BILINGUAL PUBLISHING GROUP

**Journal of International Education and Practice**

https://journals.bilpubgroup.com/index.php/jiep

ARTICLE

# Detecting Student Inattention Using Deep Learning and Behavioral Analysis

***Fatima Zedan, Rana R. Jabali, Jamal Raiyn*** *

*Computer Science Department, Al-Qasemi Academic College of Education, Baqa-El-Gharbia 30100, Israel*

## ABSTRACT

Student inattention in classrooms negatively impacts learning outcomes and academic performance, posing a significant challenge for educators. Traditional methods of monitoring engagement rely on subjective teacher observations, which can be inconsistent, labor-intensive, and prone to bias. To address these limitations, this paper presents an AI-driven framework that uses deep learning and behavioral analysis to detect student inattention in real time. The proposed system integrates computer vision techniques including facial expression recognition, posture analysis, head pose estimation, and eye-gaze analysis, employing convolutional neural networks (CNNs) to extract spatial features and recurrent neural networks (RNNs) to model temporal patterns. The framework was evaluated using annotated classroom video data collected from real teaching sessions, capturing natural student behavior under typical classroom conditions. Experimental results demonstrate that the proposed approach achieves high accuracy in distinguishing attentive from inattentive states, outperforming traditional machine learning baselines while maintaining real-time performance. Beyond detection, the system provides actionable insights for educators by highlighting patterns of disengagement across time and students. By combining CNN-based spatial analysis with RNN-based temporal modeling, the framework offers an objective, scalable, and practical solution for monitoring classroom engagement, enabling timely interventions, personalized instruction, and improved learning outcomes.

***Keywords:*** Deep Learning; Smart Classroom; Active Learning; Intelligent Teacher Assistant

*CORRESPONDING AUTHOR:

Jamal Raiyn, Computer Science Department, Al-Qasemi Academic College of Education, Baqa-El-Gharbia 30100, Israel; Email: raiyn@qsm.ac.il

ARTICLE INFO

Received: 1 November 2025 | Revised: 23 December 2025 | Accepted: 30 December 2025 | Published Online: 7 January 2026
DOI: https://doi.org/10.30564/jiep.v9i1.12512

CITATION

Zedan, F., Jabali, R.R., Raiyn, J., 2026. Detecting Student Inattention Using Deep Learning and Behavioral Analysis. Journal of International Education and Practice. 9(1): 1–17. DOI: https://doi.org/10.30564/jiep.v9i1.12512

# 1. Introduction

Student attention is a crucial factor in effective learning, significantly influencing academic performance, knowledge acquisition, and overall engagement [1]. However, maintaining consistent attention in classroom environments remains challenging due to a variety of internal and external distractions, as well as fluctuating levels of student motivation. Teachers often struggle to identify inattentive students, especially in large or diverse classrooms, making it difficult to provide timely, personalized interventions that address learning gaps as they arise. Attention and concentration are distinct yet interrelated cognitive processes critical for successful learning. Concentration is a mental state in which all cognitive resources are directed toward a specific subject, enhancing the ability to process and retain information [2,3]. It is an acquired skill that allows individuals to absorb content fully and apply it meaningfully in new contexts. Attention, meanwhile, is a cognitive process involving the encoding of sensory and language inputs, maintaining them in working memory, and retrieving them from long-term memory [4]. Effective concentration ensures that incoming information is well received and integrated, supporting higher-order thinking and problem-solving [5]. Multiple factors affect students' ability to sustain attention in class. Internal factors include psychological states such as boredom or anxiety, which can often be mitigated through engaging and varied teaching strategies. External factors encompass the broader learning environment, including the school system, classroom layout, family background, and instructional methods. For example, research highlights the influence of environmental variables like lighting, temperature, noise levels, and seating arrangements on student behavior and focus [6]. Students seated in the front rows typically demonstrate higher attention levels due to increased proximity to the teacher and reduced distractions. Furthermore, classroom attendance is a vital predictor of attention and academic success; frequent absences disrupt learning continuity and weaken students' ability to concentrate during lessons.

Emerging technologies have provided new avenues for understanding and improving student attention. Studies have demonstrated that the presence of surveillance systems, such as security cameras, can subtly influence student behavior by encouraging self-regulation [7,8]. Advanced techniques including computer vision and machine learning have been applied to analyze classroom dynamics, tracking behavioral cues such as head movements, gaze direction, posture, and facial expressions to infer attention states. These approaches offer educators objective, real-time insights into student engagement, helping them tailor instruction to maximize learning outcomes [9,10].

Recent advances in deep learning present significant opportunities to automate and enhance the detection of inattention in classrooms [11]. Deep learning models [12,13], particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown remarkable success in extracting complex spatial and temporal patterns from visual and behavioral data [14–16]. For example, gaze estimation techniques based on CNNs can accurately predict where a student is looking, while head pose analysis can indicate distraction or disengagement. Some studies have proposed architectures incorporating data fusion approaches, combining multiple gaze datasets and augmentations to improve model robustness and generalization [17].

This paper proposes a deep learning–based approach that utilizes behavioral analysis to detect student inattention and improve the learning experience. The framework integrates computer vision techniques, including facial expression recognition, posture analysis, and eye-tracking to identify behavioral indicators of inattention in real time. Our system employs CNNs to capture spatial features of student behavior and RNNs to model temporal dynamics, enabling accurate classification of attentive versus inattentive states. By providing automated, objective, and scalable detection of student inattention, our approach aims to support educators in delivering timely interventions, fostering more engaging and effective learning environments.

This paper is organized as follows: Section 2 provides an overview of deep learning technology and its applications in education. Section 3 describes the methodology, and Section 4 introduces learning behavior in the smart classroom. Sections 5–7 discuss the results, conclude the discussion, and point out directions for future research.

# 2. Related Research

Research on student attention and distraction in classrooms has received increasing focus in recent years, reflecting its critical role in learning outcomes[18]. Numerous studies have examined behavioral patterns that signal distraction, particularly among elementary and secondary school students. Demographic factors such as age, gender, and peer interactions have been found to significantly influence distraction levels, with evidence suggesting that older students often display greater engagement in social activities like talking, which can detract from instructional focus.

Visual attention research emphasizes the role of eye gaze, head movements, and cognitive load in evaluating student focus and situational awareness. Tracking these cues enables the detection of attentional lapses and provides insights into how students process instructional content. For example, head rotation, downward gaze, or eye closure can indicate disengagement, drowsiness, or confusion[19].

Traditional approaches to assessing student engagement—including surveys, teacher logs, and direct observation—often suffer from subjectivity, limited temporal resolution, and the inability to scale in real time. These methods typically provide delayed or episodic snapshots of student behavior that may fail to capture subtle, dynamic shifts in attention. As a result, recent work has turned into automated, objective monitoring systems[17] that leverage machine learning and deep learning to evaluate student focus more continuously and accurately.

Advanced monitoring systems have integrated facial expression analysis, eye-tracking, and head pose estimation to assess attentional states[20,21]. Machine learning methods have been developed to recognize signs of drowsiness, distraction, or confusion by analyzing facial landmarks and behavioral cues. Deep learning has enabled more robust and precise analysis thanks to its capacity to model complex, nonlinear patterns in high-dimensional data. Wu[22] proposed a deep learning method using an improved YOLOv3 to detect students' abnormal behaviors in smart classrooms, achieving over 90% accuracy with low false reports and minimal delay.

More recently, newer YOLO versions such as YOLOv4, YOLOv5, YOLOv7, and YOLOv8 have been increasingly adopted for gaze-related tasks, including face detection, eye-region localization, and head pose estimation, due to their improved feature representation, lightweight architecture, and suitability for real-time deployment. YOLOv-based models are often employed as the first stage in gaze estimation pipelines, enabling reliable face and eye detection under varying lighting conditions, occlusions, and camera viewpoints. These characteristics make YOLOv particularly suitable for classroom environments, where multiple students must be tracked simultaneously.

Several studies combine CNN-based gaze estimation with temporal modeling. Recent studies have introduced deep learning models, especially Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for real-time attention evaluation[9]. These models have demonstrated strong performance in recognizing behavioral patterns such as gaze direction, facial expressions of confusion, and inattentive postures. However, challenges remain regarding generalizability and consistency across diverse classroom environments, where variations in lighting, student demographics, and camera angles can degrade model performance. Their system follows three main steps: face detection, yaw and pitch estimation, and gaze zone determination. For yaw estimation, it identifies the left and right facial borders and the face center using the ellipsoidal model. For pitch estimation, it extracts novel histogram-based features and applies Support Vector Regression (SVR).

Kanade et al.[23] proposed incorporating head posture into gaze estimation models, recognizing that users' heads move freely in real-world scenarios. Their system uses Euler angles (pitch, yaw, roll) to determine head orientation and relies on commercial head trackers to detect facial landmarks, particularly around the eyes. To maintain tracking robustness, the system minimizes dependency on head sensors by using historical pupil center coordinates from prior frames. They designed a hierarchy of small, efficient CNNs to precisely locate eye regions even when local tracking fails.

Yoo et al.[24] introduced a gaze behavior–based data processing method for visualizing abstract gaze data. Their approach categorizes raw gaze information using machine learning models originally designed for image classification, including CNNs like AlexNet and LeNet. They evaluated multiple fixation identification techniques—velocity-based (I-VT), dispersion-based (I-DT), density-based, and combined velocity-dispersion (I-VDT)—and assessed their outputs across visualization formats such as attention maps, scan paths, and abstract gaze movement representations. Their pipeline uses facial landmarks and face mesh detectors to identify regions of interest, extracting features such as eye aspect ratio, mouth aspect ratio, and head pose. These features are fed into classifiers like Random Forests, Sequential Neural Networks, and Linear SVMs for engagement prediction.

Vijaypriya and Uma[20] advanced appearance-based gaze estimation using CNNs to predict gaze angles from eye images and landmark coordinates. They emphasized improving learning outcomes by training on synthetic datasets with highly accurate annotations. These architectures were adapted by replacing their final layers with fully connected regression layers outputting yaw and pitch angles, optimized with mean squared error loss. Integration of head-pose information at the feature level improved prediction accuracy, enabling reliable gaze estimation even under varying conditions.

Finally, Kar (MLGaze)[25] explored the effectiveness of machine learning in detecting and predicting gaze error patterns in consumer eye-tracking systems. Their work aimed to improve the accuracy and reliability of gaze estimation under real-world conditions by modeling systematic error patterns. They transformed raw gaze coordinates into frontal gaze angles (yaw, pitch) using ground truth datasets, which included precise screen locations. By feeding these gaze angles into predictive models, they provided insights into how user behavior and hardware limitations influence tracking accuracy.

Collectively, these studies underscore the growing maturity of gaze estimation and behavioral analysis techniques, which can be adapted to classroom settings for monitoring student attention. By leveraging CNNs, RNNs, and hybrid data fusion approaches, researchers aim to deliver real-time, scalable solutions for identifying inattention and supporting educators in creating more responsive, effective learning environments.

# 3. Methodology

Our proposed approach employs deep learning techniques to capture, analyze, and classify student behaviors indicative of attention and inattention in real-time classroom environments. The system is designed to process live video data, extract relevant behavioral features, and predict attentional states, thereby supporting educators in identifying disengaged students for timely interventions. The key components of our methodology are detailed below[26].

## 3.1. Data Collection

The system relies on live video feeds captured from strategically placed cameras in the classroom. These cameras are positioned to provide a clear view of students' faces and upper body postures while minimizing occlusions. Video data is continuously recorded during lessons to ensure comprehensive coverage of behavioral variations over time. The primary objective of data collection is to obtain rich, temporally dense visual information that reflects natural student behaviors, including eye movements, facial expressions, and body posture.

## 3.2. Preprocessing

To improve the quality and consistency of the collected video data, we implement a preprocessing pipeline with the following steps:

- Noise Reduction: Frames are filtered to remove background noise and visual artifacts that may hinder feature extraction.
- Normalization: Image data is normalized to ensure consistent lighting, contrast, and color balance across different recording sessions and classroom environments.
- Data Augmentation: Augmentation techniques such as random cropping, rotation, flipping, and brightness adjustments are applied to increase the variability of the training dataset and improve the model's generalizability to unseen classroom settings.

This preprocessing stage ensures that the model is robust to variations in environmental conditions, camera angles, and student demographics.

## 3.3. Feature Extraction

A Convolutional Neural Network (CNN) architecture is employed to extract high-level visual features from preprocessed video frames. The CNN model is trained to detect and classify specific behavioral cues that are indicative of student attention or inattention. Key features extracted include:

- Eye Movements: Direction of gaze to determine whether the student is focused on instructional materials or looking away.
- Facial Expressions: Indicators such as yawning, smiling, or frowning, which provide cues about engagement,

fatigue, or emotional state.

- Body Posture: Signs of slouching or leaning, which may suggest disengagement or drowsiness.

## 3.4. Facial Landmarks and Mesh Detection

To enhance the precision of feature extraction, the system uses facial landmarks and face mesh detectors to identify regions of interest on each student's face. These detectors locate and track key facial points across frames, enabling the extraction of fine-grained features such as:

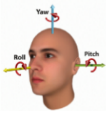
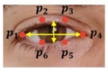
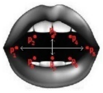
- Mouth Aspect Ratio (MAR): Used to detect yawning or speaking behavior, which can signal fatigue or participation.
- Eye Aspect Ratio (EAR): Used to assess eye openness and detect signs of drowsiness or prolonged closure.
- Head Pose Estimation: Head orientation is analyzed through yaw and pitch angles, providing information about where the student is looking and whether they are oriented toward the teacher or elsewhere.

## 3.5. Head Orientation Analysis

Among the extracted parameters, head orientation—specifically yaw (left-right rotation) and pitch (up-down tilt)—plays a critical role in evaluating attention levels as illustrated in **Table 1**. For example:

- Yaw Angle: High deviation from the front-facing direction indicates the student is looking away from instructional materials.
- Pitch Angle: Downward tilt may suggest disengagement or note-taking, requiring further contextual analysis to distinguish between them.

**Table 1.** Features usage metrics.

| Parameters | Key Facial Features |
|---|---|
| Head orientation |  |
| Drowsiness |  |
| Yawns |  |

These orientation parameters are continuously tracked and fed into the classification model to assess attention levels in real time. The gaze zone definitions categorize driver visual attention based on head orientation, where the Forward zone (heading −10° to 10°, pitch −7° to 6°) represents looking straight ahead at the road, the Left zone (heading −90° to −15°, pitch −7° to 7°) and Right zone (heading 15° to 90°, pitch −7° to 7°) indicate attention directed toward the left or right sides of the environment, respectively, and the Rear zone (heading 17° to 39°, pitch 5° to 20°) corresponds to glances toward the interior rear-view mirror, capturing backward monitoring behavior as described in **Table 2**.

**Table 2.** Interpretation of observations.

| Zone | Heading (°) | Pitch (°) | Meaning |
|---|---|---|---|
| Forward | −10 to 10 | −7 to 6 | Straight ahead |
| Left | −90 to −15 | −7 to 7 | Looking left |
| Right | 15 to 90 | −7 to 7 | Looking right |
| Rear | 17 to 39 | 5 to 20 | Interior rear-view mirror |

## 3.6. Classification and Attention Prediction

The extracted features are processed by a classification module that predicts the attentional state of each student for every frame or time window. The model is trained on labeled examples of attentive and inattentive behavior to distinguish subtle variations in gaze, facial expressions, and posture. Output predictions can be aggregated over time to generate attention profiles for individual students or the entire classroom.

## 3.7. System Output and Educator Interface

The final component of the system is an educator-facing interface that visualizes attention metrics in real time. Teachers can view summary statistics, time series plots, and heatmaps indicating levels of attention across the classroom. This feedback enables teachers to adapt their instructional strategies, re-engage distracted students, and evaluate the effectiveness of teaching interventions.

## 3.8. Concept

The proposed system evaluates student attention levels in real time by analyzing three key behavioral cues: eye gaze, head rotation, and yawning. The diagram illustrates how these cues are processed through a series of decision steps to classify students as *attentive* or *inattentive*.

- Eye Gaze Analysis
  The system detects whether the student's eyes are closed or diverted away from instructional materials. If eye closure is detected, the *glance duration* is measured to determine whether the student is blinking normally or exhibiting prolonged eye closure indicative of drowsiness or inattention.
- Head Rotation Monitoring
  The system analyzes *head pose* using rotation parameters (yaw and pitch) to track how often students turn their heads away from the teacher or screen. Frequent or prolonged deviations may indicate distraction, disengagement, or conversation with peers.
- Yawning Detection
  The system detects *mouth gaping* by measuring the mouth aspect ratio (MAR). If a yawn is detected, the *yawn duration* is computed to distinguish between brief, normal mouth movements and longer yawns that signal fatigue or boredom.

Each of these behavioral indicators feeds into a decision module that evaluates:

- Glance Duration: to capture sustained eye closure or averted gaze.
- How Often: to quantify the frequency of head pose deviations.
- Yawns Duration: to assess the severity and impact of yawning events.

The combined analysis of these metrics enables the system to classify student behavior as *attentive* or *inattentive*. The integration of eye gaze, head rotation, and yawning ensures a multi-modal approach that improves detection accuracy by accounting for different manifestations of inattention.

**Figure 1** illustrates the system's conceptual model for detecting student inattention by analyzing three main cues: eye gaze, head rotation, and yawning.

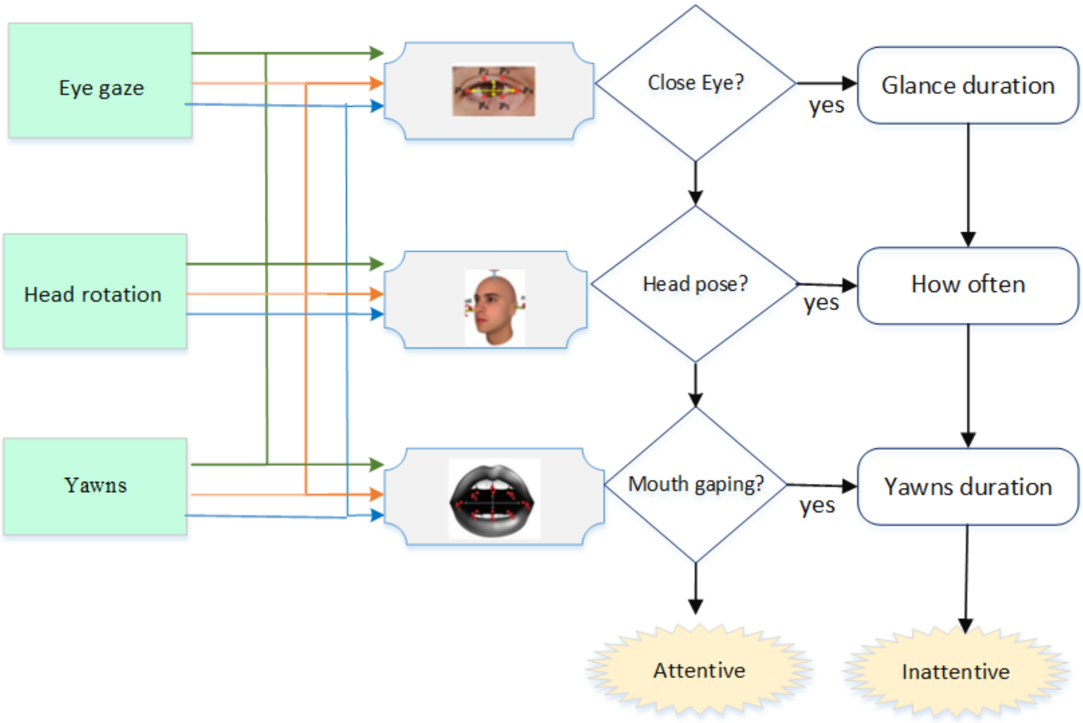


**Figure 1.** Inattentive measurements.

- Eye Gaze: If the student's eyes are closed, the system measures *glance duration* to identify potential drowsiness.
- Head Rotation: Head pose is tracked to determine *how often* the student looks away from instructional materials.
- Yawns: Mouth gaping is detected to measure *yawn duration*, indicating fatigue or boredom.

These metrics are combined to classify students as *attentive* or *inattentive*, supporting real-time monitoring and intervention.

We propose appearance-based gaze estimation approaches using convolutional neural networks (CNNs) to estimate gaze angles directly from eye images and from eye landmark coordinates. The goal is to improve learning by utilizing synthetic data with more accurate annotations.

## 3.9. Model Architecture

In this study, a deep learning model is developed to predict potential student behavior. The model is trained on a comprehensive dataset. The system integrates predictive behavioral modeling to anticipate student actions using machine learning techniques. The dataset is carefully preprocessed, with features normalized and augmented to ensure diversity in training. The model is trained using mean squared error (MSE) as the loss function and the Adam optimizer, with performance evaluated on training and validation sets based on its ability to predict collisions and recommend appropriate maneuvers. A typical Convolutional Neural Network (CNN) consists of three primary types of layers: convolutional, pooling, and fully connected layers.

- Input Data

  The input to a CNN can be in 1D, 2D, or 3D formats, originating from various sources such as sensors, audio signals, videos, or 3D images.
- Convolutional Layers

  A convolutional layer is a fundamental component of the CNN architecture, as illustrated in **Figure 2**. The weights define a convolutional kernel, which is applied to the original input. This procedure repeatedly applies multiple kernels to form an arbitrary number of feature maps, which represent different characteristics of the input tensors; different kernels can, thus, be considered different feature extractors. These layers are tasked with feature extraction. They achieve this by applying convolution operations to the input data. Convolutions use multiple filters defined by parameters such as kernel size, padding, and stride, producing feature maps. Activation functions like ReLU are then applied to enhance the feature extraction process. The output is passed to the subsequent layer for further processing. The convolutional kernels always have the same width as the time series, while their length can vary. This way, when performing a convolution, the kernel moves in one direction from the beginning of a time series towards its end. The elements of the kernel are multiplied by the corresponding elements of the time series that they cover at a given point.

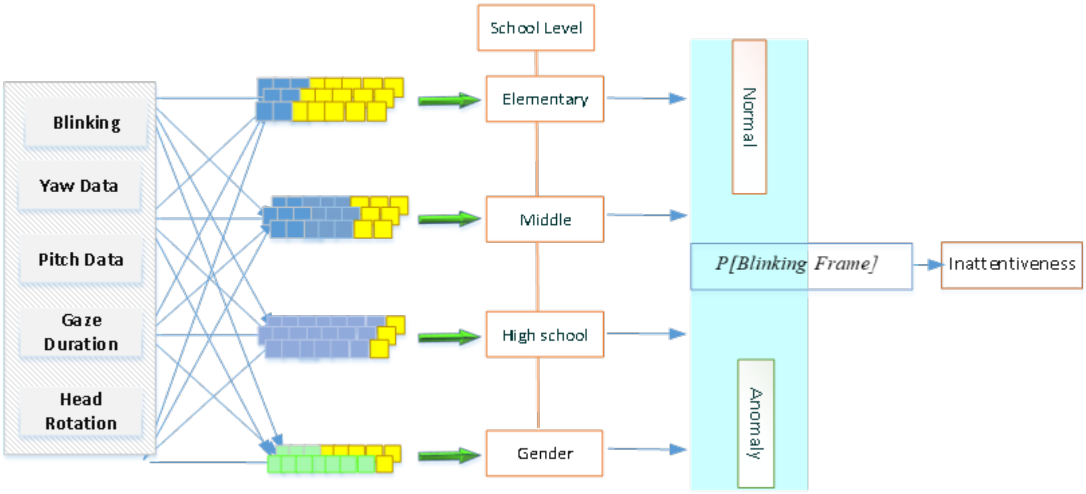


**Figure 2.** Model Architecture.

- Pooling Layers

Typically, following a convolutional layer, pooling layers condense the information from feature maps. For example, in image processing, they significantly reduce the input size, thereby decreasing computational demands and accelerating training. This also strengthens feature detection. Common pooling techniques include max pooling and average pooling. This operation is typically applied only once before the fully connected layer is engaged.

- Fully Connected Layers

The fully connected layer functions as a traditional backpropagation neural network and is used in the final stages of the neural network. It processes the features extracted by earlier layers to generate the final network output, which could be a prediction task, such as forecasting a value, or a classification task, like categorizing images into distinct classes. The output feature maps of the final convolutional or pooling layers are typically flattened. The final fully connected layer typically has the same number of output nodes as the number of classes.

- Last layer activation function

The activation function applied to the last fully connected layer is usually different from the others. An appropriate activation function must be selected according to each task. An activation function applied to the multiclass classification task is a softmax function, which normalizes output real values from the last fully connected layer to target class probabilities, where each value ranges between 0 and 1 and all values sum to 1.

## 3.10. Mathematical Description

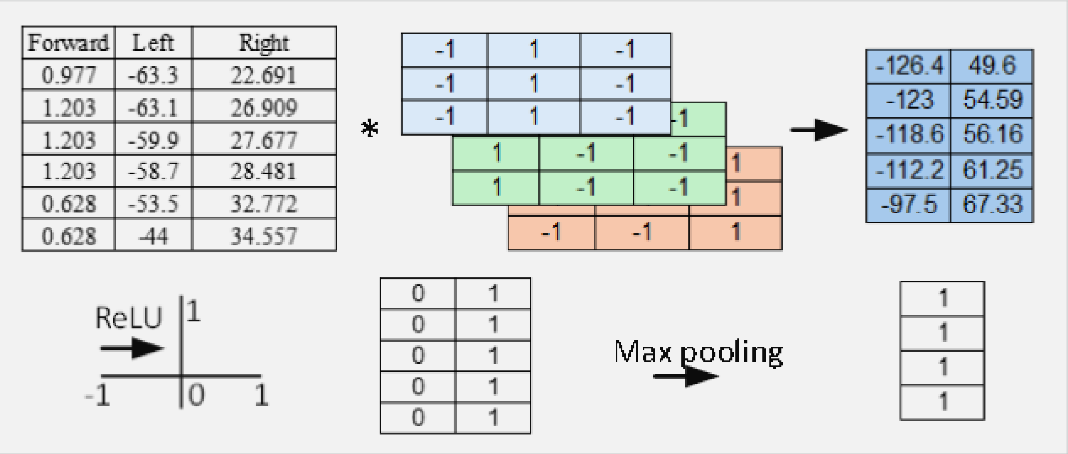
The lengths of input and output time intervals can be expressed as *F* and *P*, respectively. The model input can be written as:

$$x^i = [m_i, m_{i+1}, ..., m_{i+P-1}], i \in [1, N-P-F+1] \quad (1)$$

Where *i* is the sample index, *N* is the length of the time intervals, and $m_i$ is a column vector representing the GazeHeading data.

The CNN is applied to detect an inattentive student, which is called anomalies in the classroom. In this case, only the negative values are considered, and at the same time, present the Gaze data that are smaller than the threshold.

The extraction of features involves a combination of the convolutional and pooling layers, as illustrated in **Figure 3**.



| Forward | Left | Right |
|---|---|---|
| 0.977 | -63.3 | 22.691 |
| 1.203 | -63.1 | 26.909 |
| 1.203 | -59.9 | 27.677 |
| 1.203 | -58.7 | 28.481 |
| 0.628 | -53.5 | 32.772 |
| 0.628 | -44 | 34.557 |

**Figure 3.** Fully connected.

The output of the first convolution and pooling layers can be written as:

$$o_1^j = pool(\sigma(W_1^j x_1^j + b_1^j)), j \in [1, c_1] \quad (2)$$

and the output of the last convolutional and pooling layers can be written as

$$o_n^j = pool(\sigma(W_n^j x_n^j + b_n^j)), j \in [1, c_n]$$

Where $\sigma$ is the activation function. In the prediction, the features learned and outputted by gaze feature extraction are concatenated into a dense vector that contains the final and the highest-level features of the transportation network input. The dense vector can be written as

$$o_L^{flatten} = flatten([o_L^1, o_L^2, ..., o_L^j]), j = c_L \tag{3}$$

where $L$ is the depth of the CNN. Finally, the vector is transformed into output through a fully connected layer. The output can be written as:

$$\widetilde{y} = W_f o_L^{flatten} + b \tag{4}$$

$$= W_f(flaten(pool(\sigma(\textstyle\sum_{k=1}^{c_L-1} (W_L^j x_L^k + b_L^j))))) \\ +b_f \tag{5}$$

where $W_f$ and $b_f$ are the parameters of the fully connected layer, and $\widetilde{y}$ represents the predicted network-wide data anomalies. The CNN uses convolutional filters on its input layer and obtains local connections only where local input neurons are connected to an output neuron (in the convolutional layer). Hundreds of filters are sometimes applied to the input, and the results are merged in each layer. One filter can extract one Gaze feature from the input layer; therefore, hundreds of filters can extract hundreds of features, as illustrated in **Figure 3**. The fully connected layer expresses the negative values that represent the anomalies in each road section.

The classification report indicates strong overall performance with an accuracy of 96% across 6928 samples, as described in **Table 3**. Class 0 shows very high recall (1.00), meaning all true class 0 instances were correctly identified, though its slightly lower precision (0.91) suggests some class 1 samples were incorrectly predicted as class 0. Class 1 achieves perfect precision (1.00), so all predictions of class 1 are correct, but a slightly lower recall (0.93) indicates a small portion of true class 1 instances were missed. The F1-scores (0.95 for class 0 and 0.97 for class 1) confirm a strong balance between precision and recall for both classes. The macro and weighted averages are closely aligned, suggesting that performance is consistent across classes despite the moderate class imbalance, and the model generalizes well without being overly biased toward the larger class. The CNN model is moderately sized (≈250k parameters, ~979 KB) with almost all parameters trainable, indicating sufficient capacity to learn non-linear patterns in the data without being excessively large. The confusion matrix shows excellent discriminatory performance, with a very high number of true positives (3913) and true negatives (2739), and no false positives, implying perfect precision for the positive class. This suggests that when the model predicts a positive gaze state, it is always correct. The presence of 276 false negatives, however, indicates a tendency to miss some positive cases, reflecting a conservative prediction behavior that favors avoiding false alarms at the expense of recall. Overall, the model is highly reliable but slightly under-sensitive, making it suitable for applications where false positives are costly (e.g., incorrect attention alerts), though recall could be improved via threshold tuning or cost-sensitive training if detecting all positive events is critical.

**Table 3.** Performance analysis.

| Metrics<br>Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **0** | 0.91 | 1.00 | 0.95 | 2739 |
| **1** | 1.00 | 0.93 | 0.97 | 4189 |
| **Accuracy** | | | 0.96 | 6928 |
| **Weighted avg** | | 0.96 | 0.96 | 6928 |
| **True Positives (TP)** | | | 3913 | |
| **True Negatives (TN)** | | | 2739 | |
| **False Positives (FP)** | | | 0 | |
| **False Negatives (FN)** | | | 276 | |

# 4. Learning Behavior in the Classroom

In this research, we used artificial intelligence (AI) technologies to detect and analyze student behavior in the classroom with the goal of measuring three key categories: motivation, competition, and challenges. The AI system processed video data to identify specific behavioral cues associated with each category, enabling a detailed, objective assessment of student engagement and learning dynamics.

- Motivation

To measure motivation, AI tools were trained to recognize positive emotional and attentional indicators during lessons, as illustrated in **Figure 4**.

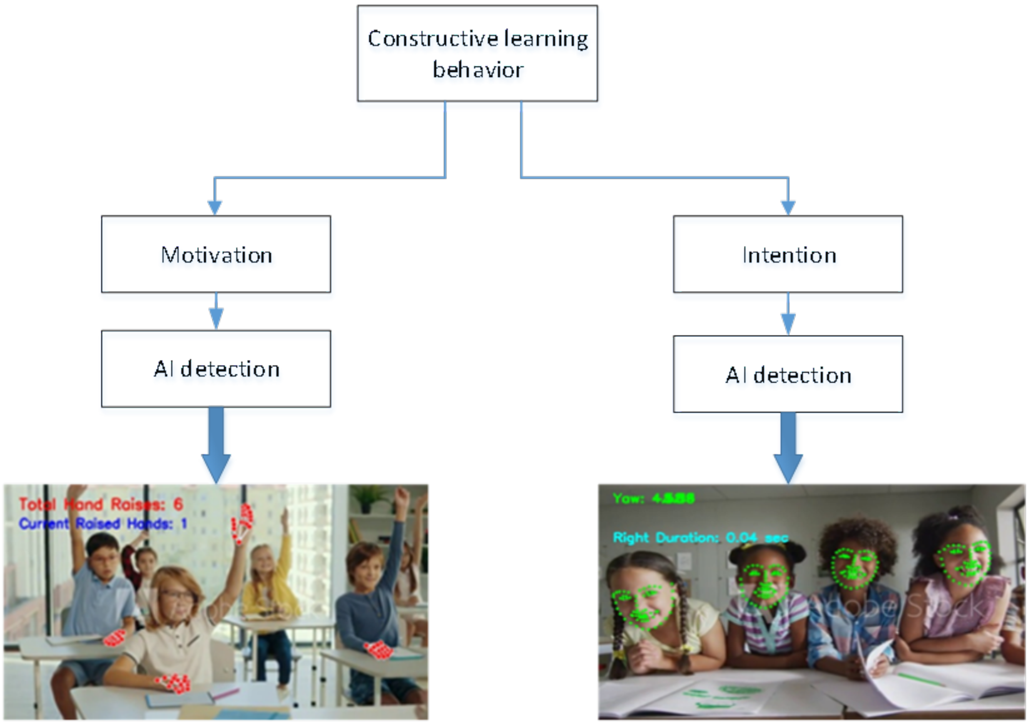


**Figure 4.** Motivation category.

Specific behaviors used as markers of motivation included:

- ○ Smiling, suggesting enjoyment and a positive attitude toward the lesson content.
- ○ Attentive gaze or intention (focused eye contact with the teacher or learning materials), indicating sustained interest and cognitive engagement.

By detecting these features automatically, the AI system provided a way to quantify levels of student motivation across the classroom in real time, complementing self-reported measures from questionnaires.

- Competition

For the category of competition[2], the AI system focused on behaviors that suggest active participation and rivalry among students. Key indicators included:

- ○ Hand-raising frequency, used to signal eagerness to answer questions or contribute before peers.
- ○ Interruptions or overlaps in speech (when available in audio), hinting at competitive dynamics during discussions.

These measures allowed us to assess the degree of competition present in the classroom, offering insights into both student enthusiasm and potential social tensions arising from competitive interactions.

- Challenges

To identify the challenges students faced during lessons, as illustrated in **Figure 5**, the AI system was designed to detect signs of disengagement, confusion, or negative emotions. Relevant behavioral cues included:

- ○ Inattentiveness, such as looking away from the teacher or materials for prolonged periods.
- ○ Head rotation to the right or left, indicating distraction or seeking help from peers instead of following the teacher.
- ○ Eye droopiness or drowsiness, signaling loss of focus or fatigue.
- ○ Facial expressions of sadness or worry, reflecting emotional distress or difficulty with the lesson content, as illustrated in **Figure 5**.

By monitoring these behaviors automatically, the AI system offered continuous, objective data on which students might be struggling, and which moments of the lesson presented the greatest challenges. Through the automated detection of these targeted behaviors, our AI-assisted approach enabled precise measurement of motivation, competition, and challenges in the classroom. This data-driven method supports teachers and researchers in understanding student needs, tailoring instruction, and improving the overall learning environment.

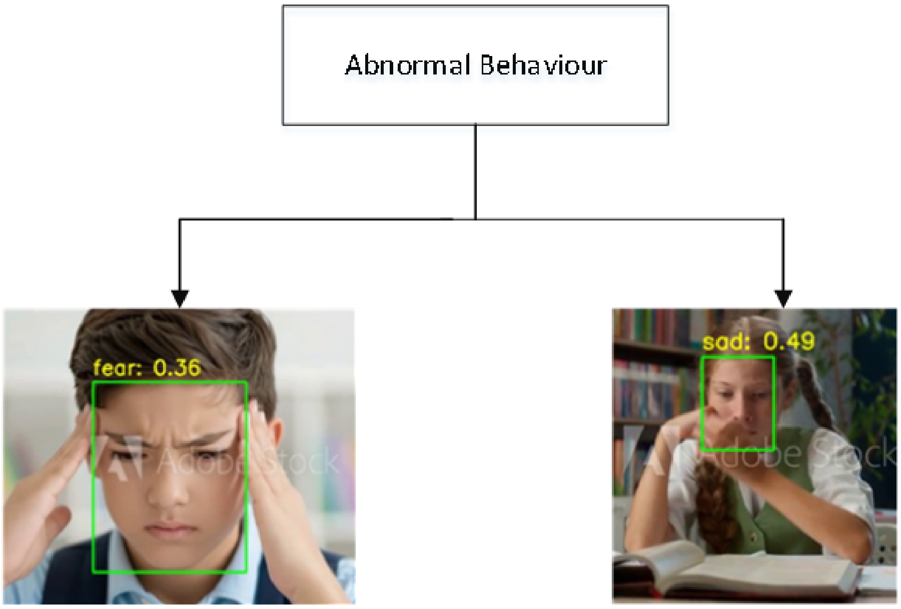


**Figure 5.** Abnormal behavior detection.

## 4.1. Student Behavior Detection

In a previous study, we have used the traditional way based on a questionnaire. We distributed a structured questionnaire designed to capture students' perceptions and experiences related to these categories. This dual approach, behavioral detection and self-report, provided a richer, more reliable understanding of student engagement and learning obstacles. One of the focuses of our study was to assess student motivation during lessons. We approached this in two ways:

### 4.1.1. AI-Based Behavioral Detection

- Motivation

AI tools were trained to recognize positive emotional and attentional indicators that signal motivation, including:

- ○ Smiling, suggesting enjoyment and a positive attitude toward lesson content.
- ○ Attentive gaze or intention (sustained eye contact with the teacher or learning materials), indicating interest and cognitive engagement.
- ○ Frequency of hand-raising and verbal participation, reflecting willingness to contribute and active involvement.

By automatically detecting these features in classroom video and audio, the AI system quantified motivation levels in real time.

- Competition

Competition among students was examined as another important category.

AI-Based Behavioral Detection:

The AI system focused on identifying signs of active participation and rivalry, such as:

- ○ Hand-raising frequency, indicating eagerness to answer questions before peers.
- ○ Interruptions or overlaps in speech (when audio was available), hinting at competitive dynamics.
- ○ Nonverbal cues suggesting tension or eagerness to "win" in group activities.

Questionnaire-Based Assessment:

Students were asked about:

- ○ Their feelings of rivalry with classmates.
- ○ Willingness to outperform peers.
- ○ Perceived pressure to compete in classroom activities.

The integration of these data sources allowed us to explore how competition influenced participation, classroom atmosphere, and social dynamics among students.

- Challenges

Finally, the study aimed to identify the challenges students faced during lessons.

AI-Based Behavioral Detection:

The AI system was designed to recognize signs of disengagement, confusion, or emotional distress, including:

- ○ Inattentiveness, such as looking away from the teacher or materials for prolonged periods.
- ○ Head rotation to the right or left, suggesting distraction or seeking help from peers instead of focusing on instruction.
- ○ Eye droopiness or drowsiness, indicating fatigue or loss of focus.
- ○ Facial expressions of sadness or worry, reflecting emotional discomfort or difficulty with the lesson content.

### 4.1.2. Questionnaire-Based Assessment

The questionnaire asked students about:

- Difficulties understanding lesson material.
- Challenges maintaining focus.
- Issues interacting with peers or the teacher.

The statements in the questionnaire were divided into three categories: motivation, competition, and challenge. The questions were closed questions to be answered on a five-point Likert scale ranging from "strongly disagree" to "Strongly agree". **Table 4** presents some items from different categories.

**Table 4.** Questionnaire categories.

| |
|---|
| **Motivation Category** |
| • I feel more excited when they asked me to write a difficult program. |
| • When I write difficult computer programs, I do not feel fun and entertained. |
| **Competition Category** |
| • I am willing to try hard to be the best in programming among my colleagues. |
| • I try hard to write programs and solve difficult issues before the rest of my colleagues. |
| **Challenge Category** |
| • I like to write computer programs that are challenging and need deep thinking |
| • If I am required to write a difficult computer program, I feel challenged and keep working on it until I finish it. |

Through the combined use of AI-assisted behavioral analysis and targeted questionnaires, this research enabled a thorough investigation into student motivation, competition, and challenges within the classroom. The dual approach provided richer, more reliable insights that can inform instructional design, support strategies, and classroom management, ultimately helping teachers tailor their methods to better meet student needs and improve learning outcomes.

## 5. Results Analysis and Discussion

The proposed deep learning-based inattention detection framework was evaluated using a dataset of recorded classroom video sessions containing annotated attention labels. The system was tested for its ability to recognize inattentive behaviors, including prolonged eye closure, frequent head turning, and yawning, using the described CNN-based feature extraction and behavioral classification pipeline. The analysis results are based on eye gaze detection. **Figure 6** illustrates a gaze heading degree measures the horizontal angle of a student's head orientation relative to facing forward (0°), which signals direct, on-task attention. The range of −100 to 75 degrees captures deviations left (negative) and right (positive), where small angles near 0° (−10 to +10) indicate focused attention, while larger deviations (beyond ±30–40°) suggest distraction or peer interaction. Extreme values (e.g., −80 or +75) often reflect significant disengagement. By analyzing the frequency and duration of these off-center headings, the model classifies attentive vs. inattentive states, enabling real-time monitoring to support improved classroom engagement.

**Figure 7** illustrates gaze pitch degree measures the vertical angle of a student's head or eye orientation relative to looking straight ahead (0°), which indicates typical attentive posture. Negative values (−30 to 0) reflect downward gaze, with mild angles (e.g., −10°) suggesting reading or writing, while steeper angles (e.g., −25° or −30°) may indicate sleepiness or disengagement. Positive values (0 to +30) capture upward gaze, where small angles can mean thinking or observing a higher display, but larger angles may suggest daydreaming. Pitch values near 0° indicate focused behavior, while extreme angles signal potential inattention. Tracking these patterns helps the system classify attentive vs. inattentive states for real-time classroom monitoring.

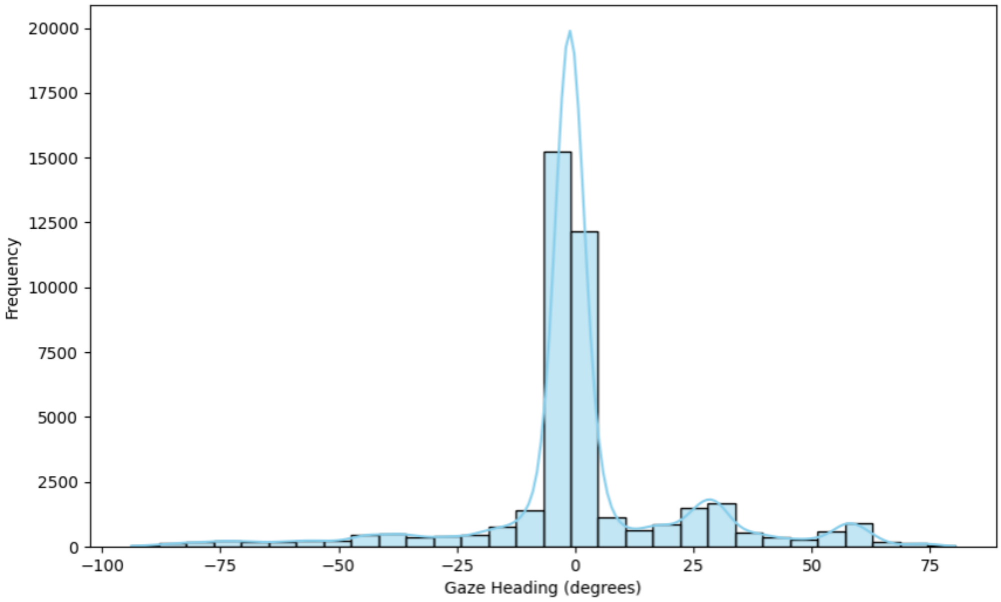


**Figure 6.** Distribution of Gaze Heading.

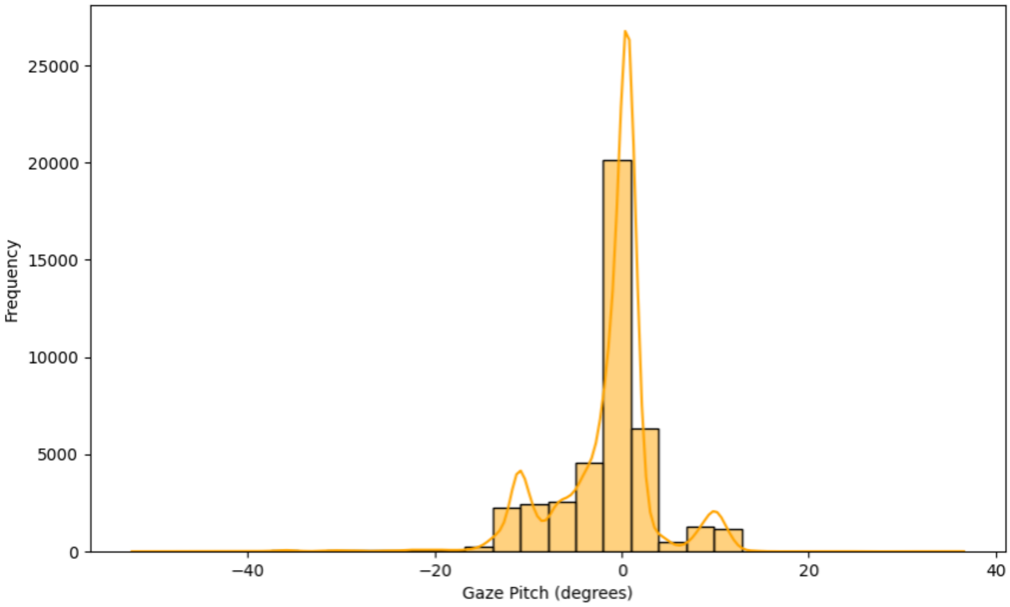


**Figure 7.** Distribution of Gaze Pitch.

**Figure 8** illustrates student gaze direction was divided into Forward (−10° to +10°), Left (angles <−10°), and Right (angles >+10°) zones to assess attentiveness. The system calculates the percentage of time each student spends in these zones, with high Forward Zone proportions indicating strong focus on the teacher or materials. Elevated time in Left or Right Zones suggests possible distractions, like peer conversations or looking away. For example, 80% forward gaze implies high attentiveness, while distributions like 50% forward, 30% left, 20% right may indicate the need for intervention. This approach provides objective, continuous metrics to help teachers identify distracted students and adapt strategies to improve classroom engagement.

**Figure 9** illustrates A histogram of eyelid opening values illustrates how frequently different levels of eyelid openness occur, grouping them into bins (e.g., 0.0–0.1, 0.1–0.2, etc.). This distribution helps assess whether students tend to have eyelids mostly open (values near 1, suggesting alertness), mostly closed (values near 0, indicating drowsiness), or if openness is evenly spread across states. For threshold analysis, eyelid opening is categorized into three states: Closed (0 to 0.1), Slightly Open (0.1 to 0.5), and Mostly Open (0.5 to 1.0). The system counts the number of samples in each category, enabling objective monitoring of alertness levels. Frequent occurrences of closed or slightly open states can signal fatigue or inattention, while a predominance of mostly open values suggests sustained attentiveness in the classroom.

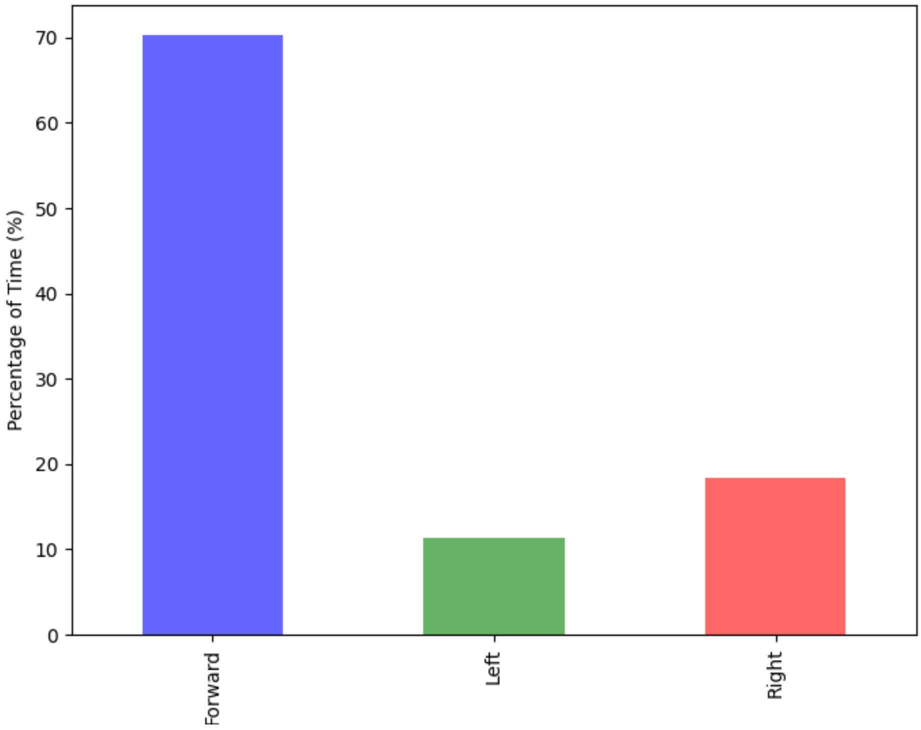


**Figure 8.** Proportion of Attention to different zones.

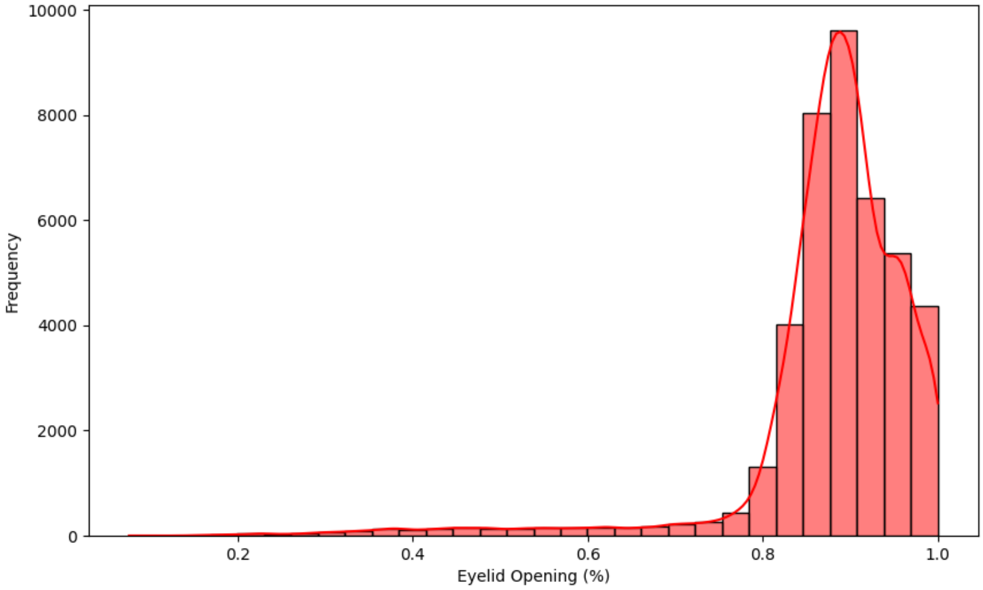


**Figure 9.** Distribution of eyelid opening.

# 6. Discussion

The experimental evaluation was conducted using annotated video data collected from real classroom environments during regular teaching sessions, capturing students seated in typical classroom layouts under natural instructional conditions. Cameras were positioned to record frontal and upper-body views while minimizing occlusion, and the recordings reflect realistic variations in student behavior, lighting conditions, seating arrangements, and interaction dynamics rather than controlled laboratory settings. Students were observed during standard instructional activities such as lectures and guided discussions, allowing attentional behaviors, including gaze shifts, head movements, posture changes, eye closure, and yawning to occur organically. Attentive and inattentive states were annotated based on observable behavioral cues (e.g., prolonged gaze diversion, repeated head rotation, sustained eye closure, and yawning duration), providing ground truth labels for supervised learning. This real-world classroom context supports the ecological validity of the proposed framework and demonstrates its applicability to dynamic, diverse smart classroom environments where student attention is influenced by instructional and environmental factors.

Within this context, the study makes several important contributions to the field of smart classrooms and educational technology. First, it introduces a multi-modal, deep learning–based framework for real-time detection of student inattention that integrates eye gaze, head rotation, facial expressions, and posture analysis. By combining CNN-based spatial feature extraction with RNN-based temporal modeling, the system captures both instantaneous behavioral cues and their temporal evolution, enabling more reliable attention classification than frame-based or single-cue approaches. Second, the proposed method moves beyond subjective and episodic engagement assessment by providing an objective, continuous, and scalable monitoring solution, addressing key limitations of traditional teacher observations and survey-based methods, particularly in large or heterogeneous classrooms. Third, the system is designed with practical deployment in mind, emphasizing real-time performance, robustness to environmental variability, and interpretability of outputs. The educator-facing interface translates model predictions into actionable insights, supporting timely instructional adaptation and targeted intervention. More broadly, this work contributes to the growing body of research applying computer vision and deep learning in education by demonstrating how behavioral analysis techniques originally developed for domains such as driver monitoring and human–computer interaction can be effectively adapted to learning environments.

Despite these strengths, several limitations should be acknowledged. The framework relies primarily on visual cues, which may not fully capture cognitive engagement; for instance, students taking notes or reading materials may be misclassified as inattentive without additional contextual information. Variations in classroom layout, camera placement, lighting conditions, and student demographics may also affect generalizability, and although data augmentation and normalization were applied, performance may degrade under extreme occlusions or non-frontal viewpoints. Furthermore, annotating attention remains inherently challenging, as attention is a latent cognitive process inferred from observable behavior; while indicators such as gaze diversion and yawning are informative, they cannot perfectly represent mental engagement. Ethical and privacy considerations related to continuous video-based monitoring were beyond the primary scope of this study and must be carefully addressed prior to large-scale deployment.

Future research can extend this work in several directions. Incorporating additional multi-modal data sources, such as audio signals, interaction logs, physiological measurements, or learning performance indicators, could improve robustness and help distinguish productive from unproductive behaviors. Adaptive and personalized models that account for individual differences in attention patterns and learning styles may further reduce false positives and enhance usability. Longitudinal studies are also needed to evaluate how real-time attention feedback influences teaching strategies, student behavior, and learning outcomes over extended periods. Finally, future systems should adopt ethical-by-design principles, including privacy-preserving techniques, explainable AI mechanisms, and transparent consent frameworks, to ensure responsible and trustworthy deployment of AI-driven attention monitoring in educational settings.

# 7. Conclusions

This paper introduced a deep learning-based framework for detecting student inattention by analyzing behavioral cues such as eye gaze direction, head rotation, and yawning. By combining convolutional neural networks (CNNs) for feature extraction with recurrent neural networks (RNNs) for temporal analysis, the system delivers real-time, objective assessments of student engagement. The results demonstrate the potential of AI to complement traditional teacher observations, offering continuous monitoring that can help identify distracted students early and support targeted interventions to improve learning outcomes. Our findings show that gaze heading and pitch analysis, along with features like eyelid closure and yawning frequency, can effectively distinguish attentive from inattentive states. The use of unsupervised clustering (e.g., KMeans) further highlights interpretable patterns in student gaze behavior, providing actionable insights for educators.

For future work, the model should be refined by incorporating richer, multi-modal data sources such as speech analysis, physiological signals (e.g., heart rate or galvanic skin response), and environmental factors (e.g., classroom layout). Expanding the dataset to include diverse classrooms, age groups, and cultural contexts will improve generalizability and robustness. Additionally, research should focus on optimizing the system for real-time deployment in live classrooms, ensuring low latency and user-friendly interfaces for teachers.

Critically, any deployment must address ethical considerations, including data privacy, student consent, transparency in monitoring practices, and strategies to avoid bias or unfair targeting. By advancing these technical and ethical dimensions, AI-powered inattention detection can become a practical, responsible tool for enhancing student engagement and learning in real-world educational settings.

## Author Contributions

Conceptualization, J.R.; methodology, J.R., R.R.J. and F.Z.; software, J.R., R.R.J. and F.Z.; validation, J.R., R.R.J. and F.Z.; formal analysis, J.R.; investigation, J.R.; resources, J.R.; data curation, F.Z., R.R.J.; writing—original draft preparation, J.R., R.R.J. and F.Z.; writing—review and editing, J.R.; visualization, J.R., R.R.J. and F.Z.; supervision, J.R.; project administration, J.R.; funding acquisition, J.R. All authors have read and agreed to the published version of the manuscript.

## Funding

This work received no external funding.

## Institutional Review Board Statement

Not applicable.

## Informed Consent Statement

Not applicable.

## Data Availability Statement

Data will be made available on request.

## Conflicts of Interest

The authors declare no conflict of interest.

## References

[1] Sahito, Z.H., Khoso, F.J., Phulpoto, J., 2025. The Effectiveness of Active Learning Strategies in Enhancing Student Engagement and Academic Performance. Journal of Social Sciences Review. 5(1), 110–127. DOI: https://doi.org/10.62843/jssr.v5i1.471

[2] Raiyn, J., 2016. The Role of Visual Learning in Improving Students' High Order Thinking Skills. Journal of Education and Practice. 7(24), 115–121.

[3] Raiyn, J., Tilchin, O., 2016. The Self-Formation of Collaborative Groups in a Problem-Based Learning Environment. Journal of Education and Practice. 7(26), 120–126.

[4] Le, H.V., 2021. An Investigation into Factors Affecting Concentration of University Students. Journal of English Language Teaching and Applied Linguistics. 3(6), 7–12.

[5] Raiyn, J., 2017. Toward Development Game-Based Adaptive Learning. Journal of Education and Practice. 8(28), 104–112.

[6] Vehlen, A., Kellner, A., Normann, C., et al., 2023. Reduced Eye Gaze during Facial Emotion Recognition in Chronic Depression: Effects of Intranasal Oxytocin. Journal of Psychiatric Research. 159, 50–56.

[7] Chang, K.-M., Chueh, M.-T.W., 2019. Using Eye Tracking to Assess Gaze Concentration in Meditation. Sensors. 19(7), 1612. DOI: https://doi.org/10.3390/s19071612

[8] Mesfin, G., Hussain, N., Covaci, A., et al., 2019. Using Eye Tracking and Heart-Rate Activity to Exam-

ine Crossmodal Correspondences QoE in Multimedia. ACM Transactions on Multimedia Computing, Communications, and Applications. 15(2), 1–22. DOI: https://doi.org/10.1145/3303080

[9] Akinyelu, A.A., Blignaut, P., 2020. Convolutional Neural Network-Based Methods for Eye Gaze Estimation: A Survey. IEEE Access. 8, 142581–142605. DOI: https://doi.org/10.1109/ACCESS.2020.3013540

[10] Hammadi, S.S., Majeed, B.H., Hassan, A.K., 2023. Impact of Deep Learning Strategy in Mathematics Achievement and Practical Intelligence among High School Students. International Journal of Emerging Technologies in Learning. 18(6), 42–52.

[11] Jan, B., Farman, H.H., Imran, M., 2019. Deep Learning in Big Data Analytics: A Comparative Study. Computers and Electrical Engineering. 75, 275–287. DOI: https://doi.org/10.1016/j.compeleceng.2017.12.009

[12] Dong, S., Wang, P., Abbas, K., 2021. A Survey on Deep Learning and Its Applications. Computer Science Review. 40, 100379. DOI: https://doi.org/10.1016/j.cosrev.2021.100379

[13] Weng, C., Chen, C., Ai, X., 2023. A Pedagogical Study on Promoting Students' Deep Learning through Design-Based Learning. International Journal of Technology and Design Education. 33, 1653–1674.

[14] Hu, Z., Li, S., Zhang, C., et al., 2020. Dgaze: CNN-Based Gaze Prediction in Dynamic Scenes. IEEE Transactions on Visualization and Computer Graphics. 26(5), 1902–1911. DOI: https://doi.org/10.1109/TVCG.2020.2973473

[15] Pereira, A.S., Wahi, M.M., 2019. Deeper Learning Methods and Modalities in Higher Education: A 20-Year Review. Journal of Higher Education Theory and Practice. 19(8), 48–71. DOI: https://doi.org/10.33423/jhetp.v19i8.2672

[16] Pathirana, P., Senarath, S., Meedeniya, D., et al., 2022. Eye Gaze Estimation: A Survey on Deep Learning-Based Approaches. Expert Systems with Applications. 199, 116894. DOI: https://doi.org/10.1016/j.eswa.2022.116894

[17] Fuhl, W., Castner, N., Kasneci, E., et al., 2018. Rule-Based Learning for Eye Movement Type Detection. In Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data (MCPMD '18), New York, NY, USA, 22 October 2018; pp. 1–6. DOI: https://doi.org/10.1145/3279810.3279844

[18] Tesch, F., Coelho, D., Drozdenko, R., 2011. The Relative Potency of Classroom Distracts on Student Concentration: We Have Met the Enemy, and He Is Us. Proceedings of the American Society of Business and Behavioral Sciences. 18(1), 886–894.

[19] Li, S., Liu, T., 2021. Performance Prediction for Higher Education Students Using Deep Learning. Complexity. 2021(1), 9958203. DOI: https://doi.org/10.1155/2021/9958203

[20] Vijaypriya, V., Uma, M., 2023. Facial Feature-Based Drowsiness Detection with Multi-Scale Convolutional Neural Network. IEEE Access. 11, 63417–63429. DOI: https://doi.org/10.1109/ACCESS.2023.3288008

[21] Xiao, L., Zhu, Z., Liu, H., et al., 2023. Gaze Prediction Based on Long Short-Term Memory Convolution with Associated Features of Video Frames. Computers and Electrical Engineering. 107, 108625. DOI: https://doi.org/10.1016/j.compeleceng.2023.108625

[22] Wu, Y., 2025. A Deep Learning Recognition Method for Students' Abnormal Behaviors in Smart Classroom Teaching Scenarios. International Journal of High Speed Electronics and Systems. 34(4), 2540291. DOI: https://doi.org/10.1142/S0129156425402918

[23] Kanade, P., David, F., Kanade, S., 2021. Convolutional Neural Networks (CNN)-Based Eye-Gaze Tracking System Using Machine Learning Algorithm. European Journal of Electrical Engineering and Computer Science. 5(2), 36–40. DOI: https://doi.org/10.24018/ejece.2021.5.2.314

[24] Yoo, S., Jeong, S., Jang, Y., 2021. Gaze Behavior Effect on Gaze Data Visualization at Different Levels of Abstraction. Sensors. 21(14), 4686. DOI: https://doi.org/10.3390/s21144686

[25] Kar, A., 2020. MLGaze: Machine Learning-Based Analysis of Gaze Error Patterns in Consumer Eye Tracking Systems. Vision. 4(2), 25. DOI: https://doi.org/10.3390/vision4020025

[26] Asad, K., Tibi, M., Raiyn, J., 2016. Primary School Pupils' Attitudes toward Learning Programming through Visual Interactive Environments. World Education Journal. 7, 20–26.