RESEARCH ARTICLE

An Integrated Deep Learning Framework for Real-Time Monitoring of Student Engagement in Smart Classrooms

Muhammad Imran Khalid[1,2,3], **Asma Komal**[1,3,*], **Usman J. Wushishi**[1], **Altaf Hussain**[1], **Nasir Hussain**[1] **and Syed Akif Hussain**[4]

[1] School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2] Chongqing Key Laboratory of Image Cognition, Chongqing university of Posts and Telecommunications, Chongqing 400065, China
[3] School of Computer Science and Technology, University of Okara, Okara 56300, Pakistan
[4] School of Economics and Management, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract

Studies have established that an ideal environment is critical for maximizing a student's learning potential. For educators, monitoring student behavior, engagement, and psychological state is essential for ensuring effective instruction. This paper introduces an automated learning analytics system designed to assist teachers by analyzing these parameters and providing actionable feedback. The system utilizes multiple cameras in conjunction with deep learning and computer vision to record and analyze classroom sessions, assessing student movement, gestures, and posture to generate summary reports. The framework integrates several high-performance models to achieve this. For facial recognition, the FaceNet technique is employed, demonstrating 99.63% accuracy on frontal profiles from labeled images and 95.12% on the YouTube Faces DB dataset. Concurrently, a Convolutional Neural Network (CNN) is used for expression monitoring, achieving competitive performance in identifying five distinct emotions. Furthermore, body posture is tracked and analyzed using OpenPifPaf, which shows a high-performance index greater than 0.8 on multiple benchmarks, including the CrowdPose and COCO datasets. By combining these technologies, the system provides a robust and accurate solution for automated, real-time student behavior monitoring.

Keywords: student behavior monitoring, smart classroom, deep learning, CNN, classroom monitoring.

1 Introduction

A student's academic performance is influenced by numerous factors that can affect their concentration and learning outcomes, including the quality of the

Submitted: 30 September 2025
Accepted: 29 October 2025
Published: 14 December 2025

Vol. 1, **No.** 4, 2025.
10.62762/JIAP.2025.377388

***Corresponding author:**
Asma Komal
imran.khalis292@gmail.com

Citation
Khalid, M. I., Komal, A., Wushishi, U. J., Hussain, A., Hussain, N., & Hussain, S. A. (2025). An Integrated Deep Learning Framework for Real-Time Monitoring of Student Engagement in Smart Classrooms. *ICCK Journal of Image Analysis and Processing*, 1(4), 172–183.

learning environment [17]. For educators, monitoring student behavior and psychological state is essential for effective instruction [9]. Among these, a student's level of focus during lectures is a critical determinant of success. Research consistently shows that student class attendance and general engagement directly impact academic results [12], making it essential for educators to identify and mitigate focus anomalies within the classroom. However, manually monitoring an entire class for signs of disengagement is a significant challenge, driving the need for automated solutions in learning analytics [3]. To address this, our system is designed to help teachers and administrators monitor students in real-time, promoting sustained focus and maximizing the retention of lecture material [26–28].

Table 1 summarizes various characteristics of inattentive students as identified in prior research [6].

In the modern era, the proliferation of personal electronic devices has become a primary source of distraction in the classroom [16]. Cognitive studies confirm that the human brain does not truly multitask but rather shifts focus rapidly between tasks. This cognitive switching prevents deep processing and hinders the transfer of information to long-term memory. A key study demonstrated this effect empirically; students who used electronic devices during lectures, despite showing similar in-class quiz performance, scored significantly lower on final exams compared to their peers who did not [15]. As illustrated in Figure 1, the variations in performance across different conditions of device usage clearly illustrate this negative impact. These findings underscore the importance of intelligent monitoring systems in classroom management [30].

This paper proposes an automated system to help observe and analyze student engagement by leveraging deep learning and computer vision. The system processes video data to provide feedback to the teacher on students' body language, movements, and facial expressions. We have developed a new, integrated platform that can simultaneously identify faces, monitor expressions, and analyze body posture in real-time. By unifying these capabilities, the system provides a holistic analytical report based on its training, improving upon the accuracy of existing, separate systems [29, 30].

To achieve this, the system integrates three core technologies. Facial recognition is performed using deep learning models that map faces to an embedding space for accurate identification [1]. Student posture and body language are analyzed using robust pose estimation and tracking techniques [2, 7]. Finally, a Convolutional Neural Network (CNN) is used to monitor facial expressions and infer students' emotional and cognitive states during the lecture [4, 11]. The primary objectives of this work are to: 1) estimate and track student posture throughout the lecture, 2) provide a mechanism for automated attendance via face recognition [10], and 3) monitor classroom expressions to help teachers gauge student comprehension and engagement levels.

1.1 Our Contributions

The key contributions of this paper are as follows:

Table 1. Those students who are not attentive in class and divide their attention [6].

Theme	CODE
Social characteristics	Family problems, lack of skills, the loser in responsibilities, the cloak in clothing, etc.
Personality characteristics	Careless, restless, careless, non-compliant, etc.
Psychological characteristics	EXTREME CONCERN, GETTING BORED, DISTRACTION, LACK OF DIRECT INVOLVEMENT, PETULANCE, LACK OF PURPOSE, SHYNESS, SILENCE, FORGETFULNESS, ETC.
Academic features	Bad grades, low learning rate, lack of interest in the course, learning disabilities, lack of initial information, etc.
Behavioral forms	Incessant speaking, bad attitudes, and behaviors to the teacher, attached with mobile phones, playing games, lack of devices, being in the back row, sleeping in class, not attending the class, etc.
Physical characteristics	Restless, weary, starving, health problems, etc.

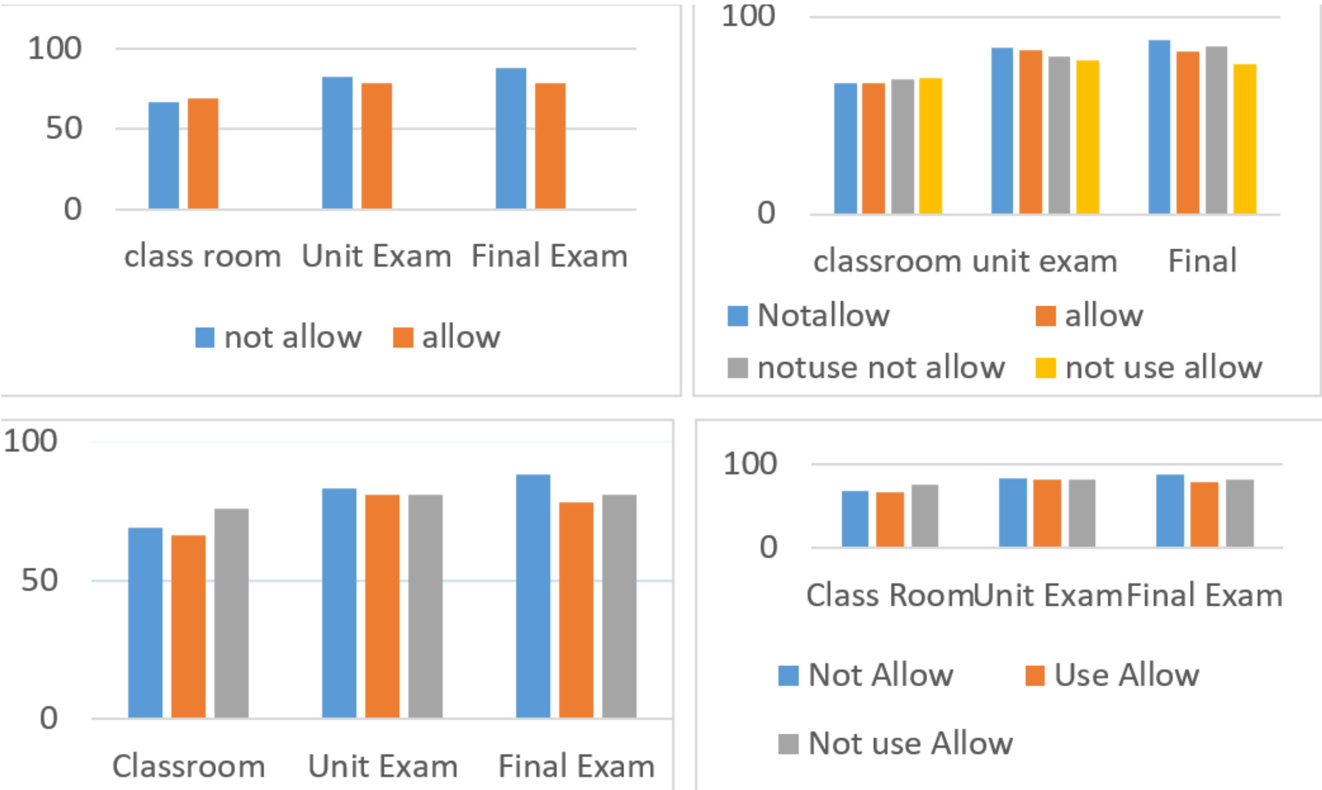

Figure 1. Comparison of student exam performance with and without the use of electronic devices during lectures demonstrates a negative impact on final exam scores [15].

1. **A Real-Time Pose Estimation and Tracking Model**: We present a model tailored for the classroom environment, capable of estimating and tracking student posture and body language throughout a lecture to gauge attentiveness.

2. **An Automated Attendance System**: We contribute an automated system that utilizes high-accuracy facial recognition to seamlessly mark student presence without manual intervention or disruption to the class.

3. **A Facial Expression Monitoring System**: We have developed a system that monitors student facial expressions in real-time, providing educators with valuable insights into student engagement, confusion, and comprehension levels.

4. **A Unified Framework**: The primary contribution is the novel integration of these three distinct models into a single, cohesive framework for the holistic, real-time analysis of student behavior.

2 Related Work

This section reviews the existing literature across three main areas: the educational basis for monitoring student engagement, advancements in computer vision for human analysis, and previous work on automated classroom systems.

2.1 Student Engagement and Classroom Challenges

The foundation of our research lies in the established link between a student's behavior and their academic success. The learning environment is a critical performance factor, and the use of academic analytics has been crucial in measuring and improving educational outcomes [3, 17]. A student's ability to engage with material is directly influenced by their study behaviors and psychological state [9], and their active presence in the classroom is a key indicator of their learning experience [12]. A primary challenge in modern education is the prevalence of distractions that divide student attention, particularly the use of personal electronic devices [16]. A pivotal study demonstrated that using electronic devices during lectures severely impairs academic performance, even on final exams, supporting the cognitive theory that rapid task-switching prevents deep learning [15]. These findings underscore the need for effective tools that can monitor and manage student focus [26–28]. Recent work has explored deep learning-based intelligent classroom teaching frameworks for quality analysis [27] and the use of the SVM algorithm for monitoring teaching dynamics in IoT-based smart classrooms [29].

2.2 Advancements in Computer Vision for Human Analysis

To address these challenges, our work leverages computer vision and deep learning. The development of deep Convolutional Neural Networks (CNNs) was pioneering [5]. Modern architectures, such as ShuffleNet V2, provide practical guidelines for efficient CNN design [24]. Techniques like Batch Normalization are essential for stabilizing and accelerating the training process of these deep networks [21]. The ability to compress deep neural networks is also an ongoing area of study [8].

Facial Recognition: Deep learning models have dramatically advanced facial recognition performance. Advanced methods, such as FaceFilter, which combine deep learning with filter algorithms, have been proposed for accurate face identification and recognition [1]. This technology is successfully applied to practical problems like automated attendance marking [10]. The Labeled Faces in the Wild (LFW) dataset serves as a fundamental benchmark for unconstrained face recognition studies [19]. Research continues to evolve, including exploring multi-view 3D methods for keypoint estimation [23].

Expression Monitoring: Beyond identity, facial expressions provide a window into a student's cognitive and emotional state. While traditional machine learning methods using Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP) have been explored [6, 13], CNNs have become the dominant approach for their ability to learn spatial hierarchies of features [4, 11]. Advanced techniques focus on emotion recognition, even from challenging visual data [22].

Human Pose Estimation: A student's posture is a strong indicator of their level of engagement. Human pose estimation has seen remarkable progress with deep learning. Approaches are often categorized as top-down, like Mask R-CNN [18], which first detects the person, or bottom-up, such as those using Part Affinity Fields (PAFs) [7], which are particularly robust in crowded scenes. Systems are often evaluated using benchmarks like PoseTrack [2]. For video analysis, pose tracking is essential, achieved by utilizing temporal flow fields to link poses across consecutive frames [14, 20].

2.3 Our Contribution in Context and State-of-the-Art Comparison

While previous studies have focused on single-modality monitoring, such as attention [26], general behavior [29], or specific network enhancements [25], the contribution of our work is the novel integration of state-of-the-art models across three distinct modalities: facial identity recognition [1], expression monitoring for emotional feedback [11], and pose estimation for engagement analysis [7]. By combining these powerful techniques into a single, real-time framework, we provide a more holistic and nuanced solution for classroom analytics than previously proposed intelligent monitoring systems [30]. (see Table 2)

3 Existing Systems and Technologies

To provide context for our work, this section reviews the existing literature and state-of-the-art technologies in three key areas of computer vision: face identification, expression monitoring, and human pose estimation.

3.1 Face Identification Models

Facial identification has evolved from a specialized tool for security to a ubiquitous technology. Early methods often relied on techniques that were less robust with small datasets. The advent of deep learning has significantly advanced the field, with models that utilize filter algorithms to improve prediction identities with high precision [1]. Other deep learning structures have demonstrated robustness against challenging conditions like image blur and varied poses [23]. Automated attendance systems are a key application, often utilizing facial recognition with methods like histogram analysis [10].

3.2 Expression Monitoring Models

Expression monitoring utilizes computer vision and deep learning to analyze facial muscle movements, providing insights into a person's emotional state. Early models often employed classifiers like Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP) with notable success [6, 13]. However, Convolutional Neural Networks (CNNs) have become the state-of-the-art for their ability to automatically learn discriminative features [4, 11]. Advanced architectures and techniques are constantly being developed for accurate recognition, even in challenging scenarios [22].

Table 2. State-of-the-Art comparison of intelligent classroom monitoring and component technologies.

Feature / Model	Key Approach	Modalities / Focus	Citation	Contribution Type	Reported Performance
Intelligent Behavior System	Hybrid Deep Learning	Student behavior recognition in the classroom	[26]	System Integration	–
Intelligent Teaching Framework	Deep Learning-based Analysis	Teaching quality and classroom framework	[27]	System Integration	–
Intelligent Monitoring System	IoT, SVM Algorithm	Smart Classroom Management, IoT integration	[29]	System / Algorithm	–
Face Identification (prior)	Deep Learning & Filter Algorithm	High-accuracy face recognition	[1]	Algorithm Refinement	–
Automated Attendance	Facial Recognition (Histogram)	Practical attendance system	[10]	Application	–
Realtime Multi-Person 2D Pose	Part Affinity Fields (Bottom-up)	Robust pose estimation in crowded scenes	[7]	Core Model	–
Multi-Person Tracking	Temporal Flow Fields	Tracking individuals across video frames	[14]	Tracking Technique	–
Emotion Recognition	Cross-Channel CNN	Expression recognition for human-robot interaction	[4]	Core Model	–
Expression Classification	SVM, KNN, MLP Classifiers	Traditional expression analysis	[13]	Baseline Method	–
Core Deep Learning Component	Batch Normalization	Stabilizing and accelerating CNN training	[21]	Optimization	–
Face Identification (Ours)	FaceNet-based DCNN (Triplet Loss)	Student identification / attendance	This work	Module / Application	99.63% (frontal); 95.12% (YTF)

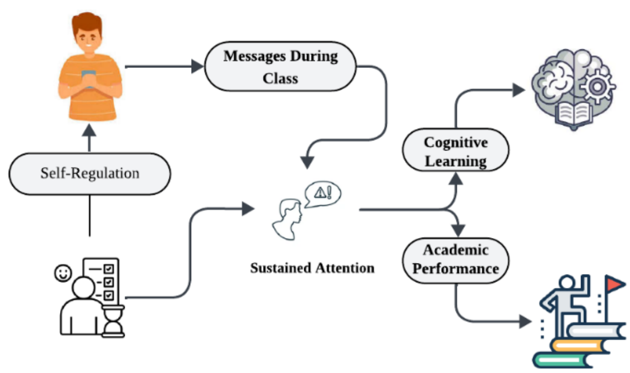

Figure 2. Conceptual model illustrating the hypothesis that Self-Regulation (SR) and Text Messaging (TM) during class influence a student's Sustained Attention, which in turn affects Cognitive Learning (CL) and Academic Performance.

3.3 Human Pose Estimation Models

Pose estimation is a computer vision technique that analyzes body posture and language from images or videos. Deep learning has driven significant advancements, with models achieving high accuracy on complex datasets [18]. Both top-down (like Mask R-CNN [18]) and bottom-up approaches (like Part Affinity Fields [7]) are used for multi-person pose estimation, with the latter being particularly effective in crowded scenes. Systems are often evaluated using benchmarks like PoseTrack [2]. For video analysis, pose tracking is essential, achieved by utilizing temporal flow fields to link poses across consecutive frames [14, 20]. This relationship between student behavior and learning outcomes is conceptualized in Figure 2.

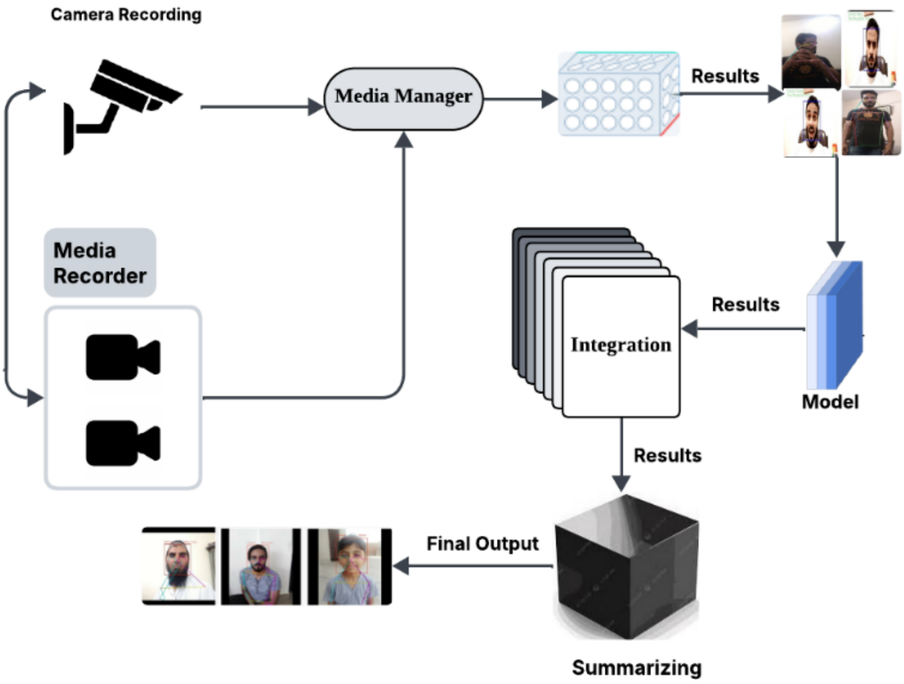

Figure 3. High-level data flow of the proposed system, from camera input to integrated analysis.

4 System Design and Methodology

This section details the architecture and core methodologies of our proposed integrated system. While the technologies described above exist as standalone solutions, our primary contribution is the novel integration of face identification, expression monitoring, and pose estimation into a single, cohesive framework designed for the classroom environment.

4.1 System Overview

Our system is designed as a real-time data retrieval and analysis pipeline. As shown in Figure 3, video streams from multiple cameras are fed into a media manager. This data is then processed by our three interconnected deep learning models. The models' outputs are integrated and summarized to produce an analytical report for the teacher, providing a holistic view of student behavior. The system is designed to be non-invasive and does not modify the raw captured data.

4.2 Face Identification Model

For robust student identification, we employ the FaceNet technique, which is based on a deep convolutional neural network (DCNN). The model learns to map a face image to a 128-dimensional vector in Euclidean space, an "embedding." The core of FaceNet's training is the triplet loss function.

$$L = \sum_{i}^{N} \left[\| f(x_i^a) - f(x_i^p) \|_2^2 - \| f(x_i^a) - f(x_i^n) \|_2^2 + \alpha \right]_+ \tag{1}$$

where $f(x)$ is the embedding generated by the DCNN, alpha is a margin enforced between positive and negative pairs, and the operation is the hinge loss function. This ensures that faces of the same person are clustered together in the embedding space, allowing for highly accurate recognition.

The triplet loss algorithm works by considering three images at a time: an "anchor" (an image of a specific person), a "positive" (a different image of the same person), and a "negative" (an image of a different person). The goal is to train the network to minimize the distance between the anchor and positive embeddings while maximizing the distance between the anchor and negative embeddings, as illustrated in Figure 4. This ensures that faces of the same person are clustered together in the embedding space, allowing for highly accurate recognition.

4.3 Expression Monitoring Model

Our system is built on a custom CNN architecture. The model is trained to classify faces into five emotional categories: angry, sad, shocked, happy, and neutral.

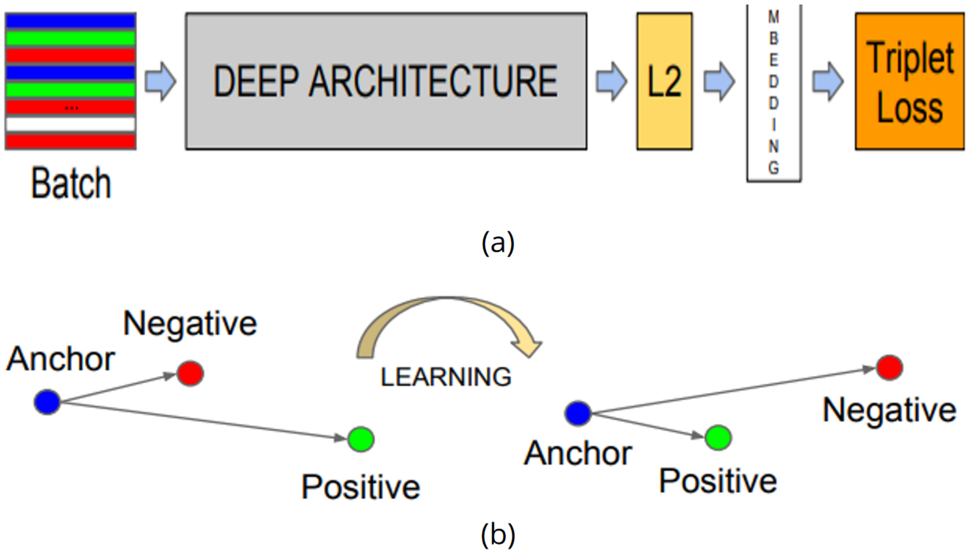

Figure 4. The DCNN model structure and its training objective: (a) a DCNN generates embeddings for face images, and (b) a loss function optimizes these embeddings by pulling the 'positive' sample closer to the 'anchor' and pushing the 'negative' sample farther away.

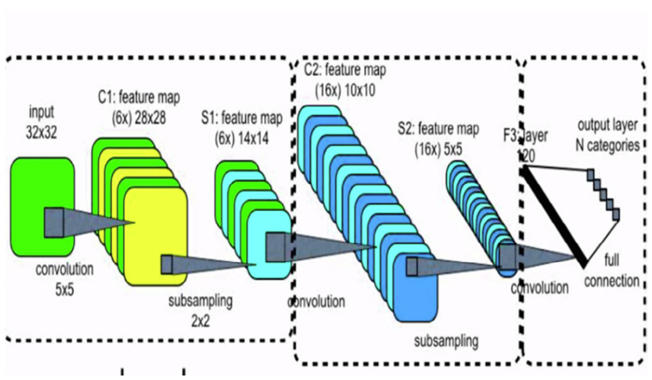

Figure 5. An illustration of the Convolutional Neural Network (CNN) architecture used for expression monitoring. The model processes an input image through a series of feature maps (convolution) and subsampling layers to a final classification output.

To prevent overfitting on our large dataset, we employ data augmentation, which programmatically creates variations of training images (e.g., rotations, shifts, zooms).

The CNN architecture, shown in Figure 5, consists of multiple convolutional layers (Conv2D) that learn hierarchical features from the input images. These are followed by max-pooling layers to reduce dimensionality while preserving important information. We utilize Batch Normalization to stabilize and accelerate the training process [21]. Finally, a fully connected layer with a softmax activation function performs the final classification.

As depicted in Figure 6, our CNN model provides a detailed visual representation of the network structure and its components. Our system is built on a custom CNN architecture trained to classify faces into five emotional categories: angry, sad, shocked, happy, and neutral. To prevent overfitting, we employ data augmentation, which programmatically creates variations of training images (e.g., rotations, shifts, zooms).

The CNN architecture consists of multiple convolutional layers that learn hierarchical features from the input images. These are followed by max-pooling layers to reduce dimensionality while preserving important information. We utilize the Rectified Linear Unit (ReLU) as our activation function to introduce non-linearity, which is essential for learning complex patterns.

$$f(x) = max(0, x) \tag{2}$$

4.4 Pose Estimation and Tracking Model

To analyze student posture and body language, we use a state-of-the-art bottom-up pose estimation method. This approach first detects all human keypoints (e.g., elbows, knees, shoulders) in an image and then groups them into person skeletons. This is highly effective in crowded scenes like a classroom. Our method is built on a CNN backbone (e.g., ResNet [1] or ShuffleNetV2 [2]) and introduces a composite loss function that simultaneously optimizes for keypoint confidence, localization, and scale.

For video analysis, we extend the single-image estimation to pose tracking. This is achieved using a

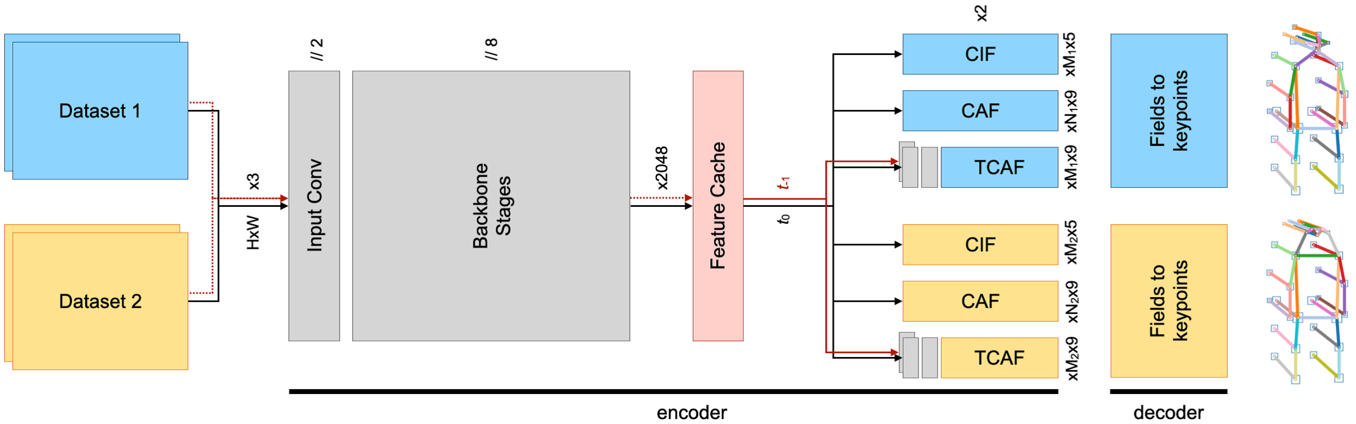

Figure 6. A visual representation of the CNN architecture used for expression monitoring.

Temporal Composite Association Field (TCAF), which links poses across consecutive frames. This allows the system to maintain a stable track of each student's posture over time, even with movement and temporary occlusions.

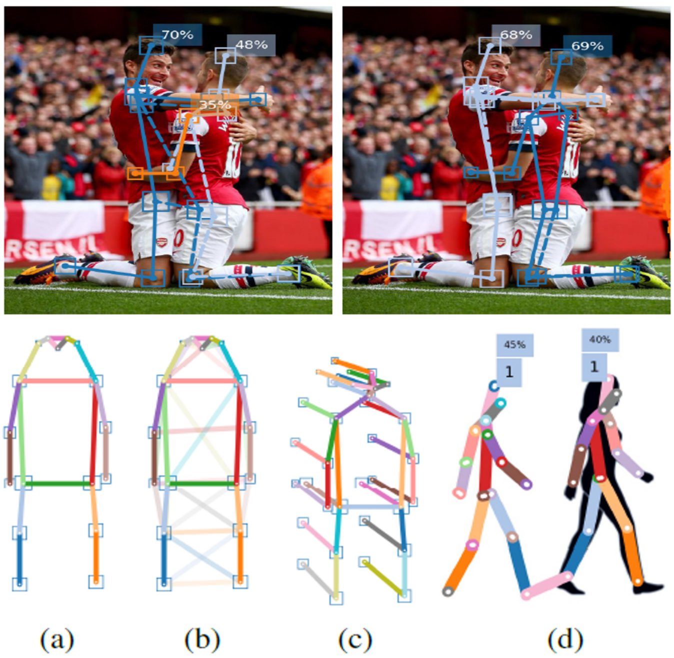

Figure 7. Pose estimation and tracking: (a) detected skeletons overlaid on individuals in a crowded scene, and (b) denser skeletons with key points first estimated per frame (pose estimation) and then linked across frames for tracking.

As demonstrated in Figure 7, the pose estimation and tracking process effectively captures student postures in classroom settings. The figure shows (a) detected skeletons overlaid on individuals in a crowded scene, and (b) the process of linking key points across frames for continuous tracking.

$$L = L_{\text{conf}} + \lambda_{\text{reg}} L_{\text{reg}} + \lambda_{\text{scale}} L_{\text{scale}} \quad (3)$$

In this equation, $\mathcal{L}_{\text{conf}}$ represents the confidence loss for detecting key points, $\mathcal{L}_{\text{reg}}$ is the L2 regression loss for refining their precise location, and $A_{\text{scale}} L_{\text{scale}}$ is a loss for predicting the correct scale of the body part. For video analysis, we extend this to pose tracking using a Temporal Composite Association Field (TCAF), which links poses across consecutive frames to maintain a stable track of each student over time.

5 Experiments and Results

This section describes the datasets, evaluation metrics, and experimental results for our integrated system.

5.1 Datasets

Our system was trained and evaluated on several standard academic datasets:

- **Face Identification**: We used the Labeled Faces in the Wild (LFW) and YouTube Faces DB datasets, which are standard benchmarks for face verification in unconstrained environments.
- **Pose Estimation**: We utilized the COCO dataset for general keypoint detection and the CrowdPose dataset, which is specifically designed for evaluating performance in crowded scenes.
- **Expression Monitoring**: Our model was trained on a custom dataset of five core emotions, with images processed to a 48x48 pixel resolution in grayscale.

5.2 Evaluation and Results

The performance of our system was evaluated on each of its core components. Pose Estimation Performance: On the challenging CrowdPose dataset, our pose estimation model (OpenPifPaf) achieves a state-of-the-art Average Precision (AP) of 70.0,

Table 3. Performance comparison of multi-person pose estimation methods on the CrowdPose dataset.

Method	AP	AP@0.50	AP@0.75	AP Easy	AP Medium	AP Hard	FPS
Mask R-CNN	60.3	69.4	57.9	45.8	29.0	60.3	29.0
AlphaPose	57.2	83.5	60.3	69.4	57.9	45.8	–
HigherHRNet-W48	61.0	∼	66.0	∼	∼	∼	∼
OpenPiFPaf (Ours)	70.0	89.1	76.1	78.4	72.1	63.8	13.7

outperforming many existing methods, especially in "hard" scenarios with an AP of 63.8. It operates at an efficient 13.7 FPS on a single GTX 1080Ti GPU. The performance on the COCO 2017 dataset is also highly competitive, as shown in Table 3.

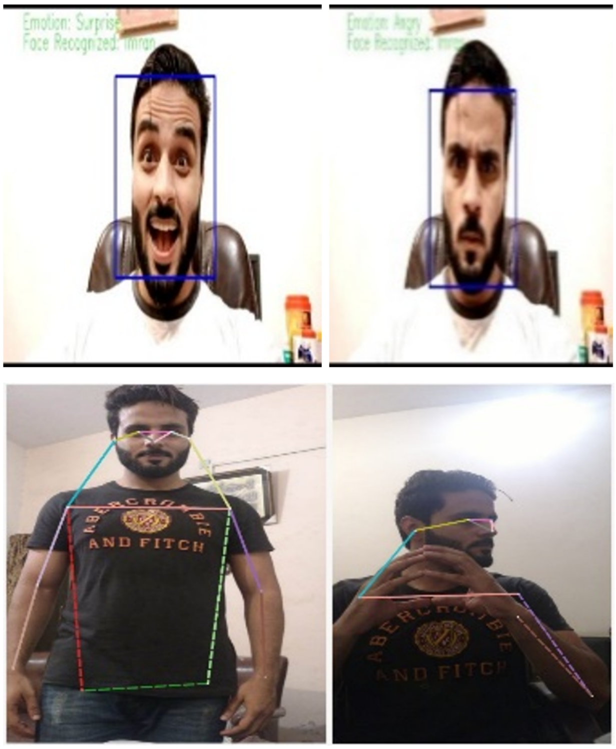

Figure 8. Demonstration of the system's multi-modal analysis capabilities. The left panels show successful recognition of "Surprise" and "Angry" expressions. The right panels show robust body pose estimation from both frontal and side views.

Figure 8 demonstrates the system's multi-modal analysis capabilities, showing successful expression recognition and robust pose estimation from various viewpoints.

Integrated Model Performance: Our final integrated system successfully combines all three models to achieve real-time analysis. The system processes a single frame in approximately 0.038 seconds, achieving a speed of nearly 30 FPS. This allows for fluid, simultaneous monitoring of student identity, posture, and emotional expression. The final performance of the integrated model is:

- Face Identification: 99.63% accuracy (frontal profile).
- On the YouTube Faces DB dataset, the proposed FaceNet-based model achieved an accuracy of 95.12%.
- Expression Monitoring: 80% accuracy (5 emotions)
- Pose Estimation: Accurately predicts with a >0.8 index on crowd datasets.

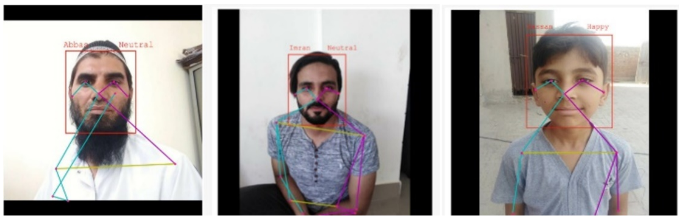

Figure 9. Final results of the integrated model on different individuals. The system correctly identifies the person's name, emotional state (e.g., Happy, Neutral), and overlays the estimated body and facial key points in real-time.

As shown in Figure 9, the integrated system provides comprehensive real-time analysis, correctly identifying individuals, their emotional states, and overlaying pose estimations simultaneously.

6 Conclusion

In this paper, we presented a novel system that integrates three distinct deep learning models for face recognition, expression monitoring, and pose estimation into a unified framework. Despite the challenge of combining these complex models efficiently, the system has demonstrated promising results in real-time monitoring, effectively tracking student behaviors, and storing data for educator review. Notably, the accuracy of the expression recognition module was improved, highlighting the benefits of an integrated approach to behavior analysis in educational settings. While some limitations remain, this work represents a significant step toward creating smarter, more responsive learning environments. Looking ahead, we plan to address current limitations by incorporating additional sensors to capture movement magnitude and environmental factors, as well as extending the system with analytical models for eye-gaze estimation, head pose tracking, and individual behavior profiling. Furthermore, results will be displayed on a real-time web-based dashboard for teachers and staff, and the integration of IoT devices will be explored to further enhance the learning process.

Data Availability Statement

All data and materials supporting the findings of this study are publicly available at: https://github.com/ImranKhalid-IT/test-dada (accessed on 13 December 2025).

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Alghaili, M., Li, Z., & Ali, H. A. (2020). Facefilter: face identification with deep learning and filter algorithm. *Scientific Programming, 2020*(1), 7846264. [CrossRef]

[2] Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., & Schiele, B. (2018, June). PoseTrack: A Benchmark for Human Pose Estimation and Tracking. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5167-5176). IEEE. [CrossRef]

[3] Baepler, P., & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning, 4*(2), 17. [CrossRef]

[4] Barros, P., Weber, C., & Wermter, S. (2015). Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)* (pp. 582–587). IEEE. [CrossRef]

[5] Bi, M., Qian, Y., & Yu, K. (2015). Very deep convolutional neural networks for LVCSR. In *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*. [CrossRef]

[6] Bilkhu, M. S., Gupta, S., & Srivastava, V. K. (2018). Emotion classification from facial expressions using cascaded regression trees and SVM. In *Computational Intelligence: Theories, Applications and Future Directions-Volume II: ICCI-2017* (pp. 585-594). Singapore: Springer Singapore. [CrossRef]

[7] Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence, 43*(1), 172-186. [CrossRef]

[8] Chandrasekhar, V., Lin, J., Liao, Q., Morere, O., Veillard, A., Duan, L., & Poggio, T. (2017, April). Compression of deep neural networks for image instance retrieval. In *2017 Data Compression Conference (DCC)* (pp. 300-309). IEEE. [CrossRef]

[9] Chang, D. F., & Sue, S. (2003). The effects of race and problem type on teachers' assessments of student behavior. *Journal of Consulting and Clinical Psychology, 71*(2), 235–242. [CrossRef]

[10] D'Souza, J. W., Jothi, S., & Chandrasekar, A. (2019, March). Automated attendance marking and management system by facial recognition using histogram. In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) (pp. 66-69). IEEE. [CrossRef]

[11] Dehghan, A., Ortiz, E. G., Shu, G., & Masood, S. Z. (2017). Dager: Deep age, gender and emotion recognition using convolutional neural network. *arXiv preprint arXiv:1702.04280*.

[12] Devadoss, S., & Foltz, J. (1996). Evaluation of factors influencing student class attendance and performance. *American Journal of Agricultural Economics, 78*(3), 499–507. [CrossRef]

[13] Dino, H. I., & Abdulrazzaq, M. B. (2019, April). Facial expression classification based on SVM, KNN and MLP classifiers. In *2019 International Conference on Advanced Science and Engineering (ICOASE)* (pp. 70-75). IEEE. [CrossRef]

[14] Doering, A., Iqbal, U., & Gall, J. (2018). Joint flow: Temporal flow fields for multi person tracking. *arXiv preprint arXiv:1805.04596*.

[15] Glass, A. L., & Kang, M. (2019). Dividing attention in the classroom reduces exam performance. *Educational Psychology, 39*(3), 395–408. [CrossRef]

[16] Grace-Martin, M., & Gay, G. (2001). Web browsing, mobile computing and academic performance. *Journal of Educational Technology & Society, 4*(3), 95–107.

[17] Grasha, A. (2002). Teaching with style: a practical guide to enhancing learning by understanding teaching and learning styles [electronic version]. *San Bernadino CA: Alliance Publishers Retrieved July, 25*, 2008.

[18] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2961–2969). [CrossRef]

[19] Huang, G. B., Mattar, M., Berg, T., & Learned-Miller,

E. (2008). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

[20] Hwang, J., Park, J., Lee, J., & So Kweon, I. (2019). Pose estimator and tracker using temporal flow maps for limbs. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. [CrossRef]

[21] Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). pmlr.

[22] Jiang, R., Ho, A. T., Cheheb, I., Al-Maadeed, N., Al-Maadeed, S., & Bouridane, A. (2017). Emotion recognition from scrambled facial images via many graph embedding. *Pattern Recognition, 67*, 245-251. [CrossRef]

[23] Liu, X., Jonschkowski, R., Angelova, A., & Konolige, K. (2020, June). KeyPose: Multi-View 3D Labeling and Keypoint Estimation for Transparent Objects. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11599-11607). IEEE. [CrossRef]

[24] Ma, N., Zhang, X., Zheng, H. T., & Sun, J. (2018, September). ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In *European Conference on Computer Vision* (pp. 122-138). [CrossRef]

[25] Diallo, M. A., Ouamri, M. A., Khalid, M. I., Komal, A., Diallo, E. M., & Adam, A. B. (2025, June). Multi-Modal Deep Reinforcement Learning Framework for Interference-Limited RIS-Assisted UAV Wireless Network. In *2025 12th IFIP International Conference on New Technologies, Mobility and Security (NTMS)* (pp. 80-85). IEEE. [CrossRef]

[26] Jasim, A. H., & Hoomod, H. K. (2025, April). Intelligent student behavior recognition system for the classroom environment using hybrid deep learning. In *AIP Conference Proceedings* (Vol. 3282, No. 1, p. 030011). AIP Publishing LLC. [CrossRef]

[27] Geng, F., John, A. D., & Chinnappan, C. V. (2023). Analysis of the teaching quality using novel deep learning-based intelligent classroom teaching framework. *Progress in Artificial Intelligence, 12*(2), 147–162. [CrossRef]

[28] Suryanarayana, K. S., Kandi, V. P., Pavani, G., Rao, A. S., Rout, S., & Krishna, T. S. R. (2024). Artificial intelligence enhanced digital learning for the sustainability of education management system. *The Journal of High Technology Management Research, 35*(2), 100495. [CrossRef]

[29] Sudrajat, S. (2025). The smart classroom: Exploring the impact of IoT on teaching dynamics using SVM algorithm. *Indonesian Journal of Multidisciplinary Science, 4*(5), 415–427. [CrossRef]

[30] Zhang, Z. (2024, September). Intelligent Learning Environment: Intelligent Monitoring System for Intelligent Classroom Management. In *Proceedings of the 2024 9th International Conference on Cyber Security and Information Engineering* (pp. 535-541). [CrossRef]

Muhammad Imran received the B.S. degree in Information Technology from the University of Education, Pakistan, in 2018, and the M.S. degree in Computer Science from the University of Okara, Pakistan, in 2022. He is currently pursuing a Ph.D. degree in Computer Science at the University of Posts and Telecommunications, China. His research during the B.S. focused on face recognition, while his M.S. research was in the field of image processing. He is currently conducting research in Explainable Artificial Intelligence (XAI). In addition, he has undertaken research and coursework in networking and cybersecurity. (Email: L202310008@stu.cqupt.edu.cn, imran.khalid292@gmail.com)

Asma Komal received the B.S. degree in Information Technology from the University of Education, Pakistan, in 2018, and the M.S. degree in Computer Science from the University of Okara, Pakistan, in 2022. She is currently pursuing a Ph.D. degree in Computer Science at the University of Posts and Telecommunications, China. Her research during the B.S. focused on machine learning and deep learning applications, while her M.S. research was in the field of machine learning and deep learning with emerging techniques. She is currently conducting research in Explainable Artificial Intelligence (XAI), large language models, and cyber attacks and prevention. In addition, she has undertaken research and coursework in image enhancement, cybersecurity, and networking. (Email: L202410011@stu.cqupt.edu.cn)

Usman J Wushishi is a PhD candidate at Chongqing University of Posts and Telecommunications, specializing in IIoT security with a focus on cybersecurity deception. He earned his Master's in Management Information Systems from CIU Nicosia. His broader research interests include artificial intelligence, deep learning, game theory, and reinforcement learning. (Email: l202310011@stu.cqupt.edu.cn)

Altaf Hussain received his Bachelor Degree in Computer Science from University of Peshawar, Pakistan in 2013 & Master Degree in Computer Science from The University of Agriculture Peshawar, Pakistan in 2017, respectively. He has more than 6 years of teaching & research experience. He worked at The University of Agriculture Peshawar in Faculty of IT as Researcher from 2017 to 2019. He has supervised many bachelor's and master's degree level students and helped them with their final year projects and research. During his Master study, he has completed his research in drone communication systems. Currently, he is a PhD Scholar in School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China, specializing in Underwater Acoustic Sensor Networks with focus on energy efficiency, channel optimization, localization and dynamic scheduling integrating 6G and quantum computing. He has served as a Lecturer in Computer Science Department in Government Degree College Lal Qilla Dir Lower, KPK Pakistan from 2020 to 2021. He has worked as Research Assistant with the Department of Accounting and Information Systems, College of Business and Economics, Qatar University, Doha, Qatar. He also worked as IT clerk in the Court of District and Session Judge Timergara Dir Lower from 2022 to 2023. He has published several notable research papers. He has reviewed many articles and is serving as reviewer for Cluster Computing, Computing, Cybernetics and Systems, Journal of Cloud Computing, Knowledge and Information Systems, Peer-to-Peer Networking and Applications, SN Applied Sciences, The Imaging Science Journal, The Journal of Supercomputing, Transactions on Emerging Telecommunications Technologies, Wireless Personal Communications, Frontiers in Big Data, CMC-Computers, Materials & Continua, and Bulletin of Electrical Engineering and Informatics (BEEI). His Research interest includes Artificial Intelligence, Machine Learning, Deep Learning, Gesture Detection, Wireless Networks, Underwater Sensor Networks, and Unmanned Drone Communication Systems. (Email: altafkfm74@gmail.com)

Nasir Hussain is a PhD candidate at Chongqing University of Posts and Telecommunications, specializing in IIoT security with a focus on cybersecurity Intrusion Detection. He earned his Master's in Computer Science from NUAA China. His broader research interests include IIoT security, artificial intelligence, cloud computing, GIS, Machine learning, and reinforcement learning. (Email: L202310005@stu.cqupt.edu.cn)

Sayed Akif Hussain received his Bachelor of Science degree in Accounting and Finance from the University of Wah, Pakistan. He is currently pursuing his Master's degree at the School of Economics and Management Sciences, where his research focuses on Financial Engineering, with a particular interest in the application of advanced computational techniques, machine learning, and data-driven models to financial markets, Corporate Finance, Financial Statement Analysis and investment strategies. His academic and research background combines expertise in finance with modern analytical methods, aiming to bridge theory and practice in innovative financial solutions. (Email: L202320034@stu.cqupt.edu.cn, syedakifhussain110@gmail.com)