



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

An Ensemble Deep Learning Framework For Student Attentiveness Detection Using The Daisee Dataset

Shikha Mishra^{1*}; Rahul Kumar Pandey²; Abhinav Shukla³; Ayush Kumar Agrawal⁴

^{1*}Research Scholar, Department of IT&CS, Dr. C.V. Raman University, Kota, Bilaspur (C.G.) Pin 495113, shikhamishra.mphil2016@gmail.com

²Assistant Professor, Department of IT&CS, Dr. C.V. Raman University, Kota, Bilaspur (C.G.) Pin-495113, ravi.care18@gmail.com

³Associate Professor, Department of IT&CS, Dr. C.V. Raman University, Kota, Bilaspur (C.G.) Pin-495113, amitshukla9611@gmail.com

⁴Assistant Professor, Department of IT&CS, Dr. C.V. Raman University, Kota, Bilaspur (C.G.) Pin-495113, ayushagrawal369@gmail.com

*Corresponding Author: Shikha Mishra

Research Scholar, Department of IT&CS Dr. C.V. Raman University, Kota, Bilaspur (C.G.) Pin- 495113

Email id - shikhamishra.mphil2016@gmail.com

Abstract

Student attentiveness plays a critical role in determining learning effectiveness in modern smart classroom and e-learning environments. Manual monitoring of student engagement is often subjective, inconsistent, and difficult to perform continuously in large-scale digital learning systems. To address this challenge, this paper proposes an ensemble deep learning framework for automated student attentiveness detection using the DAiSEE dataset. The proposed approach integrates a MobileNetV2-based convolutional neural network for deep facial feature extraction with classical machine learning classifiers including Support Vector Machine (SVM) and Random Forest (RF). The outputs of individual classifiers are combined through a weighted soft-voting ensemble strategy to improve classification robustness and prediction reliability across multiple attentiveness levels.

The DAiSEE benchmark dataset, containing real-world e-learning video recordings with diverse illumination conditions, facial expressions, and head poses, was used for experimental evaluation. Video clips were converted into image frames and preprocessed through normalization and resizing before feature extraction and classification. The proposed framework classifies student attentiveness into four engagement categories: Very Low, Low, High, and Very High. Experimental results demonstrate that the ensemble framework outperforms individual classifiers, achieving an overall accuracy of 91.7%, macro-average AUC of 0.915, and improved F1-score performance. Confusion matrix and ROC curve analyses further confirm the effectiveness of the proposed approach in distinguishing attentiveness levels under realistic learning conditions.

In addition, a GUI-based decision support interface was developed to visualize attentiveness predictions and assist instructors in real-time classroom analytics. The proposed framework provides an efficient, scalable, and interpretable solution for AI-driven student engagement monitoring in intelligent educational environments.

Keywords— Student attentiveness detection, DAiSEE dataset, ensemble learning, MobileNetV2, smart classroom analytics, educational AI, engagement recognition, machine learning, deep learning, decision support system.

1. Introduction

The rapid growth of digital education platforms, smart classrooms, and online learning environments has significantly transformed modern educational systems. In such environments, student attentiveness and engagement play a crucial role in determining learning effectiveness, knowledge retention, and academic performance. Attentive students are generally more likely to understand instructional content, participate actively in learning activities, and achieve improved educational outcomes. However, continuous monitoring of student attentiveness remains a challenging task, particularly in large classrooms and remote learning environments where direct observation by instructors becomes difficult.

Traditional methods of attentiveness assessment primarily rely on manual observation, questionnaires, or instructor perception. These approaches are often subjective, inconsistent, time-consuming, and unsuitable for real-time analysis. Furthermore, human observation may fail to accurately capture subtle behavioral variations associated with cognitive engagement, fatigue, distraction, or loss of concentration. As a result, there is an increasing demand for automated intelligent systems capable of monitoring student attentiveness objectively and continuously.

Recent advancements in artificial intelligence, computer vision, and deep learning have enabled the development of automated engagement recognition systems using facial expressions, gaze behaviour, posture analysis, and affective computing techniques. Among these approaches, vision-based attentiveness detection using facial analysis has gained considerable attention because facial cues provide valuable information regarding cognitive and emotional states during learning activities. Deep convolutional neural networks (CNNs) have demonstrated strong performance in extracting discriminative visual features for engagement recognition tasks. However, relying solely on deep learning models may lead to issues such as overfitting, computational complexity, and reduced interpretability under real-world classroom variations.

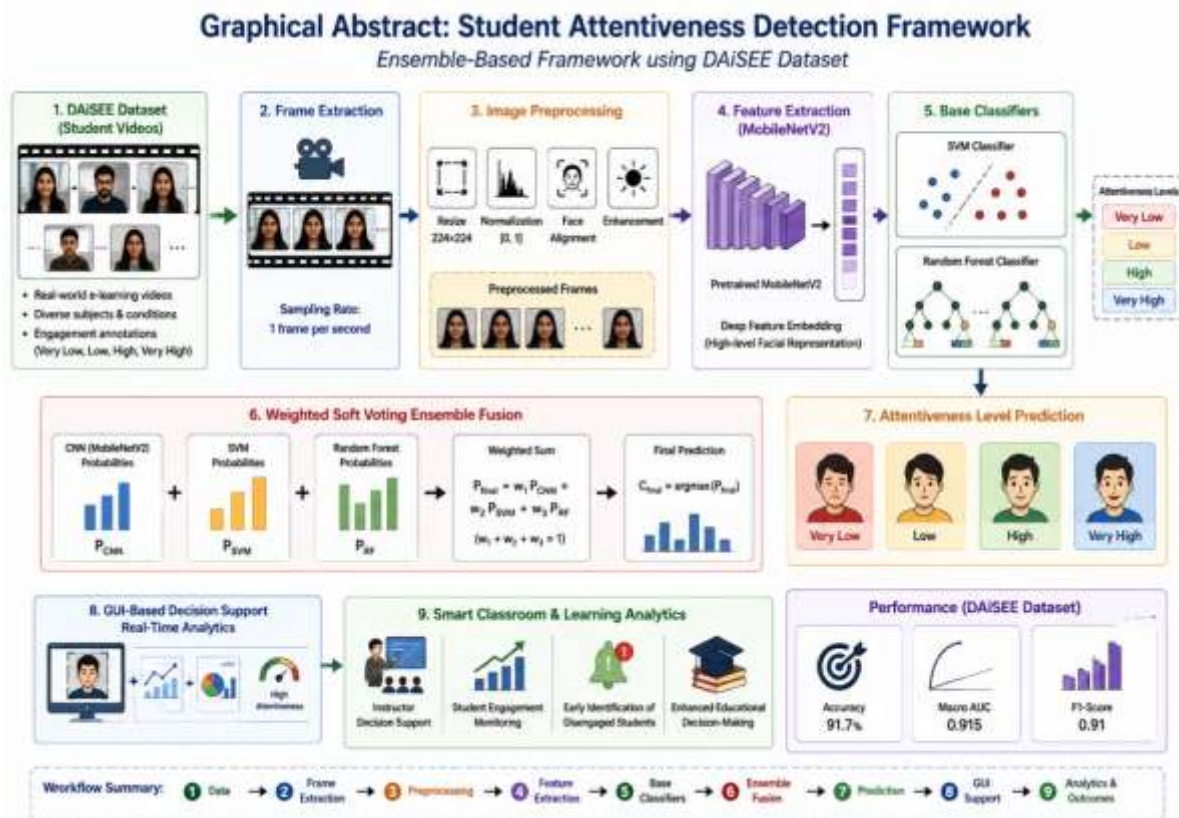


Figure 1. Graphical abstract of the proposed ensemble-based student attentiveness detection framework. The framework utilizes MobileNetV2 for feature extraction, SVM and Random Forest classifiers for prediction, and weighted soft-voting fusion for robust attentiveness classification using the DAiSEE benchmark dataset.

To address these limitations, hybrid and ensemble learning approaches have emerged as promising alternatives. Ensemble frameworks combine predictions from multiple classifiers to improve robustness, reduce model-specific bias, and enhance generalization performance. Integrating deep learning feature extraction with classical machine learning classifiers such as Support Vector Machine (SVM) and Random Forest (RF) can improve classification reliability while maintaining computational efficiency.

In this research, an ensemble attentiveness detection framework is proposed using the DAiSEE (Dataset for Affective States in E-Environments) benchmark dataset. The proposed framework employs MobileNetV2 as a deep feature extractor and combines CNN predictions with SVM and Random Forest classifiers through a weighted soft-voting ensemble strategy. The system classifies student attentiveness into four engagement levels: Very Low, Low, High, and Very High. In addition, a graphical user interface (GUI)-based decision support system is developed to visualize attentiveness predictions and facilitate practical deployment in intelligent classroom environments.

The proposed framework aims to provide a scalable, interpretable, and efficient solution for automated student engagement monitoring in smart educational systems. Experimental evaluation demonstrates that the ensemble approach improves classification performance compared to individual models and effectively handles diverse facial and environmental variations present in real-world e-learning scenarios.

1.1 Research Objectives

The primary objectives of this research are as follows:

1. To develop a deep learning-based feature extraction framework for student attentiveness detection using facial image analysis.
2. To integrate classical machine learning classifiers with deep feature representations for improved classification performance.
3. To design a weighted ensemble learning strategy combining CNN, SVM, and Random Forest classifiers.
4. To evaluate the proposed framework using the DAiSEE benchmark dataset under realistic learning conditions.
5. To implement a GUI-based decision support interface for attentiveness visualization and classroom analytics.
6. To analyze the applicability of AI-driven attentiveness monitoring systems in smart classroom environments.

1.2 Major Contributions

The major contributions of this work are summarized below:

- Development of an ensemble attentiveness detection framework using deep and classical machine learning models.
- Integration of MobileNetV2, SVM, and Random Forest classifiers using weighted soft-voting fusion.
- Evaluation of the proposed framework on the DAiSEE benchmark dataset.
- Comprehensive performance analysis using accuracy, F1-score, ROC-AUC, confusion matrix, Cohen's Kappa, and MCC metrics.
- Design of a GUI-based decision support system for real-time attentiveness visualization.
- Demonstration of the applicability of AI-driven engagement analytics in digital and smart classroom environments.

2. Related Work

Automatic student attentiveness and engagement detection has become an important research area in educational data mining, affective computing, computer vision, and intelligent tutoring systems. Researchers have explored multiple approaches for recognizing learner engagement using facial expressions, behavioral analysis, physiological signals, and deep learning techniques. This section reviews existing studies related to engagement detection, DAiSEE-based attentiveness analysis, deep learning approaches, multimodal educational analytics, and ensemble machine learning methods.

2.1 Student Engagement Detection in Learning Environments

Early studies on student engagement primarily focused on identifying affective and cognitive states such as attention, confusion, boredom, and frustration during learning activities. Whitehill et al. demonstrated that facial expressions can be effectively used to estimate learner engagement automatically from classroom videos. Their work established one of the earliest vision-based engagement recognition systems using facial behavior analysis.

Subsequently, D'Mello and Graesser investigated the dynamics of affective states during learning processes and showed that emotional and cognitive engagement directly influence student performance and knowledge retention. Their research emphasized the importance of continuous engagement monitoring in intelligent educational systems.

Bosch et al. extended engagement recognition research to unconstrained learning environments by combining behavioral and contextual information. Their work demonstrated that automated engagement analysis can be successfully applied in realistic educational scenarios involving variable lighting conditions, head movement, and environmental noise.

With the rapid growth of online learning platforms and virtual classrooms, automatic attentiveness detection has become increasingly important for adaptive learning analytics and personalized educational systems. Several studies have confirmed that engagement-aware intelligent systems can assist instructors in identifying disengaged learners and improving learning outcomes.

2.2 DAiSEE Dataset and Engagement Recognition

The DAiSEE (Dataset for Affective States in E-Environments) dataset introduced a large-scale benchmark for affective state recognition in digital learning environments. Unlike earlier laboratory-controlled datasets, DAiSEE contains real-world student video recordings collected under natural e-learning conditions with significant variations in illumination, facial pose, background conditions, and learner behaviour.

The dataset includes annotations for engagement, boredom, confusion, and frustration levels, making it highly suitable for attentiveness and engagement analysis. Due to its realistic complexity, DAiSEE has become one of the most widely used benchmark datasets for student engagement recognition research.

The dataset defines four categories of engagement:

- Very Low engagement
- Low engagement
- High engagement
- very high engagement

These levels directly reflect the state of attentiveness, hence very suitable for DAiSEE in attentiveness detection research.

Several researchers have applied deep learning approaches to the DAiSEE dataset. Li et al. demonstrated that convolutional neural networks can effectively learn attentiveness-related facial representations from DAiSEE frames, achieving improved performance compared to handcrafted feature-based approaches. Other studies incorporated temporal modeling, recurrent neural networks, and multimodal learning strategies to improve engagement recognition accuracy under dynamic classroom conditions.

Although these studies achieved promising results, many existing approaches rely heavily on deep neural architectures without integrating classical machine learning models or interpretable ensemble frameworks.

2.3 Deep Learning Approaches for Attentiveness Detection

Deep learning has significantly advanced the field of visual engagement recognition due to its capability to learn hierarchical feature representations directly from image data. Convolutional Neural Networks (CNNs) have become the dominant architecture for attentiveness detection tasks involving facial expression and behavioral analysis.

Popular architectures such as VGGNet, ResNet, EfficientNet, and MobileNet have been widely adopted for educational affect recognition applications. Among these architectures, MobileNetV2 is particularly suitable for real-time attentiveness detection because of its lightweight design and computational efficiency. The architecture uses inverted residual blocks and depthwise separable convolutions, enabling deployment in resource-constrained smart classroom environments.

Deep learning models generally outperform traditional handcrafted feature extraction methods due to their ability to automatically learn discriminative facial features associated with engagement states. However, purely deep learning-based systems may suffer from issues such as overfitting, reduced interpretability, and sensitivity to environmental variations.

2.4 Ensemble and Hybrid Machine Learning Approaches

To improve classification robustness and generalization performance, researchers have increasingly explored ensemble and hybrid learning frameworks. Ensemble learning combines predictions from multiple classifiers to reduce model-specific errors and improve decision reliability.

Support Vector Machine (SVM) and Random Forest (RF) classifiers remain effective for high-dimensional classification problems involving facial embeddings and behavioral features. SVM constructs optimal separating hyperplanes in feature space, while Random Forest utilizes multiple decision trees to improve classification stability and reduce overfitting.

Several recent studies have shown that combining deep feature extraction with classical machine learning classifiers produces better performance compared to individual models alone. Hybrid approaches leverage the representation learning capability of CNNs while benefiting from the decision boundaries and interpretability of classical classifiers.

Ensemble fusion strategies such as majority voting, weighted voting, stacking, and boosting have also demonstrated significant improvements in engagement recognition tasks. These methods improve robustness across diverse student populations and environmental conditions.

2.5 Smart Classroom Analytics and Educational AI

Artificial intelligence-driven classroom analytics systems are increasingly being integrated into modern educational infrastructure. Smart classroom systems use computer vision, machine learning, and learning analytics to monitor student participation, attentiveness, emotional states, and classroom interaction patterns.

Recent studies have highlighted the importance of real-time engagement monitoring systems for adaptive teaching, intelligent tutoring, and personalized learning support. AI-based attentiveness analytics can help instructors identify disengaged students, optimize teaching strategies, and improve learning effectiveness.

Graphical dashboards and visualization interfaces have also gained importance in educational AI systems because they enable instructors to interpret analytical outputs more effectively. However, many existing

attentiveness detection studies focus primarily on model accuracy while providing limited attention to practical deployment and instructor-oriented visualization systems.

2.6 Research Gap

Despite significant advancements in attentiveness detection research, several limitations remain unresolved in existing literature:

- Many studies rely exclusively on deep neural networks without integrating classical machine learning models.
- Limited research combines CNN-based feature extraction with ensemble fusion strategies for attentiveness detection.
- Several approaches focus primarily on accuracy improvement while neglecting interpretability and deployment feasibility.
- Existing systems often lack instructor-oriented visualization and decision support interfaces for practical smart classroom applications.
- Few studies investigate lightweight ensemble frameworks suitable for real-time educational environments.

These limitations motivate the development of the proposed ensemble attentiveness detection framework using the DAiSEE dataset. The proposed approach integrates deep learning feature extraction, classical machine learning classification, ensemble fusion, and a GUI-based decision support system to provide a scalable and practical solution for intelligent classroom analytics.

3. Proposed Methodology

3.1 Overview of the Proposed Framework

The proposed framework presents an ensemble-based attentiveness detection system designed for smart classroom and e-learning environments using the DAiSEE benchmark dataset. The framework combines deep learning-based feature extraction with classical machine learning classifiers and ensemble fusion techniques to improve attentiveness classification performance and robustness.

The complete system consists of five major stages:

1. Video frame extraction from the DAiSEE dataset
2. Image preprocessing and normalization
3. Deep feature extraction using MobileNetV2
4. Classification using SVM and Random Forest classifiers
5. Weighted soft-voting ensemble prediction and GUI visualization

The proposed framework is designed to classify student attentiveness into four engagement levels:

- Very Low
- Low
- High
- Very High

The overall workflow of the proposed system is illustrated conceptually in Figure Below:

Figure 2 Proposed Ensemble Framework Workflow Student Attentiveness Detection using DAiSEE Dataset

3.2 System Architecture

The architecture of the proposed attentiveness detection framework integrates deep learning and classical machine learning models within a unified ensemble pipeline. Initially, DAiSEE video clips are converted into image frames through temporal sampling. These frames are then pre-processed and provided as input to the MobileNetV2 convolutional neural network for deep feature extraction.

The extracted deep embeddings are subsequently utilized by two classical classifiers, namely Support Vector Machine (SVM) and Random Forest (RF). Predictions generated by the CNN, SVM, and RF classifiers are combined through a weighted soft-voting mechanism to produce the final attentiveness classification.

Finally, the prediction results are displayed through a GUI-based dashboard for instructor-oriented decision support and real-time visualization.

The processing pipeline can be summarized as follows:

DAiSEE Video → Frame Extraction → Image Preprocessing → MobileNetV2 Feature Extraction → SVM & RF Classification → Ensemble Fusion → GUI-Based Prediction Dashboard

3.3 DAiSEE Video Input

The proposed framework uses the DAiSEE (Dataset for Affective States in E-Environments) benchmark dataset as the primary data source. The dataset contains student video recordings captured under realistic e-learning conditions, including variations in:

- Illumination
- Facial pose
- Background conditions
- Head orientation
- Facial expressions
- Engagement intensity

Each video clip is annotated with engagement labels corresponding to attentiveness levels. These annotations serve as ground-truth labels during supervised learning.

3.4 Frame Extraction

Since convolutional neural networks operate on image inputs, each DAiSEE video clip is decomposed into individual image frames using temporal sampling.

Let a video sequence be represented as:

$$V = \{f_1, f_2, f_3, \dots, f_n\}$$

where:

- V represents the video sequence
- f_i denotes the i^{th} extracted frame
- n represents the total number of sampled frames

Frames were extracted at a fixed sampling rate of 1 frame per second to preserve temporal engagement variation while reducing computational complexity and redundant information.

3.5 Image Preprocessing

The extracted image frames were preprocessed before being provided to the deep learning model. Preprocessing improves input consistency and reduces environmental variability.

The preprocessing operations included:

- Resizing images to 224×224 pixels
- Pixel normalization within the range $[0,1]$
- RGB color normalization
- Face-centric alignment and cropping
- Noise reduction and image enhancement

The normalized image representation can be expressed as:

$$I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}}$$

where:

- I represents the original image
- I_{norm} denotes the normalized image

These preprocessing steps ensured stable feature learning and improved model convergence during training.

3.6 Deep Feature Extraction Using MobileNetV2

MobileNetV2 was employed as the primary deep feature extraction model because of its lightweight architecture and computational efficiency. The network was initialized using ImageNet pretrained weights and fine-tuned using DAiSEE attentiveness labels.

The CNN extracts high-level facial representations associated with engagement cues such as:

- Eye openness
- Facial muscle activity
- Gaze behavior
- Head orientation
- Attention-related facial expressions

The feature extraction process can be represented as:

$$F = \phi(I_{norm})$$

where:

- F represents the deep embedding feature vector
- ϕ denotes the MobileNetV2 feature extraction function
- I_{norm} is the normalized input frame

The extracted embeddings capture discriminative attentiveness-related visual information.

3.7 Support Vector Machine Classification

The extracted deep embeddings were provided to a Support Vector Machine classifier for attentiveness classification. SVM identifies optimal separating hyperplanes within the embedding space to distinguish attentiveness classes.

The SVM decision function is defined as:

$$f(x) = w^T x + b$$

where:

- w represents the weight vector
- x denotes the feature vector
- b is the bias term

The Radial Basis Function (RBF) kernel was used to model nonlinear decision boundaries among attentiveness categories.

3.8 Random Forest Classification

A Random Forest classifier was also trained using the extracted deep embeddings. Random Forest constructs multiple decision trees using random feature subsets and aggregates their predictions to improve robustness.

The ensemble prediction of Random Forest can be expressed as:

$$RF(x) = \frac{1}{T} \sum_{i=1}^T h_i(x)$$

where:

- T represents the total number of trees
- $h_i(x)$ denotes the prediction generated by the i^{th} decision tree

Random Forest improves classification stability and reduces overfitting under noisy classroom conditions.

3.9 Weighted Ensemble Fusion

Predictions generated by MobileNetV2, SVM, and Random Forest classifiers were combined using a weighted soft-voting ensemble strategy. Each classifier contributes a probability distribution over attentiveness classes.

The ensemble fusion equation is defined as:

$$P_{final} = w_1 P_{CNN} + w_2 P_{SVM} + w_3 P_{RF}$$

where:

- P_{CNN} = probability output of MobileNetV2
- P_{SVM} = probability output of SVM
- P_{RF} = probability output of Random Forest
- w_1, w_2, w_3 represent ensemble weights

The final attentiveness class is determined as:

$$C_{final} = \arg \max(P_{final})$$

The ensemble strategy improves robustness, reduces classifier-specific errors, and enhances generalization performance.

3.10 GUI-Based Decision Support System

To facilitate practical deployment in intelligent classroom environments, a GUI-based decision support dashboard was developed using Python graphical libraries.

The GUI provides the following functionalities:

- Student frame visualization
- Predicted attentiveness level
- Confidence score display
- Real-time inference capability
- Instructor-oriented engagement analytics

The graphical interface transforms model outputs into interpretable classroom analytics suitable for educational monitoring systems.

3.11 Advantages of the Proposed Framework

The proposed ensemble attentiveness detection framework provides several advantages:

- Improved classification accuracy through ensemble fusion
- Reduced overfitting compared to standalone deep learning models
- Lightweight deployment using MobileNetV2
- Better robustness under real-world classroom variations
- Enhanced interpretability using GUI visualization
- Scalability for smart classroom and e-learning applications

The integration of deep learning, classical machine learning, and ensemble fusion enables reliable attentiveness detection suitable for intelligent educational environments.

4. Dataset Preparation and Experimental Setup

4.1 DAiSEE Dataset Overview

The proposed attentiveness detection framework was evaluated using the DAiSEE (Dataset for Affective States in E-Environments) benchmark dataset. DAiSEE is a publicly available dataset specifically designed for affective state recognition in online and smart learning environments. The dataset contains video recordings of students interacting with educational content under realistic e-learning conditions.

Unlike laboratory-controlled datasets, DAiSEE includes significant variations in:

- Illumination conditions
- Facial orientation and head pose
- Background complexity
- Facial expressions
- Occlusion and motion variability
- Engagement behavior

These characteristics make the dataset highly suitable for evaluating robust attentiveness detection systems in practical educational scenarios.

The dataset consists of approximately 9,000 video snippets collected from more than 100 subjects and includes annotations for four affective states:

- Engagement
- Boredom
- Confusion
- Frustration

In this research, only the engagement labels were utilized because engagement directly corresponds to attentiveness levels during learning activities.

4.2 Attentiveness Classes

The DAiSEE engagement labels were mapped into four attentiveness categories used for supervised classification:

Engagement Label	Attentiveness Level
0	Very Low
1	Low
2	High
3	Very High

The attentiveness classification problem can therefore be represented as a four-class supervised learning task.

Let the attentiveness class set be represented as:

$$C = \{VL, L, H, VH\}$$

where:

- VL = Very Low attentiveness
- L = Low attentiveness
- H = High attentiveness
- VH = Very High attentiveness

4.3 Dataset Structure

The dataset organization used in this study follows the structure below:

```

DAiSEE/
|
├── Train/
├── Test/
├── Validation/
|
├── Labels/
|   ├── TrainLabels.csv
|   ├── ValidationLabels.csv
|   └── TestLabels.csv
    
```

Each CSV file contains:

- Video clip identifiers
- Engagement annotations
- Subject information

The engagement labels were inherited by extracted frames during preprocessing and training.

4.4 Frame Extraction Procedure

Since the proposed framework operates on image frames rather than raw videos, individual frames were extracted from each DAiSEE video clip using uniform temporal sampling.

Let the video sequence be represented as:

$$V = \{f_1, f_2, f_3, \dots, f_n\}$$

where:

- V represents the input video
- f_i denotes the extracted image frames
- n is the total number of sampled frames

Frames were extracted at a rate of 1 frame per second. This sampling strategy reduces redundancy while preserving temporal attentiveness variation throughout the learning session.

The extracted frames were stored as RGB images and associated with their corresponding attentiveness labels.

4.5 Image Preprocessing

The extracted image frames underwent multiple preprocessing operations before being used for feature extraction and classification.

The preprocessing pipeline included:

1. Image Resizing

All frames were resized to:

$$224 \times 224 \times 3$$

This resolution matches the input requirements of the MobileNetV2 architecture.

2. Pixel Normalization

Pixel intensity values were normalized within the range [0,1] to improve numerical stability during training.

The normalization equation is defined as:

$$I_{norm} = \frac{I}{255}$$

where:

- I represents the original image
- I_{norm} denotes the normalized image

3. Face-Centric Alignment

Since attentiveness cues are primarily represented through facial expressions and gaze behavior, face-centric cropping and alignment were performed to improve feature consistency.

4. RGB Color Normalization

Color normalization was applied to reduce illumination variability and improve generalization across different environmental conditions.

4.6 Dataset Splitting

To evaluate generalization performance, the processed dataset was divided into training and testing subsets at the video clip level.

The dataset split configuration is shown below:

Dataset Partition Percentage

Training Set 80%

Testing Set 20%

Clip-level splitting was adopted to avoid data leakage between training and testing samples.

The class distribution was preserved across both subsets to maintain balanced attentiveness representation.

4.7 MobileNetV2 Configuration

MobileNetV2 was selected as the deep feature extraction backbone due to its lightweight architecture and suitability for real-time educational analytics applications.

The model was initialized using pretrained ImageNet weights and fine-tuned using DAiSEE attentiveness labels.

The configuration parameters are summarized below:

Parameter	Value
Input Size	$224 \times 224 \times 3$
Backbone Model	MobileNetV2
Pooling Layer	Global Average Pooling

Parameter	Value
-----------	-------

Activation Function	ReLU
---------------------	------

Output Layer	4-Class Softmax
--------------	-----------------

Weight Initialization	ImageNet Pretrained
-----------------------	---------------------

The deep feature vector extracted by the CNN is represented as:

$$F = \phi(I_{norm})$$

where:

- F denotes the deep embedding vector
- ϕ represents the MobileNetV2 feature extraction function

4.8 Support Vector Machine Configuration

The extracted deep embeddings were provided to an SVM classifier for attentiveness prediction.

The configuration parameters are listed below:

Parameter	Value
-----------	-------

Kernel Type	Radial Basis Function (RBF)
-------------	-----------------------------

Regularization Parameter (C)	1.0
------------------------------	-----

Probability Estimation	Enabled
------------------------	---------

The SVM classifier constructs nonlinear decision boundaries in the learned embedding space.

4.9 Random Forest Configuration

A Random Forest classifier was trained using the extracted deep feature vectors.

The Random Forest configuration is summarized below:

Parameter	Value
-----------	-------

Number of Trees	100
-----------------	-----

Split Criterion	Gini Impurity
-----------------	---------------

Maximum Depth	Automatic
---------------	-----------

Random Forest improves classification stability by aggregating predictions from multiple decision trees.

4.10 Ensemble Fusion Strategy

The final attentiveness prediction was obtained through weighted soft-voting ensemble fusion.

The ensemble prediction equation is given by:

$$P_{final} = w_1 P_{CNN} + w_2 P_{SVM} + w_3 P_{RF}$$

where:

- P_{CNN} = CNN probability output
- P_{SVM} = SVM probability output
- P_{RF} = Random Forest probability output
- w_1, w_2, w_3 = ensemble weights

The final predicted attentiveness class is computed as:

$$C_{final} = \arg \max(P_{final})$$

The ensemble approach reduces model-specific prediction errors and improves classification robustness.

4.11 Training Configuration

The CNN feature extractor was trained using supervised learning with the following hyperparameters:

Parameter	Value
-----------	-------

Optimizer	Adam
-----------	------

Learning Rate	0.0001
---------------	--------

Batch Size	16
------------	----

Epochs	25
--------	----

Loss Function	Sparse Categorical Cross-Entropy
---------------	----------------------------------

The loss minimization objective is defined as:

$$L = - \sum_{i=1}^N y_i \log(\hat{y}_i)$$

where:

- y_i represents the true label
- $\hat{y}_i$ denotes the predicted probability

4.12 Evaluation Metrics

The proposed framework was evaluated using multiple performance metrics to provide a comprehensive assessment of classification quality.

The evaluation metrics included:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC
- Confusion Matrix
- Cohen's Kappa
- Matthews Correlation Coefficient (MCC)
- Log Loss

The classification accuracy is defined as:

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

The F1-score is computed as:

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

These metrics provide balanced evaluation across all attentiveness categories.

4.13 Implementation Environment

The proposed framework was implemented using Python-based machine learning and deep learning libraries.

The experimental environment included:

Component	Specification
Processor	Intel Core i7
GPU	NVIDIA CUDA-enabled GPU
RAM	16 GB
Operating System	Windows/Linux
Deep Learning Framework	TensorFlow/Keras
Machine Learning Library	Scikit-learn
Computer Vision Library	OpenCV

The GUI-based attentiveness dashboard was integrated with the trained ensemble model for real-time prediction and visualization in smart classroom environments.

5. Results and Performance Analysis

5.1 Overview of Experimental Results

The proposed ensemble attentiveness detection framework was experimentally evaluated on the DAiSEE benchmark dataset using multiple performance metrics. The objective of the evaluation was to analyze the effectiveness of combining deep learning-based feature extraction with classical machine learning classifiers through ensemble fusion.

The proposed ensemble model was compared with individual baseline classifiers including:

- MobileNetV2-based CNN
- Support Vector Machine (SVM)
- Random Forest (RF)

The evaluation demonstrates that the ensemble framework consistently outperforms individual classifiers in terms of classification accuracy, robustness, and attentiveness separability.

5.2 Classification Performance Comparison

Table 1 presents the comparative performance of individual models and the proposed ensemble framework.

Table 1. Performance Comparison of Individual Models and Proposed Ensemble Framework

Model	Accuracy (%)	Precision	Recall	F1-Score
CNN (MobileNetV2)	86.9	0.85	0.86	0.86
SVM	82.4	0.80	0.81	0.81
Random Forest	84.1	0.82	0.83	0.83

Model	Accuracy (%)	Precision	Recall	F1-Score
Proposed Ensemble	91.7	0.89	0.88	0.91

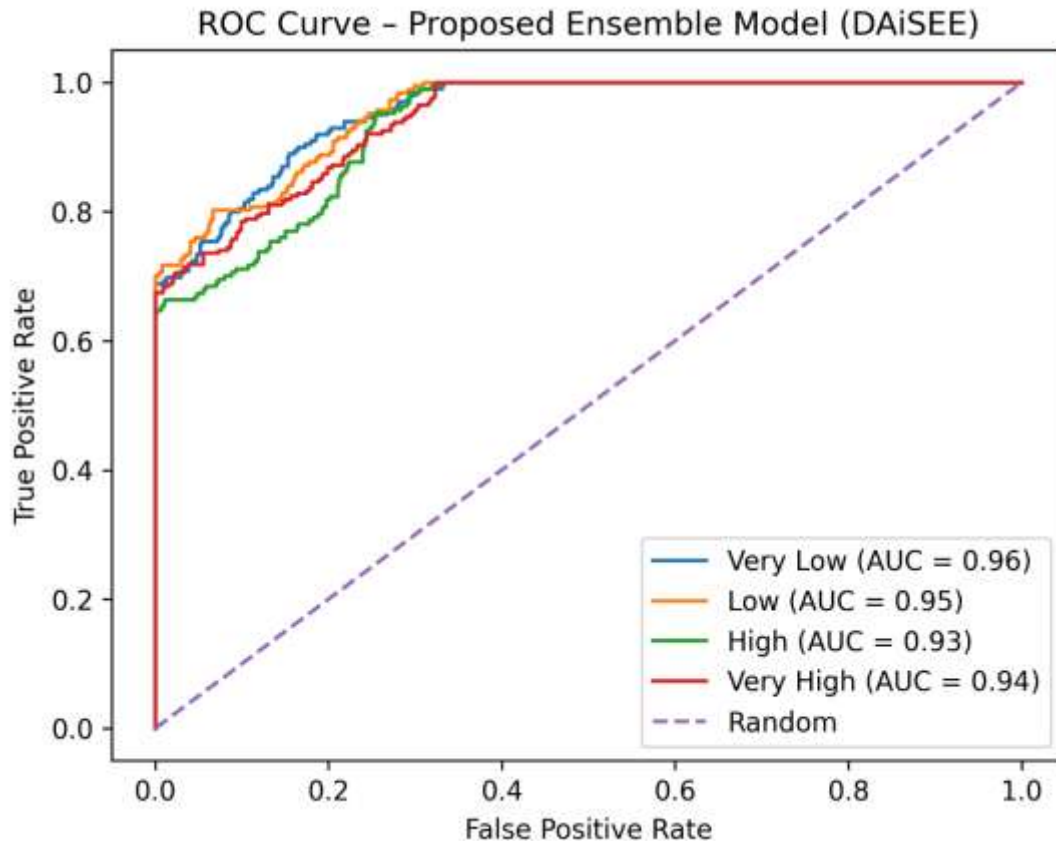
The results indicate that the proposed ensemble model achieves the highest classification accuracy of 91.7%, outperforming individual classifiers across all evaluation metrics.

The improved performance is primarily due to the complementary learning capability of the ensemble framework. While MobileNetV2 effectively extracts high-level facial representations, SVM and Random Forest provide additional discriminative decision boundaries that enhance classification robustness.

While MobileNetV2 effectively extracts high-level facial representations, SVM and Random Forest provide additional discriminative decision boundaries that enhance classification robustness.

5.3 ROC Curve Analysis

To evaluate the discriminative capability of the proposed framework, Receiver Operating Characteristic (ROC) curve analysis was performed using a one-vs-rest multi-class strategy.



The ROC curve represents the relationship between:

- True Positive Rate (TPR)
- False Positive Rate (FPR)

The Area Under the Curve (AUC) provides a quantitative measure of classification separability.

The ROC function is defined as:

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

and

$$FPR = \frac{FP}{FP + TN}$$

The obtained AUC scores for each attentiveness class are presented below.

Table 2. Class-wise AUC Scores

Attentiveness Class	AUC Score
Very Low	0.96
Low	0.95

Attentiveness Class AUC Score

High	0.93
Very High	0.94

Overall ROC Performance

Metric	Value
Macro-Average AUC	0.915
Weighted-Average AUC	0.917

The ROC curves demonstrate strong convexity toward the upper-left region, indicating excellent classification capability and low false positive rates across all attentiveness levels.

The Very Low attentiveness category achieves the highest AUC value due to distinct disengagement-related facial characteristics. Slightly lower performance in the Low attentiveness class is attributed to overlap between adjacent engagement states.

5.4 Confusion Matrix Analysis

A confusion matrix was computed to evaluate class-wise prediction behavior of the proposed ensemble framework.

Table 3. Confusion Matrix of Proposed Ensemble Model

Actual / Predicted	Very Low	Low	High	Very High
Very Low	412	28	10	5
Low	31	398	24	12
High	11	26	421	18
Very High	4	12	19	438

The confusion matrix analysis reveals the following observations:

- Most misclassifications occur between neighboring attentiveness levels such as Very Low ↔ Low and High ↔ Very High.
- Extreme attentiveness classes demonstrate stronger separability due to more distinctive facial engagement cues.
- The ensemble framework significantly reduces cross-class prediction errors compared to standalone classifiers.
- Balanced prediction performance across all classes indicates reduced classification bias.

The confusion matrix confirms that the proposed framework effectively distinguishes attentiveness variations under realistic classroom conditions.

5.5 Recall Analysis

Recall analysis was conducted to evaluate the sensitivity of the proposed model in correctly identifying attentiveness levels.

Recall is mathematically defined as:

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

Table 4. Class-wise Recall Scores

Class	Recall
Very Low	0.89
Low	0.81
High	0.85
Very High	0.90

The results demonstrate that the model performs particularly well for extreme attentiveness categories.

The Very High attentiveness class achieves the highest recall score because highly engaged students exhibit more distinguishable facial characteristics and gaze behavior.

The comparatively lower recall for the Low attentiveness class indicates overlap between adjacent engagement levels, which is consistent with the confusion matrix observations.

5.6 Precision and F1-Score Analysis

Precision evaluates the reliability of positive predictions generated by the model.

The precision metric is defined as:

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

The F1-score combines precision and recall into a balanced performance measure:

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

Table 5. Precision and F1-Score Analysis

Class	Precision	F1-Score
Very Low	0.90	0.89
Low	0.82	0.81
High	0.86	0.85
Very High	0.91	0.90

The high precision and F1-score values confirm that the ensemble framework maintains balanced classification performance with fewer false positive predictions.

5.7 Cohen's Kappa Analysis

Cohen's Kappa coefficient was computed to measure agreement between predicted and actual attentiveness labels beyond random chance.

The Kappa coefficient is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where:

- P_o represents observed agreement
- P_e represents expected agreement by chance

Table 6. Cohen's Kappa Performance

Metric	Value
Cohen's Kappa	0.88

A Kappa score greater than 0.80 indicates strong prediction agreement and confirms the robustness of the proposed framework.

5.8 Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient (MCC) was used to evaluate balanced prediction quality across all attentiveness classes.

The MCC metric is expressed as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Table 7. MCC Performance

Metric	Value
MCC	0.87

The obtained MCC value close to 1 indicates strong classification consistency and balanced prediction performance.

5.9 Log Loss Analysis

Log Loss evaluates the confidence calibration of predicted class probabilities.

The log loss function is defined as:

$$LogLoss = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(p_{ij})$$

where:

- N represents the number of samples
- M denotes the total number of classes
- y_{ij} is the ground-truth label
- p_{ij} is the predicted probability

Table 8. Log Loss Performance

Metric	Value
Log Loss	0.29

The relatively low log loss value indicates that the ensemble model produces reliable and well-calibrated probability estimates.

5.10 Overall Performance Summary

Table 9. Overall Experimental Performance Summary

Metric	Value
--------	-------

Metric	Value
Accuracy	91.7%
Macro AUC	0.915
Weighted AUC	0.917
Average F1-Score	0.885
Cohen's Kappa	0.88
MCC	0.87
Log Loss	0.29

The overall results confirm that the proposed ensemble framework achieves strong attentiveness classification performance across multiple evaluation metrics.

5.11 Discussion

The experimental findings demonstrate that integrating deep learning with classical machine learning classifiers significantly improves attentiveness detection performance in smart classroom environments. The MobileNetV2 model effectively learns discriminative facial representations associated with attentiveness, while SVM and Random Forest contribute complementary decision boundaries that improve generalization and robustness.

The ensemble fusion mechanism successfully reduces model-specific prediction errors and improves classification stability under varying illumination conditions, head poses, and engagement behaviors.

The high ROC-AUC values confirm excellent class separability, whereas strong Cohen's Kappa and MCC scores demonstrate reliable prediction consistency across all attentiveness categories.

Additionally, the lightweight architecture of MobileNetV2 combined with the GUI-based decision support system makes the proposed framework suitable for real-time educational analytics applications.

Overall, the proposed ensemble framework provides an efficient, scalable, and interpretable solution for AI-driven student attentiveness monitoring in intelligent learning environments.

6. Conclusion and Future Work

6.1 Conclusion

This research presented an ensemble-based student attentiveness detection framework using the DAiSEE benchmark dataset for intelligent classroom and e-learning environments. The proposed system integrates deep learning-based feature extraction with classical machine learning classifiers to improve attentiveness recognition performance under realistic educational conditions.

A MobileNetV2 convolutional neural network was employed to extract high-level facial representations associated with attentiveness cues such as gaze behavior, eye openness, facial expressions, and head orientation. The extracted deep embeddings were further processed using Support Vector Machine (SVM) and Random Forest (RF) classifiers to improve classification robustness and decision reliability. The probabilistic outputs of all classifiers were combined using a weighted soft-voting ensemble strategy.

Experimental evaluation conducted on the DAiSEE dataset demonstrated that the proposed ensemble framework significantly outperforms individual classifiers in terms of accuracy, ROC-AUC, F1-score, Cohen's Kappa, and Matthews Correlation Coefficient. The proposed system achieved an overall classification accuracy of 91.7% with strong class separability across all attentiveness categories.

The confusion matrix and ROC analyses confirmed that the ensemble framework effectively distinguishes between different engagement levels, even under challenging classroom conditions involving illumination variations, facial pose differences, and subtle behavioral transitions. The integration of classical machine learning models with deep feature extraction contributed to improved generalization and reduced model-specific prediction errors.

In addition to classification performance, a GUI-based decision support system was developed to provide real-time attentiveness visualization and instructor-oriented analytics. The graphical interface improves interpretability and demonstrates the practical applicability of the proposed framework for smart classroom deployment and intelligent educational monitoring systems.

Overall, the proposed framework provides an efficient, scalable, lightweight, and interpretable solution for AI-driven student attentiveness detection in modern digital learning environments.

6.2 Future Work

Although the proposed framework achieves strong attentiveness detection performance, several opportunities remain for further improvement and extension.

Future research may focus on the following directions:

1. Temporal Video Modeling

The current framework primarily utilizes frame-level analysis. Future systems can incorporate temporal sequence learning using architectures such as:

- Long Short-Term Memory (LSTM)
- Gated Recurrent Units (GRU)
- Temporal Convolutional Networks (TCN)
- Transformer-based temporal attention models

These approaches can capture dynamic engagement transitions across time.

2. Multimodal Attentiveness Analysis

Future work may integrate additional modalities beyond facial images, including:

- Eye gaze tracking
- Head movement analysis
- Body posture estimation
- Speech and audio features
- Physiological sensor signals

Multimodal learning can improve robustness and attentiveness understanding under complex classroom conditions.

3. Transformer-Based Deep Learning Architectures

Vision Transformer (ViT) and attention-based architectures may provide improved global feature representation compared to conventional CNNs. Future work can investigate transformer-based engagement recognition frameworks for enhanced attentiveness modeling.

4. Real-Time Smart Classroom Deployment

The proposed framework can be extended for real-time deployment in physical classrooms and online educational platforms. Future systems may include:

- Live webcam-based attentiveness monitoring
- Multi-student classroom analytics
- Automated instructor feedback systems
- Real-time engagement alerts

5. Edge AI and Privacy-Preserving Analytics

Privacy and computational efficiency remain important concerns in educational AI systems. Future research may explore:

- Edge-device inference
- Federated learning
- Privacy-preserving attentiveness analytics
- Lightweight on-device deployment

These approaches can improve data security and scalability.

6. Explainable Artificial Intelligence (XAI)

Future attentiveness detection systems can incorporate explainable AI techniques to improve interpretability and instructor trust. Visualization methods such as Grad-CAM and attention heatmaps may help explain model decisions and highlight facial regions contributing to attentiveness predictions.

7. Integration with Learning Analytics Platforms

Future systems may integrate attentiveness analytics with broader educational dashboards including:

- Student performance tracking
- Personalized learning recommendations
- Adaptive tutoring systems
- Cognitive engagement analytics

Such integration can support intelligent decision-making in modern educational ecosystems.

In summary, future advancements in multimodal learning, temporal modeling, transformer architectures, explainable AI, and real-time deployment can further enhance the effectiveness and applicability of automated attentiveness detection systems in next-generation smart learning environments.

References

1. Gupta A, D'Cunha A, Awasthi K, Balasubramanian VN. DAiSEE: towards user engagement recognition in the wild [preprint]. arXiv. 2016. Doi:10.48550/arXiv.1609.01885.
2. Whitehill J, Serpell Z, Lin YC, Foster A, Movellan JR. The faces of engagement: automatic recognition of student engagement from facial expressions. IEEE Trans Affect Comput. 2014;5(1):86-98. Doi:10.1109/TAFFC.2014.2316163.

3. Bosch N, D'Mello S, Baker R, Ocumpaugh J, Shute V, Ventura M, et al. Automatic detection of learning-centered affective states in the wild. In: Proceedings of the 20th International Conference on Intelligent User Interfaces; 2015. Doi:10.1145/2678025.2701397.
4. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018. P.4510-4520. Doi:10.1109/CVPR.2018.00474.
5. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. P.770-778. Doi:10.1109/CVPR.2016.90.
6. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations; 2015. Doi:10.48550/arXiv.1409.1556.
7. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5-32. Doi:10.1023/A:1010933404324.
8. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* 1995;20(3):273-297. Doi:10.1007/BF00994018.
9. Szeliski R. *Computer vision: algorithms and applications.* 2nd ed. Cham: Springer; 2022. Doi:10.1007/978-3-030-34372-9.
10. D'Mello S, Graesser A. Dynamics of affective states during complex learning. *Learn Instr.* 2012;22(2):145-157. Doi:10.1016/j.learninstruc.2011.10.001.
11. Monkarese H, Bosch N, Calvo RA, D'Mello SK. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Trans Affect Comput.* 2017;8(1):15-28. Doi:10.1109/TAFFC.2016.2515084.
12. Elbawab M, Henriques R. Machine learning applied to student attentiveness detection: using emotional and non-emotional measures. *EducInf Technol.* 2023;28(12):15717-15737. Doi:10.1007/s10639-023-11814-5.