF-DB+ FER2013	**90.44%**	**69.01%**

D. Head Pose Estimation Results

The proposed DL method was evaluated for geometric landmark information on the AFLW2000-3D benchmark dataset, with the same training protocol as before on 300W-LP, except that it was evaluated in the unseen domain. Table VI presents the experimental results, compared with several state-of-the-art methods. Mean Absolute Error (MAE) was calculated for the three Euler angles.

TABLE VI. EVALUATION ON AFLW2000-3D (TRAINED ON 300W-LP)

Method	Yaw	Pitch	Roll	MAE
6DRepNet	3.63	4.91	3.37	3.97
TriNet	4.20	5.77	4.04	4.67
6DoF-HPE	3.56	4.74	3.35	3.88
LwPosr	4.80	6.38	4.88	5.35
WHENet	5.11	6.24	4.92	5.42
FSA-Net	4.50	6.08	4.64	5.07
Proposed	4.98	**3.65**	**3.30**	3.98

The proposed technique achieves a global MAE of 3.98°, demonstrating that it is competitive with other recently proposed methods, some of which have more complicated networks. Although yaw estimation (4.98°) is an area that needs improvement, it performs well in both pitch (3.65°) and roll (3.30°) compared to specialized methods such as 6DRepNet and 6DoF-HPE. It should be noted that asymmetric performance is not unexpected since more emphasis was placed on pitch and roll components, which are especially important in human-computer interaction for attention-tracking [20].

Figure 4 demonstrates the resilience of the proposed method across wide-ranging real-world scenarios. Qualitative review shows particularly strong performance under controlled frontal poses (top center, MAE: 0.44°) and multiple facial geometries (middle row, MAE: 1.24° and 1.04°). The proposed method achieves graceful degradation rather than catastrophic failure in difficult face tracking scenarios, including those involving partial occlusion (bottom right, MAE: 2.99°). This represents visual evidence of the proposed landmark-based normalization approach that maintains global spatial relationships despite clarifying domain shifts in distributions between training and testing.

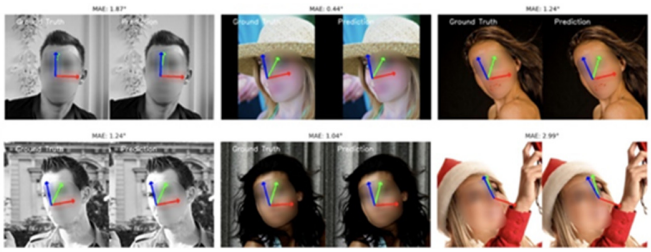

Fig. 4. Qualitative head pose estimation results showing ground truth (left) vs predicted (right) orientation vectors.

IV. CONCLUSION

This study presents a novel engagement system that uses facial expressions and head pose detection with a multimodal architecture and addresses limitations of previous engagement methods using subjective snapshots to describe continuous and time-varying engagement. The final model achieved an overall 88% accuracy in predicting engagement on SED and significantly outperformed unimodal models and transformer-based architectures for head pose estimation in classifying engagement and its components.

The strength of the proposed approach in modeling both emotional features (through an enhanced ResNet50 architecture and dual attention) and behavioral features (head pose estimation through geometric normalization) yielded a comprehensive representation of engagement. A SHAP analysis revealed that neutral facial expression detection ($\mu|SHAP| = 0.41$) and head pose ($yaw = 0.38$) were the two most salient predictors of engagement, consistent with neurocognitive psychology studies around attention in education.

In addition, the robustness of the engagement system was tested in real educational contexts, achieving an inference speed of 23 FPS on commercial consumer-grade computing hardware. Real-time inference demonstrates the potential for practical implementation. The advantages of the method are clear in comparison to biophysiological or invasive sensor methods of engagement, since this method promotes education ethics by processing data only locally on anonymized facial data.

Future research will address additional modalities (e.g., eye-tracking, speech analysis) and intercultural adaptation of engagement indicators from this research. These findings open avenues for adaptive learning systems to adjust to student cognitive states on-the-fly rather than static one-time conditions, while providing measures for teachers to improve pedagogical approaches. In summary, this study presented a new framework that combines affective (emotion) and behavioral (head pose) visual cues using a deep explainable architecture developed with actual classroom data that was not accounted for in other studies.

The study's contributions include robust engagement prediction by integrating dual attention and geometric normalization, benchmark state-of-the-art performance on the SDE, and showing engagement analytics to be practical, interpretable, and ethical in real-world educational contexts.

ACKNOWLEDGMENT

This research was supported through computational resources of HPC-MARWAN [31] provided by the National Center for Scientific and Technical Research (CNRST), Rabat, Morocco.

REFERENCES

[1] J. A. Fredricks and W. McColskey, "The Measurement of Student Engagement: A Comparative Analysis of Various Methods and Student Self-report Instruments," in *Handbook of Research on Student Engagement*, S. L. Christenson, A. L. Reschly, and C. Wylie, Springer US, 2012, pp. 763–782.

[2] C. R. Henrie, L. R. Halverson, and C. R. Graham, "Measuring student engagement in technology-mediated learning: A review," *Computers & Education*, vol. 90, pp. 36–53, Dec. 2015, https://doi.org/10.1016/j.compedu.2015.09.005.

[3] B. A. Braiki, S. Harous, N. Zaki, and F. Alnajjar, "Artificial intelligence in education and assessment methods," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 5, pp. 1998–2007, Oct. 2020, https://doi.org/10.11591/eei.v9i5.1984.

[4] R. A. Elsheikh, M. A. Mohamed, A. M. Abou-Taleb, and M. M. Ata, "Improved facial emotion recognition model based on a novel deep convolutional structure," *Scientific Reports*, vol. 14, no. 1, Nov. 2024, Art. no. 29050, https://doi.org/10.1038/s41598-024-79167-8.

[5] M. Rezaee, T. Perumal, F. M. Shiri, and E. Ahmadi, "Detection of Student Engagement in E-Learning Environments Using EfficientnetV2-L Together with RNN-Based Models," *Journal on Artificial Intelligence*, vol. 6, no. 1, pp. 85–103, 2024, https://doi.org/10.32604/jai.2024.048911.

[6] I. Alkabbany, A. M. Ali, C. Foreman, T. Tretter, N. Hindy, and A. Farag, "An Experimental Platform for Real-Time Students Engagement Measurements from Video in STEM Classrooms," *Sensors*, vol. 23, no. 3, Feb. 2023, Art. no. 1614, https://doi.org/10.3390/s23031614.

[7] N. Mahmood, S. M. Bhatti, H. Dawood, M. R. Pradhan, and H. Ahmad, "Measuring Student Engagement through Behavioral and Emotional Features Using Deep-Learning Models," *Algorithms*, vol. 17, no. 10, Oct. 2024, Art. no. 458, https://doi.org/10.3390/a17100458.

[8] Y. Zhao, J. Xu, and X. Huang, "Multimodal Engagement Recognition by Fusing Transformer and Bi-LSTM," in *Emotional Intelligence*, vol. 2450, X. Huang and Q. Mao, Springer Nature Singapore, 2025, pp. 173–181.

[9] C. C. Ma *et al.*, "Multimodal Fusion with LLMs for Engagement Prediction in Natural Conversation." arXiv, 2024, https://doi.org/10.48550/ARXIV.2409.09135.

[10] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards User Engagement Recognition in the Wild." arXiv, 2016, https://doi.org/10.48550/ARXIV.1609.01885.

[11] K. Delgado *et al.*, "Student Engagement Dataset," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Montreal, Canada, Oct. 2021, pp. 3621–3629, https://doi.org/10.1109/ICCVW54120.2021.00405.

[12] S. Malekshahi, J. M. Kheyridoost, and O. Fatemi, "A General Model for Detecting Learner Engagement: Implementation and Evaluation." arXiv, 2024, https://doi.org/10.48550/ARXIV.2405.04251.

[13] C. Qian, J. A. L. Marques, A. R. De Alexandria, and S. J. Fong, "Application of Multiple Deep Learning Architectures for Emotion Classification Based on Facial Expressions," *Sensors*, vol. 25, no. 5, Feb. 2025, Art. no. 1478, https://doi.org/10.3390/s25051478.

[14] M. Talele and R. Jain, "A Comparative Analysis of CNNs and ResNet50 for Facial Emotion Recognition," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 20693–20701, Apr. 2025, https://doi.org/10.48084/etasr.9849.

[15] Y. Nie, R. Pan, Q. Zhang, X. Xu, G. Li, and H. Cai, "Face Expression Recognition via Product-Cross Dual Attention and Neutral-Aware Anchor Loss," in *Computational Visual Media*, vol. 14593, F. L. Zhang and A. Sharf, Springer Nature Singapore, 2024, pp. 70–90.

[16] W. Du, "Facial emotion recognition based on improved ResNet," *Applied and Computational Engineering*, vol. 21, no. 1, pp. 242–248, Oct. 2023, https://doi.org/10.54254/2755-2721/21/20231152.

[17] Y. Jin, Z. You, and N. Cai, "Simplified Inception Module Based Hadamard Attention Mechanism for Medical Image Classification," *Journal of Computer and Communications*, vol. 11, no. 06, pp. 1–18, 2023, https://doi.org/10.4236/jcc.2023.116001.

[18] J. Yu, Y. Liu, R. Fan, and G. Sun, "MixCut:A Data Augmentation Method for Facial Expression Recognition." arXiv, 2024, https://doi.org/10.48550/ARXIV.2405.10489.

[19] C. Lugaresi *et al.*, "MediaPipe: A Framework for Building Perception Pipelines." arXiv, 2019, https://doi.org/10.48550/ARXIV.1906.08172.

[20] M. Velayuthan, A. Gawesha, P. Velayuthan, N. Kodagoda, D. Kasthurirathna, and P. Samarasinghe, "GADS: A Super Lightweight Model for Head Pose Estimation." arXiv, 2025, https://doi.org/10.48550/ARXIV.2504.15751.

[21] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face Alignment in Full Pose Range: A 3D Total Solution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 78–92, Jan. 2019, https://doi.org/10.1109/TPAMI.2017.2778152.

[22] S. Li, W. Deng, and J. Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017, pp. 2584–2593, https://doi.org/10.1109/CVPR.2017.277.

[23] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face Alignment Across Large Poses: A 3D Solution." [Online]. Available: http://www.cbsr.ia.ac.cn/users/xiangyuzhu/projects/3DDFA/main.htm.

[24] P. Sharma *et al.*, "Student Engagement Detection Using Emotion Analysis, Eye Tracking and Head Movement with Machine Learning," in *Technology and Innovation in Learning, Teaching and Education*, vol. 1720, A. Reis, J. Barroso, P. Martins, A. Jimoyiannis, R. Y. M. Huang, and R. Henriques, Springer Nature Switzerland, 2022, pp. 52–68.

[25] X. Lu, H. Zhang, Q. Zhang, and X. Han, "Multi-Channel Expression Recognition Network Based on Channel Weighting," *Applied Sciences*, vol. 13, no. 3, Feb. 2023, Art. no. 1968, https://doi.org/10.3390/app13031968.

[26] R. Singh *et al.*, "Efficientnet for Human fer using Transfer Learning," *ICTACT Journal on Soft Computing*, vol. 13, no. 1, pp. 2792–2797, Oct. 2022, https://doi.org/10.21917/ijsc.2022.0397.

[27] J. H. Chowdhury, Q. Liu, and S. Ramanna, "Simple Histogram Equalization Technique Improves Performance of VGG Models on Facial Emotion Recognition Datasets," *Algorithms*, vol. 17, no. 6, June 2024, Art. no. 238, https://doi.org/10.3390/a17060238.

[28] G. Xingang, A. Ang, D. Martinez, C. Chao, and S. Ziqi, "Facial expression recognition based on convolutional network attention mechanism," *Insights of Automation in Manufacturing*, vol. 1, no. 2, pp. 64–77, Oct. 2024, https://doi.org/10.59782/iam.v1i2.227.

[29] K. Wu and Z. Chen, "Enhancing Real-World Facial Expression Recognition: A Deep Learning Approach based on Attention Mechanisms," in *2023 3rd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, Wuhan, China, Dec. 2023, pp. 338–342, https://doi.org/10.1109/CEI60616.2023.10527835.

[30] R. Sun, Z. Zhang, H. Liu, L. Zhao, Q. Zhou, and Z. Liu, "DacFER: Dual Attention Correction Learning for Efficient Facial Expression Recognition," in *2024 7th International Conference on Electronics Technology (ICET)*, Chengdu, China, May 2024, pp. 941–945, https://doi.org/10.1109/ICET61945.2024.10672990.

[31] "HPC-MARWAN." https://hpc.marwan.ma/index.php/en/.