AI-POWERED REAL-TIME CLASSROOM ATTENTIVENESS ANALYSIS USING DEEP LEARNING

[1]Kunapureddy Maneesha, [2]Nenavath Abhishek Naik, [3]Marachu Ganesh, [4]Kunapareddy Mohan Venkata Sai
[1234]Department of Electronics and Communication Engineering
[1234]RVR & JC College of Engineering, India

Abstract : An AI-driven system for real-time analysis of student attentiveness in classroom environments is proposed using a vision-based framework. Live video streams are processed using the YOLO object detection model to detect and localize student faces. The extracted facial regions are analyzed using a deep learning–based emotion recognition model trained on custom and publicly available datasets such as KDEF. The model classifies emotions including happiness, neutrality, surprise, sadness, and boredom. These emotional states are mapped to attentive and inattentive categories based on predefined criteria. The system computes overall classroom attentiveness in real time and visualizes it using an attentiveness-versus-time graph. The proposed approach provides a non-intrusive and scalable solution for monitoring engagement. It supports educators in making data-driven decisions to improve teaching effectiveness and learning outcomes.

***Indexterms* - Smart Classroom, Emotion Recognition, YOLO, Deep Learning models, Computer Vision, Student Engagement Monitoring**

I. INTRODUCTION

The concept of intelligent classrooms has gained increasing attention in recent years. Traditional attention detection methods often struggle with pose variation and real-time performance[2]. The YOLO object detection algorithm addresses these limitations by providing fast and accurate detection. Modern smart classroom systems integrate sensors, computer vision, and artificial intelligence to assist instructors in understanding classroom dynamics and student participation. One of the most important indicators of effective teaching is student attentiveness during lectures. Student attentiveness reflects the level of cognitive engagement while interacting with educational content. When students remain attentive, they are more likely to understand explanations, participate in discussions, and retain knowledge effectively. However, measuring attentiveness objectively is challenging.

Teachers often rely on their intuition based on facial expressions or body posture; however, these methods cannot provide quantitative information regarding engagement levels. In this project, Computer vision-based monitoring systems provide an alternative approach by automatically analyzing visual information captured by classroom cameras. Facial emotion recognition is particularly useful because emotional expressions often reflect students' reactions to instructional materials. By combining face detection with emotion recognition algorithms, it becomes possible to estimate attentiveness using facial cues observed in classroom video streams. In the proposed system, YOLO-based face detection identifies student faces in real time, while the EfficientNetB0 deep neural network analyzes facial features to determine emotional states. The overall framework consists of video acquisition, face detection, emotion recognition, attentiveness classification, and engagement metric visualization. This architecture enables real-time analysis of classroom engagement and provides instructors with objective feedback on student attentiveness.

II. LITERATURE REVIEW

A. Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student's Behavior Recognition

The researchers proposed a deep learning framework that utilized classroom video recordings to recognize student behavior patterns associated with attentiveness. The system employed convolutional neural networks to extract facial features and body posture information from video frames. By analyzing these features, the model classified students into attentive and inattentive categories in real time. The study demonstrated that deep learning models could effectively automate engagement monitoring and reduce the burden on instructors. The performance of the system depends on environmental conditions such as lighting intensity, camera placement, classroom layout[1].

B. Student Recognition and Activity Monitoring in E-Classes Using Deep Learning in Higher Education

Alruwais et al. [2] proposed a deep learning-based system for monitoring student behavior in online classroom environments. The system utilizes webcam video feeds to perform student recognition and analyze activities during virtual lectures. It integrates facial recognition techniques with activity analysis methods to evaluate student participation and engagement levels. By processing real-time visual data, the system is capable of identifying students and tracking their attentiveness throughout the session.

C. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild

Mollahosseini et al. [3] introduced AffectNet, a large-scale facial expression dataset designed for emotion recognition in real-world conditions. The dataset contains a wide variety of facial images collected from the internet, labeled with different emotional categories as well as valence and arousal information. This diversity allows deep learning models to learn more generalized and robust features for emotion classification. AffectNet has been widely used for training and evaluating facial emotion recognition systems. However, the dataset itself does not provide a direct framework for attentiveness analysis, but it serves as a valuable resource for improving the accuracy of emotion-based models.

D. ***Smart Classroom: A Deep Learning Approach Towards Attention Assessment Through Class Behavior Detection***

Parambil et al. [4] proposed a smart classroom framework that employs deep learning techniques to assess student attention through behavior detection. The system analyzes classroom video data to identify patterns in student activities and posture, enabling the estimation of engagement levels during lectures. By focusing on behavioral cues, the approach provides a non-intrusive method for monitoring classroom dynamics. However, the system primarily relies on activity-based analysis and does not extensively incorporate facial emotion recognition, which can offer more detailed insights into student attentiveness.

III. PROPOSED METHODOLOGY

The proposed system begins with video acquisition using a camera installed in the classroom environment. The captured video stream is divided into individual frames that are processed sequentially. Each frame is analyzed by the YOLO object detection model which identifies student faces within the image. Once faces are detected, the corresponding facial regions are cropped and provided as input to the emotion recognition module. The emotion recognition network is based on the EfficientNetB0 architecture which uses residual connections to enable the training of deep convolutional networks.

The network extracts hierarchical features from facial images and classifies them into emotional categories including happy, neutral, bored, surprised, and sad. These emotional states are mapped to attentiveness categories where happy and neutral expressions are interpreted as attentive while bored or sad expressions indicate reduced engagement. The attentiveness statistics are aggregated across multiple frames to compute overall classroom engagement. Temporal graphs are generated to visualize how attentiveness changes throughout the lecture session. As illustrated in Fig. 1: Pipeline of the System, the detected facial regions are then processed through multiple analytical modules to extract behavioral features. These modules include facial emotion recognition, head pose estimation, and optionally eye gaze analysis. Facial emotion recognition analyzes expressions to determine emotional states such as engagement or boredom, while head pose estimation identifies the direction of the student's head orientation. If implemented, eye gaze analysis further refines the estimation of where the student is focusing.

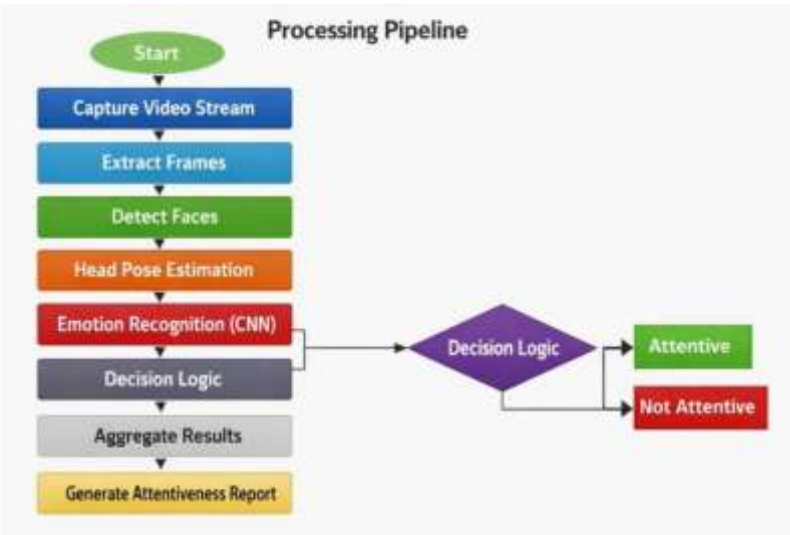

Fig 1: Proposed System Processing Pipeline

The system combines the outputs from multiple modules using a rule-based attentiveness decision mechanism. By evaluating several behavioral cues simultaneously, the system improves the accuracy and reliability of attentiveness classification.

3.1 Face Detection

Face detection serves as the foundational step of the entire methodology, as all subsequent analysis depends on accurately locating student faces within each video frame. The purpose of this stage is to identify human facial regions and isolate them from the background so that only relevant visual information is processed further. Once faces are detected, the corresponding regions are cropped and extracted from the original frame. Cropping helps eliminate unnecessary background noise, thereby improving computational efficiency and model accuracy. The extracted facial images are then forwarded to the emotion recognition and head pose estimation modules.

3.2 Head Pose Estimation

Head pose estimation is an important component used to determine whether a student is visually oriented toward the instructor or distracted by surrounding activities. Human attentiveness is strongly associated with head orientation, making this feature highly valuable for behavioral analysis. The system estimates head orientation by detecting facial landmarks such as eye corners, nose tip, chin position, and mouth edges. These landmarks provide geometric reference points that allow the calculation of head rotation angles relative to the camera. Three primary rotational parameters are considered: pitch, yaw, and roll. Pitch represents upward or downward head movement, yaw represents left or right turning, and roll indicates sideways tilting of the head. Using these parameters, the system evaluates whether a student's head is aligned toward the front of the classroom. Threshold values are defined experimentally to classify orientations as attentive or distracted.

3.3 Facial Emotion Recognition Using Deep Learning

Facial emotion recognition forms the core intelligence of the proposed system. This module analyzes facial expressions to identify emotional states associated with attentiveness or disengagement. A deep learning approach based on a Convolutional Neural Network (CNN) utilizing the EfficientNet-B0 architecture was employed for emotion classification.

Fig 2: EfficientNet-B0 Architecture

The model was trained using the custom dataset containing approximately 350 labeled classroom images. Each image was resized to a standardized input dimension before training. Rectified Linear Unit (ReLU) activation functions were used in hidden layers to introduce non-linearity, while the output layer used Softmax activation to classify images into multiple emotional categories. Cross-entropy loss was selected as the optimization objective, and the Adam optimizer was applied to accelerate convergence during training. Training was conducted over multiple epochs with a moderate batch size to balance learning stability and computational efficiency. The model achieved high training accuracy, indicating successful learning of dataset patterns. However, testing accuracy remained significantly lower, revealing overfitting caused primarily by limited dataset size and insufficient diversity. Despite this limitation, the model demonstrated the feasibility of emotion-based attentiveness estimation. EfficientNet uses optimized scaling of depth, width, and resolution for better performance.

Table 1: Emotion Classification and Attentiveness Scoring Parameters

Emotion	**Weights Assigned**	**Contribution**	**Attentiveness Category**	**Description**
Happy	1.3	+ Positive	Attentive	Indicates high engagement and positive interest in the lecture
Neutral	1.2	+ Positive	Attentive	Represents normal focus without strong emotional expression
Surprise	1.1	+ Positive	Attentive	Suggests curiosity or reaction to new information
Angry	0.7	− Negative	Not Attentive	Reflects discomfort or distraction from the learning process
Bored	0.6	− Negative	Not Attentive	Indicates lack of interest
Motion	--	−40 Penalty	Not Attentive	Significant movement

				reduces attentiveness score due to possible distraction

A weighted correction mechanism is applied to raw emotion probabilities to prioritize engagement-related expressions such as happiness and neutrality, while reducing the influence of negative emotions. This improves the robustness of attentiveness classification in real-world classroom conditions.

3.4 Attentiveness Decision Logic Implementation

The attentiveness decision module combines outputs from multiple analytical components to produce the final classification result. Instead of depending on a single parameter, the system applies rule-based logic integrating head pose estimation, emotion recognition results, and optional gaze estimation. When the student's head orientation is forward-facing and the detected emotion indicates engagement or neutrality, the system classifies the student as attentive. Otherwise, the student is labeled as not attentive. This decision is computed for each detected face within every processed frame. The use of multi-factor decision logic improves robustness by reducing false classifications caused by temporary facial expressions or brief movements.

The attentiveness decision module integrates outputs from multiple analytical components to determine the final classification of each student. Rather than relying on a single parameter, the system employs a rule-based decision mechanism that combines emotion recognition results with motion analysis to improve reliability and robustness.

Initially, facial regions detected using the YOLO model are processed through a deep learning–based emotion recognition model, which outputs probability scores for various emotional states, including happiness, neutrality, surprise, sadness, fear, anger, and disgust. To enhance classification accuracy, a weighted correction scheme is applied to these emotion scores, assigning higher importance to engagement-related emotions such as happiness, neutrality, and surprise, while reducing the influence of negative emotions. The dominant emotion is then determined based on the maximum corrected score. ubsequently, an attentiveness score is computed by aggregating the confidence values of emotions categorized as attentive (happy, neutral, and surprise). This score is normalized and capped to ensure consistency across frames. In parallel, motion detection is performed using frame differencing techniques, where significant pixel variations between consecutive frames indicate student movement. If motion exceeds a predefined threshold, a penalty is applied to the attentiveness score to account for potential distraction.

The final classification is derived using a rule-based logic: a student is labeled as attentive only if the dominant emotion belongs to the predefined attentive category and no significant motion is detected. Otherwise, the student is classified as not attentive. This decision process is executed for each detected face in every processed frame, enabling continuous and real-time monitoring.

By incorporating both emotional and behavioral cues, the proposed multi-factor decision logic reduces false positives caused by transient facial expressions or brief movements, thereby improving the overall robustness and reliability of attentiveness estimation in dynamic classroom environments.

3.5 Graph Generation

To visualize attentiveness patterns, the stored statistical data is used to generate a time-series graph representing engagement variations during the lecture. The horizontal axis represents time intervals, while the vertical axis represents the number or percentage of attentive students. Graphical visualization provides intuitive understanding of classroom dynamics by highlighting moments of increased or decreased attention.

IV. Result And Discussion

The This section presents the experimental results and performance evaluation of the proposed Student Attentiveness Monitoring System. The performance of the system is analyzed at two levels:

1. **Deep learning model performance**(emotion Recognition accuracy)
2. **Overall Classroom attentiveness monitoring effectiveness**

The evaluation includes training and testing results, overfitting analysis, confusion matrix interpretation, performance metrics, and classroom-level attentiveness analytics generated from video input.

4.1 Emotion Recognition Model Performance

The facial emotion recognition module was trained using a custom dataset containing approximately 350 labeled classroom images. The model was evaluated using training and testing datasets to measure learning capability and generalization performance.

4.2 Training Results

The training phase plays an important role in evaluating how effectively the proposed model learns patterns from the labeled facial emotion dataset. During this stage, the model is exposed to training images and learns to identify relationships between facial features and corresponding emotion labels. Training on multiple datasets helped the model learn a broader range of facial expressions and improved its ability to recognize emotions under varying conditions. During training, performance metrics such as training accuracy and training loss were monitored to evaluate the learning progress of the model. Training accuracy represents the percentage of correctly classified images in the training dataset, while training loss indicates how well the model is minimizing prediction errors during the optimization process. The model achieved an approximate training accuracy of 77.93%, indicating that it successfully learned the patterns and facial features present in the training data.

4.3 Testing Results

The testing phase evaluates the model's generalization ability, which refers to how well the trained model performs on unseen images that were not part of the training dataset. This stage is critical for assessing the real-world applicability of the proposed attentiveness monitoring system. For testing, a separate portion of the dataset was used to evaluate the model's prediction performance. The testing results showed an approximate testing accuracy of around 72.27%. The difference between training and testing performance suggests the presence of overfitting, where the model learns the training data too closely but struggles when exposed to new images.

4.4 Comparative Training Results on Multiple Datasets

To improve the robustness of the emotion recognition module, experiments were conducted using different datasets available from Kaggle and the custom classroom dataset collected for this project. These datasets contain various facial expressions captured under different conditions, helping the model learn more generalized facial features. The performance of the model across different datasets was analyzed and summarized in Table.

Table 2: Model Performance on Different Datasets.

Dataset	**Model**	**Training Accuracy**	**Testing Accuracy**
KDEF	MobileNetV2	73.19%	74.83%
KDEF	EfficientNetB2	57.62%	85.20%
Collected data set	EfficientNet-B0	77.93%	72.27%

The custom-collected dataset, specifically created for this project, was trained using the EfficientNetB0 architecture for facial emotion recognition. The model achieved a training accuracy of 77.93% and a testing accuracy of 72.27%. The observed gap between training and testing accuracy suggests a moderate level of overfitting, where the model performs well on the training data but shows reduced performance on unseen samples.

4.5 Accuracy and Loss Curves

During the training process, the performance of the deep learning models was monitored using accuracy and loss metrics across multiple epochs. These metrics help evaluate how effectively each model learns patterns from the dataset and whether the training process converges properly. Since multiple datasets and model architectures were used in this study, separate accuracy and loss curves were generated for each experiment. The training accuracy represents the percentage of correctly classified images during the training phase. As the training progresses through successive epochs, the models gradually improve their ability to recognize facial expression patterns present in the dataset. An increasing accuracy curve indicates that the model is successfully learning relevant features from the training images. The training loss, on the other hand, measures the difference between the predicted outputs of the model and the actual labels in the dataset. A lower loss value indicates that the model predictions are closer to the true labels. During effective training, the loss curve typically decreases over time, showing that the optimization algorithm is minimizing prediction errors and guiding the model toward convergence.

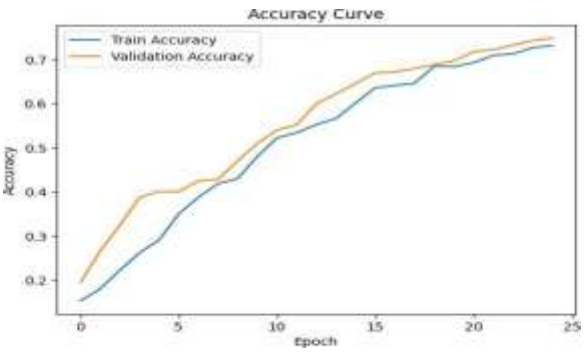

Fig 3: Training Accuracy vs Epochs for MobileNetV2 on

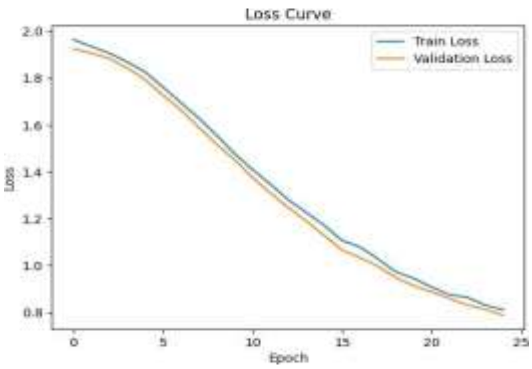

Fig 4: Training Loss vs Epochs for MobileNetV2 on KDEF KDEF

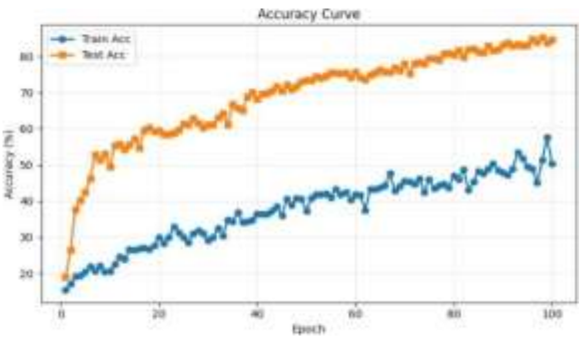

Fig 5: Training Accuracy vs Epochs for EfficientNetB2 KDEF

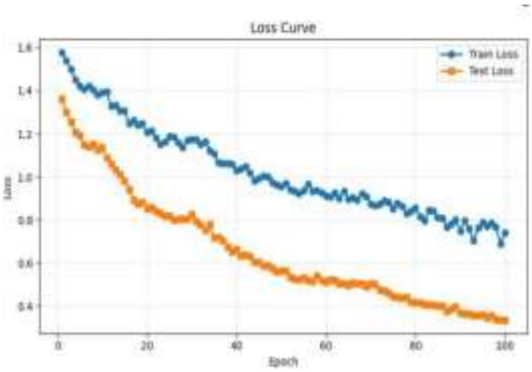

Fig 6: Training Loss vs Epochs for EfficientNetB2 on KDEF on

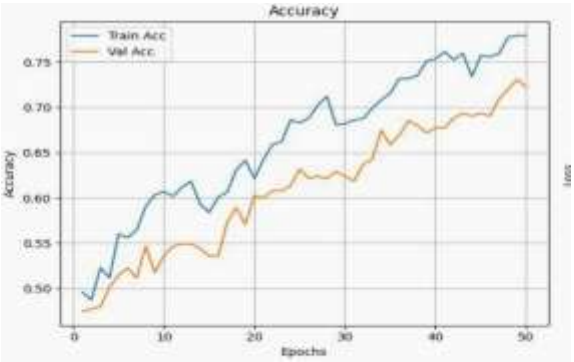

Fig 7: Training Accuracy vs Epochs for EfficientNetB0 Collected Dataset

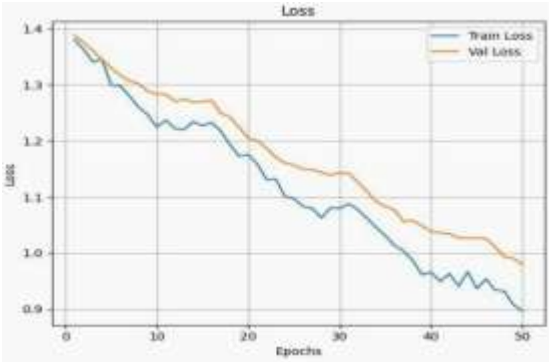

Fig 8: Training Loss vs Epochs for EfficientNetB0 on Custom on Custom Collected Dataset

These graphs help visualize the learning behavior of each model during training. Models trained on larger datasets such as KDEF generally show more stable learning patterns, while the model trained on the custom collected dataset achieved higher training accuracy but may exhibit signs of overfitting due to the smaller dataset size. The corresponding training curves are shown in Fig. 3 to Fig. 8, where each model includes two graphs representing accuracy and loss trends across training epochs. By analyzing these plots, it becomes easier to understand how each model performs during the training process and how effectively it learns facial expression features from different datasets.

4.6 Confusion Matrix Analysis

A confusion matrix is an important evaluation tool used to analyze the classification performance of a machine learning model. It provides a detailed comparison between the actual class labels and the predicted class labels generated by the model. This allows for a deeper understanding of how well the model distinguishes between different emotion categories. The confusion matrix consists of rows representing the actual classes and columns representing the predicted classes. Each cell in the matrix indicates the number of instances where a particular class was predicted correctly or incorrectly. Diagonal values in the matrix represent correct predictions, while off-diagonal values indicate misclassifications. Confusion Matrix of Emotion Recognition Model presents the classification results for the emotion recognition system used in the attentiveness monitoring framework. The matrix shows how accurately the model predicted different facial emotion categories such as Neutral, Happy, Bored, Surprised, and Upset.

From the confusion matrix, it can be observed that the model performs well in identifying certain expressions, while some confusion occurs between similar emotional states. For example, expressions such as neutral and bored may sometimes be misclassified due to similarities in facial appearance. Despite these minor misclassifications, the confusion matrix helps highlight the overall effectiveness of the model in recognizing facial emotions. By analyzing the confusion matrix, it becomes easier to identify areas where the model can be improved, such as increasing dataset size, improving class balance, or applying additional training techniques to enhance classification performance.

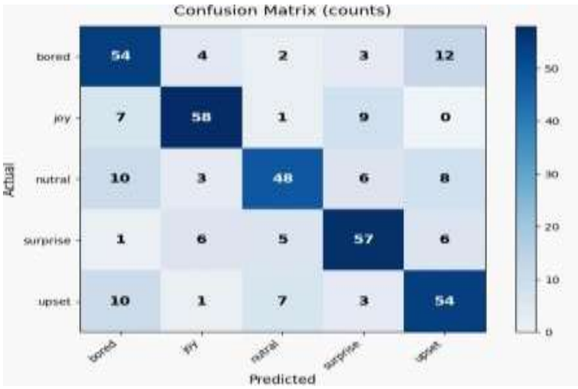

Fig 9: Confusion Matrix of the EfficientNetB0 Model Custom Collected Dataset

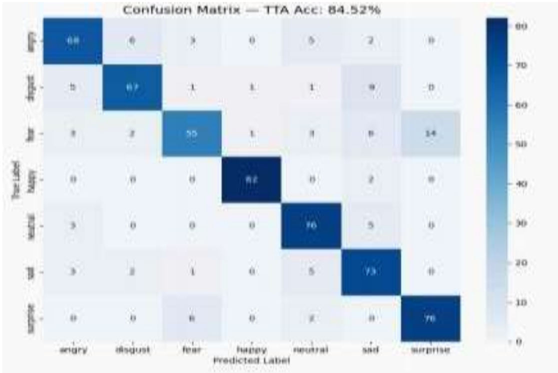

Fig 10: Confusion Matrix of EfficientNetB2 on KDEF Dataset on

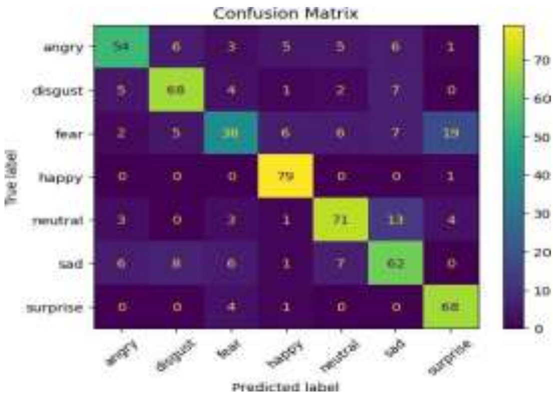

Fig 11: Confusion Matrix of MobileNetV2 on KDEF Dataset

4.7 Overfitting Analysis

Overfitting occurs when a model memorizes training data instead of learning generalized patterns applicable to new data.

Reasons for Overfitting

- Small dataset size (~350 images)
- Limited diversity of student faces
- Lighting and environmental similarity
- Possible class imbalance
- Limited augmentation variation

Effects of Overfitting

- Very high training accuracy
- Lower testing accuracy
- Reduced real-world reliability

A confusion matrix was used to evaluate the classification performance of the proposed model in detail. It provides a comprehensive understanding of prediction behavior by comparing actual and predicted classes. The confusion matrix consists of four components: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). True positives represent correctly identified attentive students, while true negatives indicate correctly classified non-attentive students. False positives correspond to non-attentive students incorrectly classified as attentive, and false negatives represent attentive students misclassified as non-attentive.

1. Accuracy

Accuracy measures the overall correctness of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Explanation: Represents the proportion of total correctly classified samples.

2. Precision

Precision measures prediction reliability for attentive classification.
DR

$$Precision = \frac{TP}{TP + FP}$$

Explanation: It indicates how many students predicted as attentive were truly attentive. High precision -> fewer false alarms.

3. Recall (Sensitivity)

Recall measures detection capability.

$$Recall = \frac{TP}{TP + FN}$$

Explanation: Shows how many truly attentive students were correctly detected. High recall → fewer missed attentive students.

4. F1-Score

F1-Score balances precision and recall.

$$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4.8 Classroom-Level Attentiveness Analysis

The complete system was tested using classroom video input to evaluate real-world functionality. The system generated the following outputs:

- Number of students detected per frame
- Number of attentive students
- Number of non-attentive students
- Frame-wise attentiveness classification
- Time-based engagement variation
- Overall lecture attentiveness score

4.9 Average Attentiveness Calculation

Over all lecture attentiveness is calculated using aggregated frame predictions.

$$\text{Average Attentiveness} = \frac{TotalObservations}{TotalAttentiveInstances} \times 100$$

4.10 Output and Visualization

The system was evaluated using a 10-minute classroom video to analyze its real-time performance in detecting student attentiveness. During execution, the system continuously captures video frames and identifies student faces using the detection model. Each detected face is classified as either attentive or non-attentive based on facial expressions and behavioral cues. The output interface displays the total number of students, along with the count of attentive and non-attentive individuals, and computes the overall attentiveness percentage dynamically.

In addition to real-time detection, the system provides meaningful visualizations to support analysis. It generates a time-based graph illustrating variations in attentiveness throughout the session, allowing identification of periods with high and low engagement. Furthermore, a final report is produced that summarizes key statistics, including average, maximum, and minimum attentiveness levels.

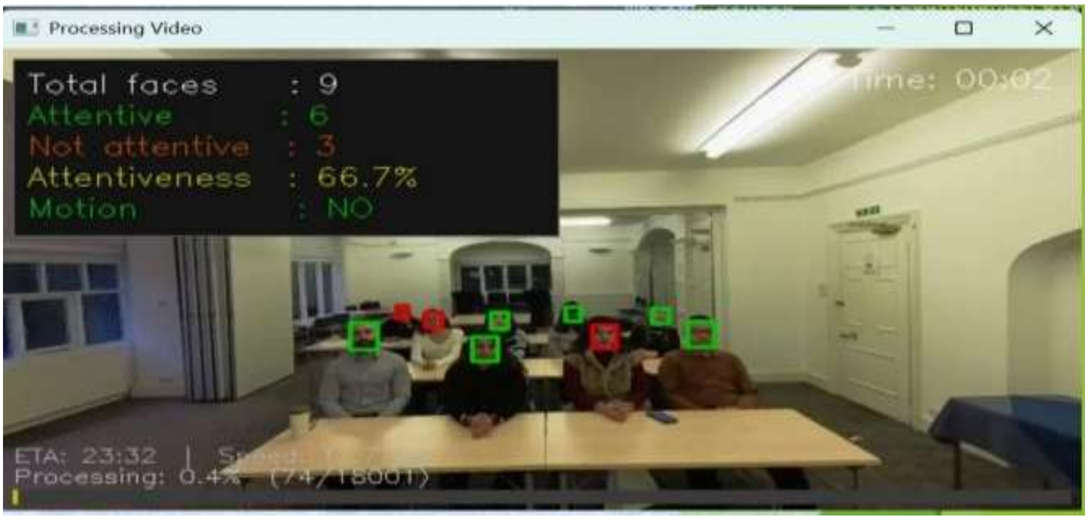

Fig 12: Real-Time Student Attentiveness Monitoring Output

In the visual output, the system clearly distinguishes student attention levels using color-coded bounding boxes. Green boxes represent students who are classified as attentive, indicating that they are focused and engaged with the lecture. In contrast, red boxes highlight students identified as non-attentive, suggesting reduced engagement or distraction. This intuitive visualization allows for quick interpretation of classroom behavior, enabling instructors to easily identify which students are actively participating and which may require additional attention.

Class Attentiveness Over Time:

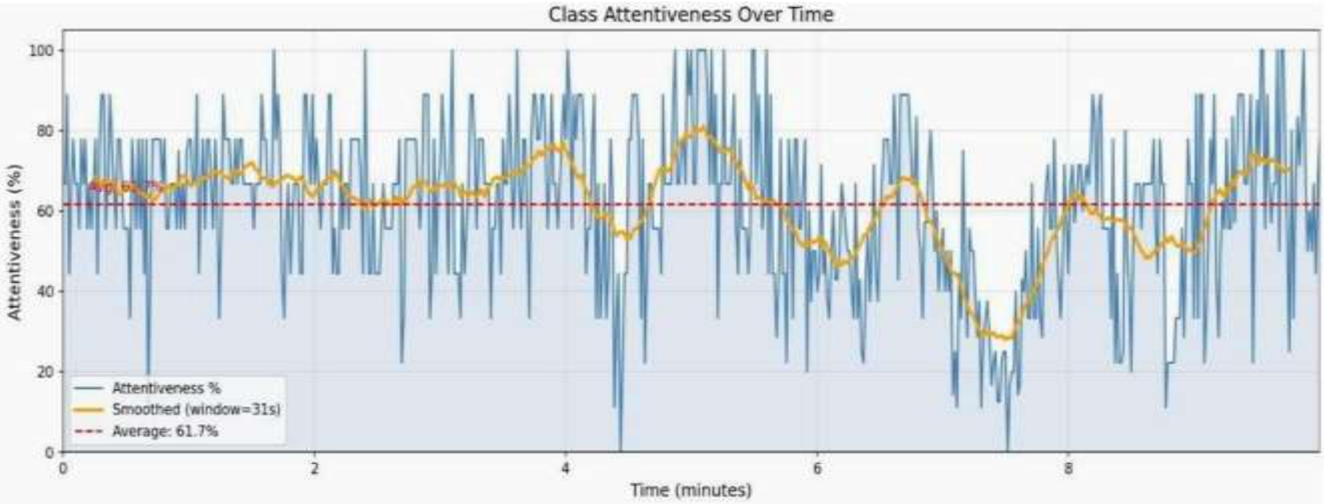

Fig 13: Class Attentiveness Over Time

The class attentiveness over time graph illustrates how student engagement varies throughout the lecture duration. The blue line represents the instantaneous attentiveness percentage calculated for each frame, showing fluctuations due to changes in student behavior. To provide a clearer trend, a smoothed curve is plotted, which highlights the overall pattern attentiveness during the session. The graph also includes an average attentiveness line, indicating the general engagement level of the class. It can be observed that attentiveness rises and falls at different intervals, reflecting moments of high focus and reduced attention, thereby offering valuable insights into classroom dynamics over time.

Furthermore, the analysis of the graph reveals that certain time intervals exhibit noticeable drops in attentiveness, which may correspond to reduced interest, fatigue, or less engaging portions of the lecture. In contrast, periods of increased attentiveness suggest moments where the content is more interactive or engaging for students. These variations highlight the importance of adaptive teaching strategies that can maintain consistent student focus. By examining such trends, instructors can identify critical points during the lecture and modify their delivery methods to improve overall classroom engagement and learning effectiveness.

V. CONCLUSION

This The Student Attentiveness Monitoring System was successfully developed using computer vision and deep learning techniques to analyses classroom engagement. The proposed system utilizes face detection, facial emotion recognition, and head pose estimation to classify students as attentive or non-attentive in real time. By providing an automated and objective approach, the system addresses the limitations of traditional observation methods and enables effective monitoring of student behavior in classroom environments.

Experimental results demonstrate that the system can effectively estimate attentiveness levels and generate meaningful insights such as engagement trends and attentiveness variations over time. However, the performance is influenced by dataset size and diversity, leading to challenges such as overfitting. Future improvements can include the use of larger datasets and integration of additional modalities such as physiological signals to enhance accuracy and robustness.

VI. REFERENCES

[1] Peter K Trabelsi Z., Alnajjar F., Parambil M.M.A., Gochoo M., Ali L., Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student's Behavior Recognition, Big Data and Cognitive Computing, 2023.

[2] N. M. Alruwais and M. Zakariah, "Student Recognition and Activity Monitoring in E-Classes Using Deep Learning in Higher Education," IEEE Access, vol. 12, 2024.

[3] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," IEEE Transactions on Affective Computing, 2019.

[4] M. M. A. Parambil et al., "Smart Classroom: A Deep Learning Approach Towards Attention Assessment Through Class Behavior Detection," in Proc. ASET, 2022.

[5] F. C. Lin, H. H. Ngo, C. R. Dow, K. H. Lam, and H. L. Le, "Student Behavior Recognition System for the Classroom Environment Based on Skeleton Pose Estimation and Person Detection," Sensors, 2021.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. CVPR, 2016.

[7] E. Murphy-Chutorian and M. M. Trivedi, "Head Pose Estimation Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 4, pp. 607–626, 2009.