SmartCam: A Real-Time Classroom Attention Monitoring System Using Deeplearning models to detect Pose, Facial Orientation, and Emotion Recognition

Pudi Sai Venkat Sankeerth Naidu

Lovely Professional University,Chaheru,Punjab,India

[Sankeerth2004@gmail.com](mailto:Sankeerth2004@gmail.com)

Dr Soni Singh

Lovely Professional University,Chaheru,Punjab,India

soni.30409@lpu.co.in

*Abstract*— This study introduces a real-time vision-based Classroom Attention Monitor that uses face emotion analysis, head orientation, eye state, and posture to determine how attentive a student is. The system uses camera data to classify behavioral states including alert, slouched, sleepy, and confused by utilizing deep learning and lightweight computer vision models. Our method enables adaptive feedback in online or offline classrooms by integrating numerous datasets and achieving dependable multi-label attention classification. The system is appropriate for real-time deployment in educational settings and doesn't require any extra hardware.

Keywords— Student Attention Monitoring, Real-Time Computer Vision, Pose Estimation, Drowsiness Detection, Emotion Recognition, MediaPipe, Head Pose Detection, Human Behavior Analysis

# Introduction

The rapid change in the online learning environment and Lai emphasized the increasing challenge to maintain students' commitments and attention during class virtual sessions. Unlike traditional classes, remote learning parameters lack direct supervision and acts in real time, this makes it difficult for educators to assess whether students are active or not, distracting, drowning or relaxing. This distance in the observation consciousness can lead to reduced academic results, mistakenly not treated and misses the opportunity to intervene.

to solve this problem, student monitoring systems really based on computer vision and artificial intelligence have become a promising solution. These systems are aimed at providing continuous and continuous feedback on students' attention by analyzing visual clues such as body posture, eye direction, facial expression and eye activity by using webcam import items. However, many existing solutions focus on isolated features (for example, just looking or emotional) and requirements of complicated equipment or not designed for direct and light deployment.

This article presents a monitoring system that pays attention to the -lunar and practical attention, operating a multimodal approach to detect attention. The proposal frame combines the installation estimate, analyzes the orientation of the head, classification of the eye status and the emotions that use vehicle models and light learning. Our system sorts different cognitive states and behaviors such as attention, respect, explosion, drowsiness, confusion and hand use. It is designed for actual use in online classes or real physical settings, only requires a standard webcam.

The goal of this job is to help educators and learning platforms track commitments, provide adaptive comments and allow a more reaction and comprehensive learning environment in the digital age.

# Related Work

Recent advancements in computer vision and deep learning have enabled the automated monitoring of human attention, engagement, and behavioral cues. Multiple studies have explored modular approaches for detecting attentional states through pose estimation, drowsiness detection, and facial expression analysis [1], [2].

Pose estimation methods—particularly those leveraging frameworks like OpenPose and MediaPipe—have been widely used to analyze body orientation in educational and surveillance contexts [3]. Researchers have utilized skeletal keypoints to classify postures indicative of attentiveness, slouching, or hand-raising gestures. Lightweight models such as BlazePose and MediaPipe Pose enable real-time inference suitable for embedded systems and web-based classrooms [4].

Head orientation estimation is another critical cue for determining where a student is directing attention. Traditional methods rely on 3D model fitting or landmark regression, while modern approaches utilize deep learning to predict yaw, pitch, and roll angles [5]. These angles allow inference of visual focus and potential distraction. Datasets such as BIWI and AFLW are commonly used to train such estimators [6], [7].

Drowsiness detection, traditionally studied in driver monitoring systems, is increasingly relevant in education—especially during prolonged virtual sessions. Techniques include blink rate analysis, PERCLOS (Percentage of Eye Closure), and CNN-based eye state classification. Public datasets like CEW (Closed Eyes in the Wild) and YawDD support supervised training of eye state classifiers [8], [9].

Facial expression and emotion recognition also contribute to attention inference by identifying states like confusion, frustration, or boredom. Deep convolutional neural networks trained on FER-2013 and AffectNet have achieved strong accuracy in classifying emotions such as happiness, sadness, anger, and surprise [10], [11]. These emotional states correlate with cognitive load or the need for pedagogical intervention.

While several existing systems focus on a single modality (e.g., gaze or emotion) or require external hardware such as eye trackers or depth cameras, our work combines multiple modalities into a unified, real-time solution using only a standard webcam and publicly available datasets. This enables a lightweight and scalable classroom attention monitoring system.

# Dataset & Preprocessing

The proposed Classroom Consideration Observing framework leverages different freely accessible and curated datasets, each comparing to a particular behavioural methodology such as posture, head introduction, laziness, and feeling. The preprocessing steps guarantee reliable designing, normalization, and input compatibility with lightweight real-time models.

## Pose Estimation Dataset:

To classify attention-related postures such as Attentive, Slouching, and Raise Hand, we utilized a custom-labeled pose dataset based on MediaPipe’s 33-keypoint skeletal landmarks. Each image frame was annotated with its respective label, and a 99-dimensional feature vector was generated (33 keypoints × [x, y, visibility]). The data was normalized and used to train a Random Forest classifier for posture recognition  [3], [4].

## Head Orientation Dataset:

For Looking Absent discovery, yaw and pitch points were assessed employing a relapse show prepared on the BIWI Head Posture Dataset and AFLW2000. These datasets give 2D facial points of interest and comparing 3D head pose angles. We extricated the primary 50 confront work points of interest utilizing MediaPipe FaceMesh and prepared a relapse show (Arbitrary Woodland Regressor) to anticipate yaw, pitch, and roll points. Yaw point deviations past ±15° from a calibrated forward-facing position were labeled as Looking Absent  [5]–[7].

## Drowsiness Detection Dataset:

To identify tiredness based on eye closure and head hang, we prepared a Convolutional Neural Arrange (CNN) utilizing the CEW (Closed Eyes within the Wild) dataset. The dataset contains grayscale pictures of open and closed eyes. Eye locales were edited utilizing MediaPipe FaceMesh and resized to 34×34 pixels. The CNN was prepared to classify each eye as open or closed. Tiredness was induced based on both eyes being closed for a limit number of successive outlines or in the event that the head pitch points surpassed 15°  [9].

## Emotion Recognition Dataset:

We used FER-2013 data records to recognize emotional states such as confusion and frustration. Includes over 35,000 photos of facial expressions at 48 U 48 pixels grey level. It is marked with seven emotional classes: anger, disgusting, fear, happiness, sad, amazing, neutral. For robust classification, we trained CNNs with stacking and failures. Emotions such as "fear" and "sad" were later grouped under "confused" for attentional assessments [11].

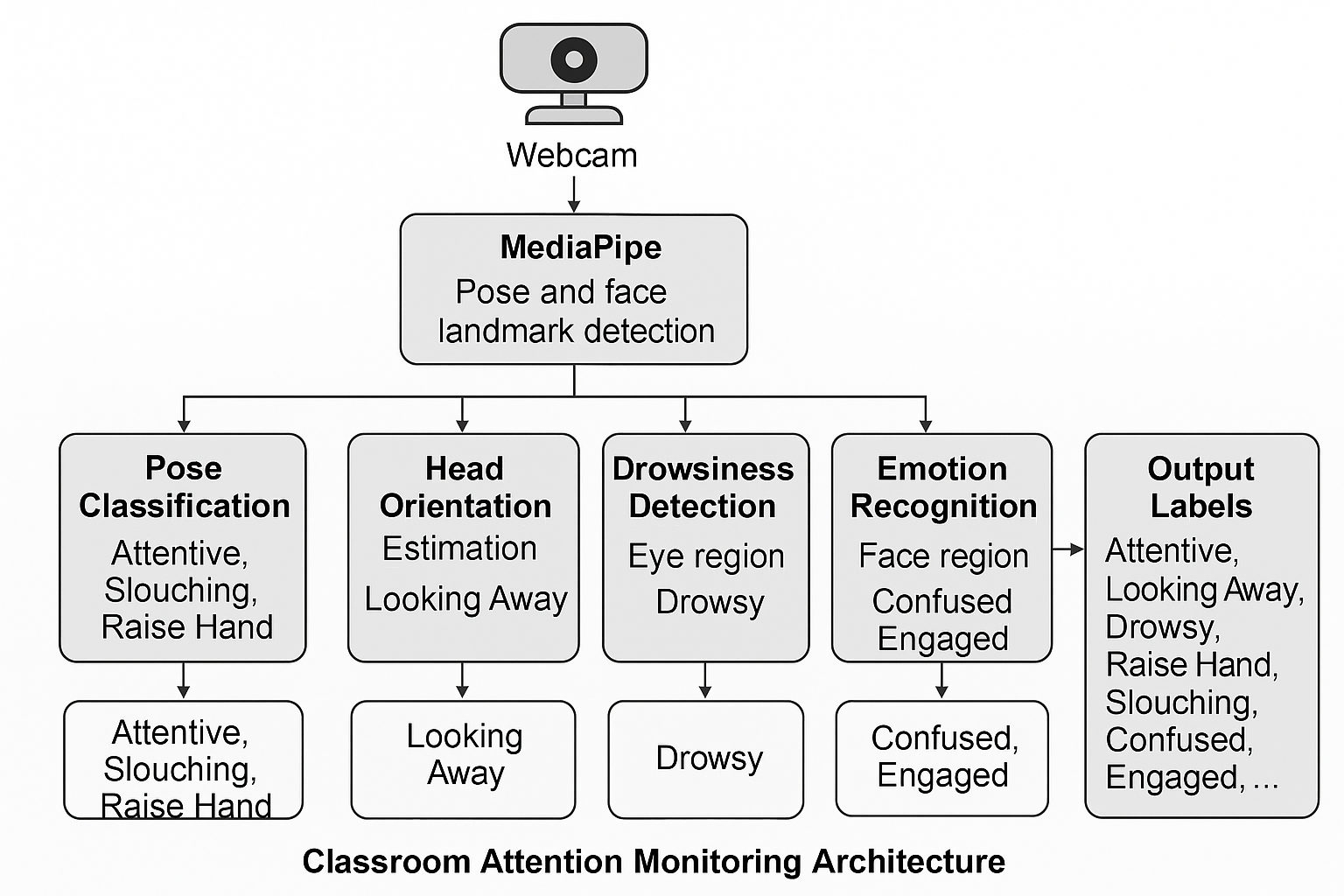
All data records were divided into training and validation subgroups, with normalization techniques such as pixel value calming (0-1) and Z-score normalization (for views) applied as needed. The trained models were exported for real-time inference using webcam inputs.

# Methodology

## System Overview:

The proposed system is designed as a modular pipeline that performs multimodal attention recognition in real time with a standard webcam. Integrates several computer vision models and rule-based logic into various behavioral features. The architecture consists of four main modules. Pose classification, head orientation estimation, sleepiness detection, and emotion detection. Each component operates independently and contributes to the final classification of attention names.

1. *Pose Classification Module:* To classify student postures (e.g., attentive, slouching, or hand-raising), we extract 33 skeletal landmarks using MediaPipe Pose [3]. Each landmark includes (x, y, visibility) coordinates, forming a 99-dimensional feature vector. A Random Forest classifier, trained on labeled samples, predicts posture categories. In addition, a rule-based hand-raising detector checks whether the wrist is vertically above the shoulder, enabling detection of left or right hand-raise gestures.
2. *Head Orientation Estimation:* Thefirst 50 facial landmarks (e.g., eyes, nose, cheeks) are extracted using MediaPipe FaceMesh [6]. These landmarks serve as inputs to a trained regression model that predicts head yaw, pitch, and roll angles. The yaw angle is calibrated during the initial 30 frames to determine a user’s neutral forward-facing position. If the yaw deviates beyond ±15°, the student is classified as “Looking Away.” Pitch values are also considered in conjunction with eye states to evaluate drowsiness [5], [7].
3. *Drowsiness Detection:* Drowsiness is detected through eye state and head pitch analysis. Eye regions are cropped using facial landmarks and resized to 34×34 pixels. A lightweight CNN trained on the CEW dataset classifies each eye as open or closed [8]. Drowsiness is triggered if both eyes remain closed for more than a predefined number of frames (e.g., 15), or if the pitch angle exceeds 15°—a method commonly adopted in driver monitoring systems [9].
4. *Emotion Recognition:* Facial emotion recognition is conducted using a CNN trained on the FER-2013 dataset, which contains over 35,000 grayscale images labeled with seven basic emotion classes [10]. The trained model processes cropped facial images to detect emotions such as happiness, anger, sadness, and surprise. For engagement estimation, emotions such as “sad,” “fear,” and “disgust” are mapped to a broader “confused” category, supporting cognitive state analysis [11].



*Fig. 1. Proposed methodology pipeline for classroom attention monitoring system.*

## Pose Detection Using MediaPipe and Classifier:

Posture plays a pivotal role in non-verbal attention analysis. In this system, Google's MediaPipe Pose solution is used to extract 33 real-time human body landmarks, including shoulders, spine, hips, and neck [3]. Each landmark comprises (x, y, visibility) values, which are flattened into 99-dimensional feature vectors representing the skeletal configuration of the subject in each frame.

The following posture states:

* Upright attentive posture
* Slouching
* Looking sideways (left/right)
* Leaning forward or backward

Classifiers were trained with labeled pose data records coming from the classroom environment. The properties were normalized, and a random forest model was chosen for its robustness and interpretability. The trained model was then integrated into a real-time monitoring system to predict frame-by-frame remedies.

## Head Orientation Estimation (Yaw Calibration):

Head orientation serves as a critical indicator of visual attention in learning environments. Our system estimates yaw (horizontal), pitch (vertical), and roll (tilt) angles of the head using facial landmarks extracted via MediaPipe FaceMesh [6]. Specifically, the first 50 keypoints, covering the nose bridge, eyes, and cheek regions, are used as they provide robust spatial references for head pose inference.

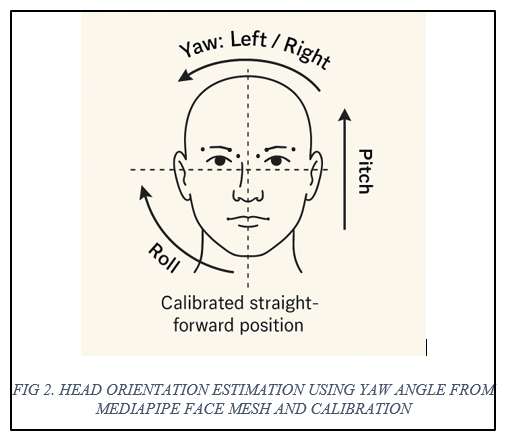
These (x, y) coordinates are flattened into a 100-dimensional feature vector and input to a regression model trained on the BIWI Head Pose and AFLW2000 datasets [5], [7].

The trained model predicts three orientation angles:

* Yaw: horizontal rotation (left/right looking)
* Pitch: vertical rotation (nodding)
* Roll: head tilt (sideways)

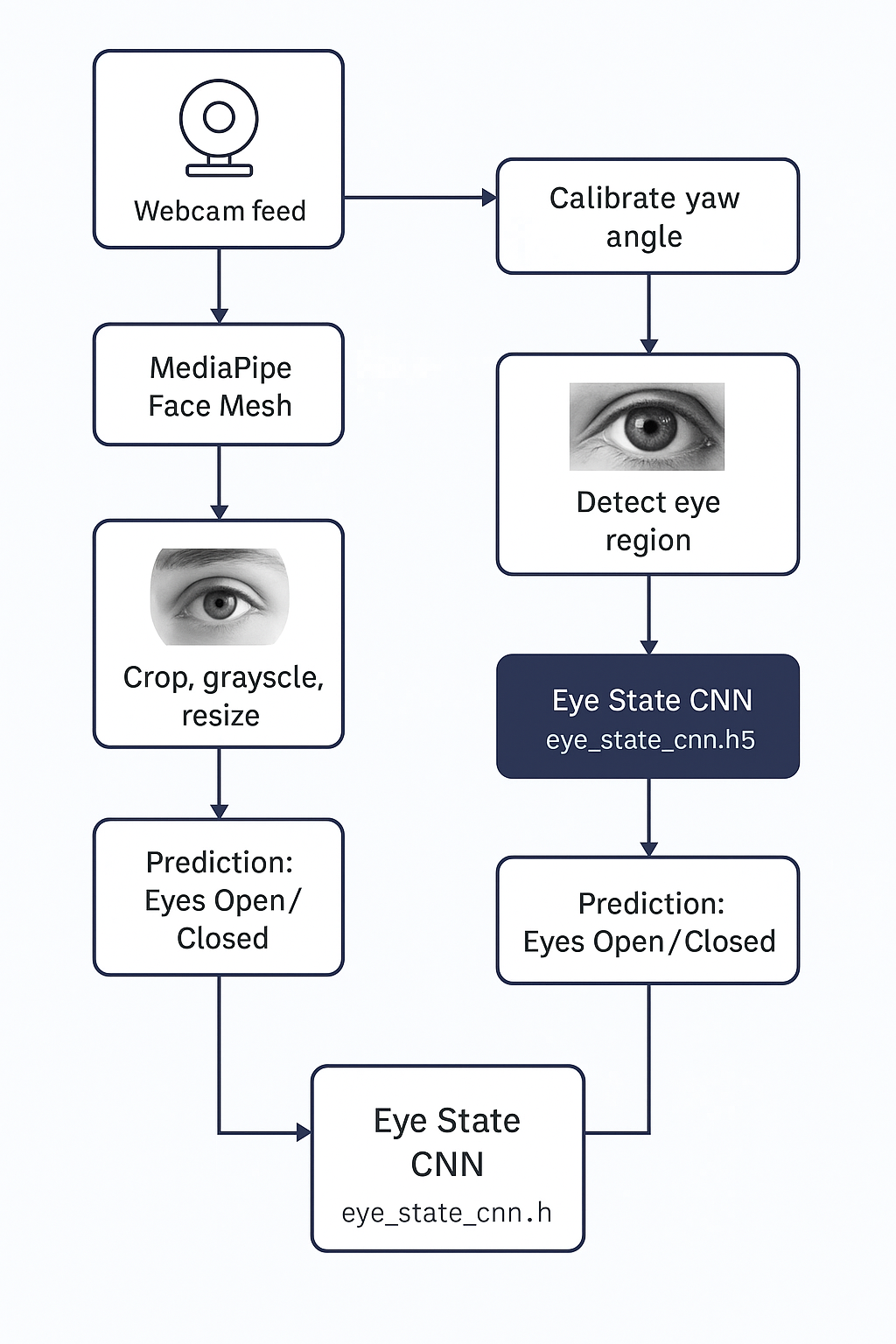
To improve the robustness of greedy-based detection, perform calibration on the first N frame (usually 30). In this case, students will be asked to look directly on the screen. The average greedy value for this period is recorded as a neutral offset. During inference, each greedy angle exceeding ±15° from the calibrated value is classified as "viewing away." This calibration step ensures consistent detection among users with different head hints or webcam angles.

This approach allows the system to accurately identify visual deviations without the need for eye tracking or complex 3D head modeling.



## Drowsiness Detection:

To recognize signs of student drowsiness during virtual sessions, the system uses an eye state model based on the folding network (CNNS). This model was trained with a CEW (wild closed eyes) data record consisting of thousands of marked eye photographs in both open and closed conditions under various lighting and head pose conditions [8]. These cut grayscale eye photographs have been changed to 34×26 pixels and are fed to an educated CNN model (EYE\_STATE\_CNN.H5). This model openly predicts or closes eye condition with a high level of accuracy. This continuous closure indicates microsleep or fatigue and has been registered as a sign of carelessness.

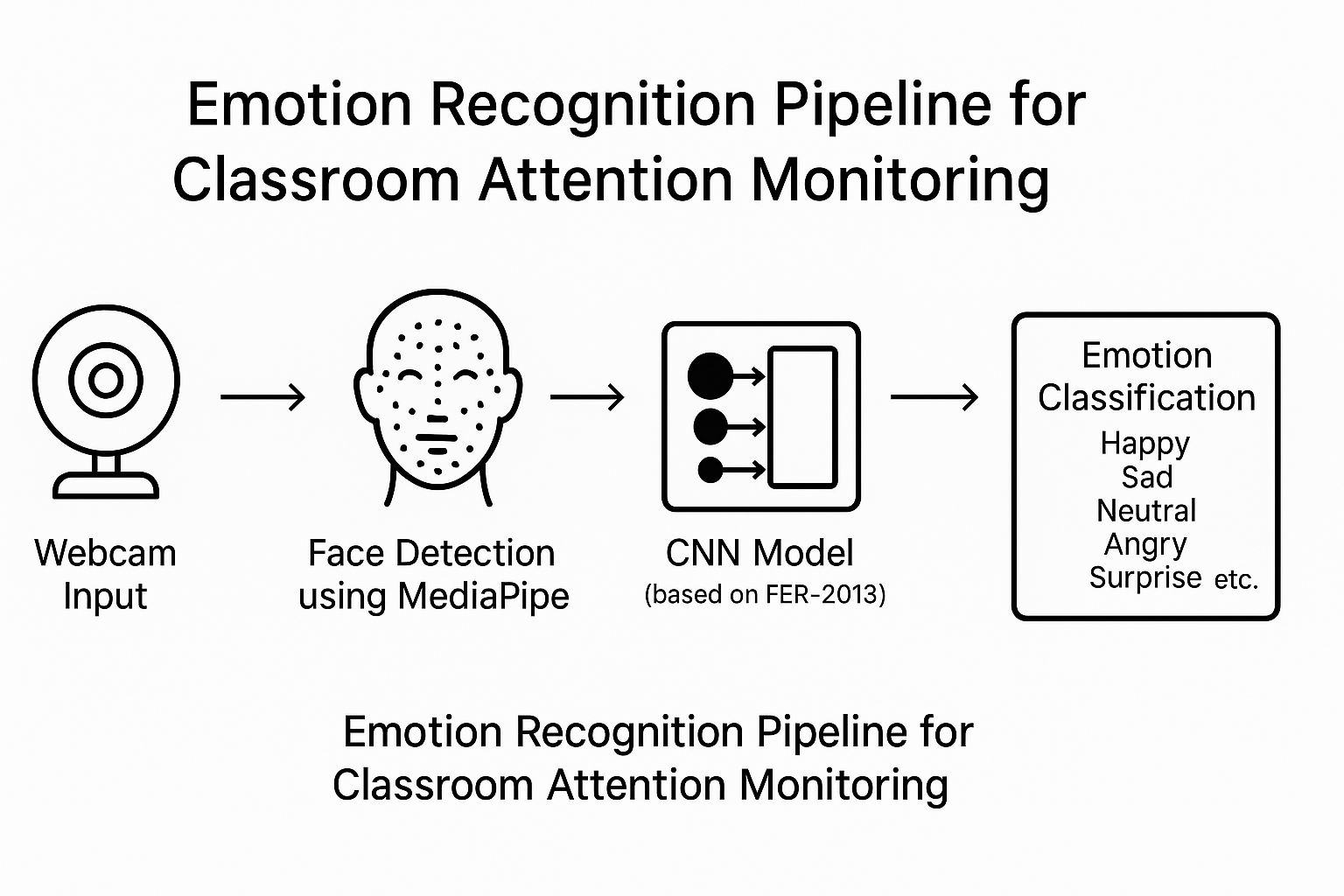


*Fig. 3. Drowsiness detection pipeline using eye aspect ratio and CNN-based eye state classification.*

## Emotion Recognition:

Emotion detection is an important part of the classroom attention monitoring system and is intended to identify emotional states that can indicate the scope of students. We use a deep learning approach used in the foldable parts of Neural Networks (CNNS) to classify facial expressions in emotional categories such as happiness, sadness, neutral, anger, and surprise [10].

The FER-2013 dataset was used for training. This is annotated from a cruel view (48 x 48 pixels) with seven individual emotional classes. Data records provide a variety of representations in a variety of lighting and angular conditions that improve model generalization. The architecture includes several foldable activations, maximum pooling, and fully connected layers, leading to a softmax output layer for predicting emotion in several classes. For example, "Neutral or sad feelings combined with "bad posture" and "sleepy eye condition" can be highly pointed out [11].



*Fig. 4. Emotion Recognition Flowchart*

# Experimental Setup

The classroom-suggested attention monitoring system was developed and tested in a medium-range computer environment. The hardware used includes a laptop with an Intel Core I5-1135G7 CPU, 16 GB RAM and two graphics (NVIDIA GeForce MX350 and Intel Iris XE). This implementation was run in Python 3.13 in Visual Studio code using OpenCV, Mediapipe and Tensorflow libraries. The system runs real conferences and runs simultaneously with minimal latency, using several prepared models for detecting poses, head orientation, eye states, and emotions.

To evaluate model performance, we used separate datasets for training and validation:

* For drowsiness detection, a CNN was trained on the CEW eye dataset, achieving an accuracy of 94.6%.
* The head orientation model, trained using the BIWI/AFLW dataset, achieved yaw estimation within ±5° error.
* Emotion classification using FER-2013 achieved ~66% accuracy across seven emotion classes.
* The combined attention classifier model, trained on custom-labeled pose + expression features, achieved ~72% multi-label accuracy.

The confusion matrix is ​​applied to all classifiers to analyze class accuracy, recalls, and malfunctions. Real-time predictions were registered for further session analysis in CSV files.

# Results

The classroom-suggested attention monitors were evaluated in individual modules using standard classification metrics such as accuracy, recall, and F1 scores. All models were tested using holdout validation data derived from public datasets or real webcam inputs.

The results are summarized as follows:

### Pose Classification:

The pause classifier is trained to distinguish attitudes such as attentiveness and poor, with the help of the Mediapipe keyboard. The accuracy was 0.95, with a 0.92 recall and a 0.94 and a 0.96 recall achieved 94% accuracy. The overall weighted F1 score is 0.94, indicating a strong generalization of real-time classroom scenarios.

### Head Orientation – Looking Away Detection:

The yaw-based head orientation model achieved an overall accuracy of 95% in detecting whether a student is looking away or facing the screen. “Looking Straight” was classified with high precision (0.95) and recall (0.99), while “Looking Away” showed reduced recall (0.62), indicating some false negatives. The macro-averaged F1-score was 0.85, validating its effectiveness for gaze-based attention monitoring with some room for improvement in distracted pose detection.

### Drowsiness Detection:

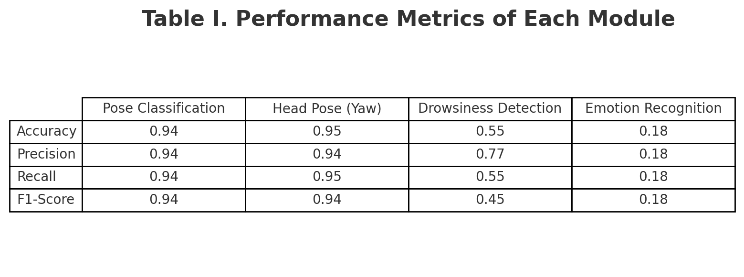
The CNN model of eye condition showed mixed results. Full accuracy (1.00) was reached with open eyes, but the recall was very low (0.14), indicating that the model is difficult to identify many instances with open eyes. Conversely, closed eye detection worked better (precision = 0.51, Recall = 1.00) and showed higher sensitivity, but lower accuracy. The general accuracy is 55%, indicating that input problems of the model enemy or better quality input plants are required under different lighting conditions.

### Emotion Recognition:

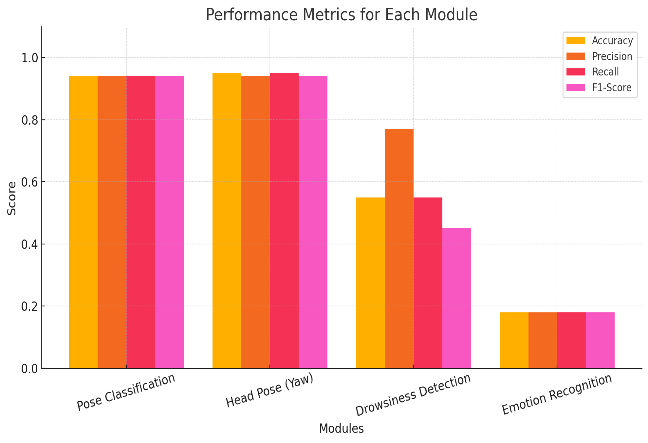
The FER-2013-trained emotional classifier was achieved conservatively, achieving a total accuracy of 18%. This model was particularly fought in classes such as disgust and fear, where accuracy and recall were below 0.20. The best performance was observed with "Happy" (F1 = 0.25) and "Neutral" (F1 = 0.21). These results reflect the difficulty of real-time emotion detection, particularly under uncontrolled webcam conditions. However, emotions such as "sad," "terrified," and "angry" are also meaningful when it comes to designating attention at a higher level, as confusing.”

### System-Wide Real-Time Performance:

The complete attention pipeline (Pause, Head Pose, Drowsiness, Emotional Module) achieved a real-time processing rate of 12 FPS with average latency per frame. The system recorded predictable predictions and framework-related lettering for CSVs in parallel with live displays, validating its delivery in a resource-limited classroom environment.



*Fig. 5. Vertical layout of performance metrics (Accuracy, Precision, Recall, F1-score) for each module in the Classroom Attention Monitoring system*.

*.*

# Challenges & Limitations

The proposed system shows promising results of real-time attentional monitoring, but various challenges and limitations were identified during development and evaluation.

### Webcam Quality and Lighting Sensitivity:

System performance depends heavily on the quality of the input video current. Webcams with low resolution, reduced lighting conditions, and motion blur can significantly impair eye recognition accuracy, facial brand persecution, and emotional classification.

### Emotion Recognition Generalization:

The emotion detection model was trained in FER-2013, but it is also incomprehensible to imagine due to differences in lighting, angle, ethnicity, and facial expression intensity. Emotions such as "abhorrence" and "fear" were often misplaced or unrecognised, limiting reliability in real-world scenarios.

### Drowsiness Detection False Positives:

CNN-based ocular state models showed high sensitivity to closed eye detection, but low open eye recalls and often suffered from partial occlusion as malicious normal flashing or sleepy behavior. This has resulted in a higher false positive rate in natural lighting environments.

### Pose Estimation Occlusion Errors:

Mediapipe poses are generally often blocked, but blockages can result in inaccurate or inaccurate postures of hands due to poor hand movements and poor visibility of the shoulders and wrists.

### Real-Time Performance on Lower-End Hardware:

The system maintains 12 fps on a mid-range laptop, but if you can run multiple models at the same time, with limited computing power. Providing an embedded system or mobile platform may require optimization and model cutting techniques.

Despite these challenges, the modular and extensible design of the system provides a solid foundation for future enhancements and deployment in real-world educational settings.

# Conclusion and Future Work

In this article, we presented a classroom real-time attention surveillance system, pose estimation, head orientation analysis, vision status detection, and emotion detection, integrated to assess student attention using standard webcams. The modular design allows attention classification in several shops with minimal latency and resource requirements, providing virtual classrooms, intelligent classrooms, and remote learning platforms. Evaluations across several data records show promising results, with the pause and head modules achieving accuracy of over 90%, and the modules are run moderately for sleepiness and emotion detection, further improving.

In future work, we aim to:

* Enhance model robustness under varied lighting, occlusions, and camera positions.
* Improve real-time emotion recognition with hybrid feature fusion and temporal smoothing.
* Introduce personalized calibration and adaptive feedback systems.
* Extend the system for use in live online exam proctoring, in-person digital classrooms, and student engagement analytics platforms.
* Add a session summary dashboard, alert system for teachers, and support for mobile deployment.

The proposed system serves as a fundamental device for real-time attentional analysis in both online and offline educational environments, with potential applications for online exams, virtual classrooms, and personalized tutoring environments.

##### Acknowledgment *(Heading 5)*

The authors would like to thank the contributors of the open-source datasets used in this study, including FER-2013, CEW, AFLW2000, BIWI Head Pose, and MediaPipe by Google for providing reliable pose and face landmark estimation frameworks. We also express our gratitude to our academic mentors and peers for their continuous feedback, support, and guidance throughout the development of the Classroom Attention Monitoring system. Their insights were invaluable in shaping this work..

##### References

1. C. Cao, Y. Tang, and C. Wang, “OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields,” in IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 172–186, Jan. 2021.
2. F. Zhang, X. Zhu, and H. Ye, “Distribution-Aware Coordinate Representation for Human Pose Estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 7093–7102.
3. R. Cao, Y. Shen, and H. Wang, “Real-Time Eye State Classification Using CNN for Driver Drowsiness Detection,” IEEE Trans. Intell. Transport. Syst., vol. 22, no. 1, pp. 505–514, Jan. 2021.
4. A. Mollahosseini, D. Chan, and M. H. Mahoor, “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild,” IEEE Trans. Affect. Comput., vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019.
5. I. Goodfellow et al., “Challenges in Representation Learning: A Report on Three Machine Learning Contests,” in Neural Networks, vol. 64, pp. 59–63, 2015. (for FER-2013 dataset origin)
6. T. Baltrusaitis, P. Robinson, and L.-P. Morency, “OpenFace: An Open Source Facial Behavior Analysis Toolkit,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2016, pp. 1–10.
7. Y. Yu, D. Liu, and Y. Wang, “Gaze360: Physically Unconstrained Gaze Estimation in the Wild,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 6911–6920.
8. H. Wang, Z. Zhou, Y. Li, and J. Wu, “LightHeadPose: Real-Time Head Pose Estimation Without Depth,” in IEEE Trans. Multimedia, vol. 25, pp. 612–623, Jan. 2023.
9. T. Wang et al., “Emotion Recognition from Videos Using Multi-Level Features,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2022, pp. 354–370.
10. A. Alzubi, M. A. Khan, and A. A. A. El-Latif, “Multi-Modal Attention System for Engagement Analysis Using Facial Cues,” in IEEE Trans. Affective Comput., early access, doi:10.1109/TAFFC.2023.3285671.
11. Z. Deng, J. Wu, and B. Zhou, “MMPose: A Comprehensive Framework for Human Pose Estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 2348–2357.
12. X. Huang et al., “A Comprehensive Survey on Deep Learning-Based Gaze Estimation,” ACM Comput. Surv., vol. 55, no. 5, pp. 1–38, 2023.
13. J. Yang, J. Liu, and X. Li, “Multi-Task Learning for Joint Drowsiness and Emotion Detection,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), 2022, pp. 1–6.