Student Engagement Analysis using Computer Vision

Project Team

Student 1 22I-1148
Student 2 22I-0788
Student 3 22I-0962

Session 2022-2026

Supervised by

Dr Qaisar Shafi

Co-Supervised by

Sir Adil-ur-rahman

Department of Computer Science

National University of Computer and Emerging Sciences
Islamabad, Pakistan

June, 2022

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the research project titled AI-Based Student Engagement Analysis. It establishes the contextual foundation, explains the purpose and objectives, and defines the problem domain and specific research problem being addressed. The chapter also outlines the system's intended functionality and its relevance within modern educational technology.

1.1 Problem Domain

The education sector is transforming with AI integration in classrooms, yet measuring student engagement—a key driver of academic success—remains challenging using traditional, subjective observation methods. Computer vision and deep learning enable automated, real-time analysis of facial expressions, gestures, and behaviors, but existing systems often limit focus to emotions, ignoring posture, hand movements, and attentiveness etc. This study bridges AI, affective computing, and educational psychology to deliver multimodal feedback on emotional, cognitive, and behavioral engagement, enabling data-driven, inclusive teaching through lecture-to-lecture analysis for teachers.

1.2 Research Problem Statement

Current AI-based classroom analytics suffer from key limitations: over-reliance on facial expressions, ignoring gaze, posture, and gestures; the majority offer only binary engagement classifications, limited to engaged and disengaged states. Moreover, they rarely support lecture-to-lecture analysis on a student-by-student basis, hindering the tracking of individual progress and long-term trends. This research develops a multimodal AI system integrating emotion recognition and action detection for comprehensive engagement analysis from classroom videos, supporting lecture-to-lecture comparisons. It provides multi-modal, individualized insights—including lecture-to-lecture analysis per student—to help educators identify engagement and disengagement, track personalized trends, and refine strategies, advancing educational AI while enabling teachers to adapt and improve their strategies for the betterment of students.

Chapter 2

Literature Review

This chapter critically examines existing literature relevant to AI-based student engagement analysis, highlighting key studies, their methodologies, contributions, and limitations. Each review concludes by connecting the findings to the proposed multimodal engagement detection system.

2.1 Related Research

The following research items reflect foundational and state-of-the-art contributions in automated student engagement detection. For each item, a summary, critical analysis, and relevance to the present project are provided.

2.1.1 Detecting Student Engagement with CNN and FER[2]

Summary: Alruwais and Zakariah (2025) present a CNN model trained on the FER-2013 dataset (48$\times$48 grayscale images, augmented), which recognizes static facial expressions tied to student engagement.

Critical analysis: Strengths include effective use of a standard facial-expression dataset and a robust CNN pipeline. The main limitations are the focus on static facial data only, absence of body posture or gesture analysis, and lack of real-time feedback.

Relation to proposed work: This study's approach to facial-expression classification informs the emotion stream within our multimodal fusion system, which is further extended to include dynamic and behavioral inputs.

2.1.2 Measuring Student Engagement through Behavioral and Emotional Features[5]

Summary: Mahmood et al. (2024) design a hybrid deep-learning system using ResNet-50 for feature extraction and TCN for temporal analysis, working on video clips from classrooms for engagement detection.

Critical analysis: The model's strength lies in its combination of spatial and temporal features, allowing robust analysis of video sequences. However, it is limited by basic emotion categories, and excludes detailed gestures, postures, and nuanced affective states.

Relation to proposed work: Their fusion of behavioral and temporal cues motivates the proposed use of TCNs for behavioral streams, which will be integrated with our emotion network.

2.1.3 Improving Student Engagement Detection with ResNet + TCN Hybrid[1]

Summary: Abedi and Khan (2021) introduce a multimodal engagement metric, combining behavioral gating and emotional fusion—using transfer learning approaches (ResNet50/VGG16/Inceptic

Critical analysis: While leveraging multiple pretrained models and introducing an innovative engagement metric, the model's sensitivity to subtle cues (e.g., note-taking, posture shifts) is limited.

Relation to proposed work: The multimodal metric concept influences our own design, which explicitly includes fine-grained behavioral cues for richer student analysis.

2.1.4 Multimodal Engagement Analysis from Facial Videos[11]

Summary: This work deploys a ResNet-50 Attention-Net with Affect-Net on classroom video data, performing multi-label engagement classification through advanced deep learning.

Critical analysis: Strengths involve deployment on real classroom datasets and the use of multi-label output. Limitations are a persistent focus on facial features and testing in limited settings.

Relation to proposed work: We expand beyond facial-only models, adding body posture, gesture, and action detection to target multimodal fusion and real-classroom scenarios.

2.1.5 Student Engagement Detection Using YOLOv4[7]

Summary: A YOLOv4-based system (with GAN-augmented data, 1,276 images) performs binary classification of engagement and includes a feedback mechanism.

Critical analysis: Its principal advantages are speed and the provision of feedback, but it is constrained by binary labeling, missing multimodal signals, and high computational demands.

Relation to proposed work: YOLO-based action detection is incorporated, but extended to multi-class and multimodal context with scalable and efficient implementation.

2.1.6 Student Engagement Detection in Classrooms with CV[8]

Summary: This empirical system uses YOLOv4 and GAN-based augmentation to monitor engagement in real classrooms in real time.

Critical analysis: Real-world deployment is a core strength. Limitations include binary engagement labeling, small data size, and no multimodal analysis.

Relation to proposed work: Real-time tracking and augmentation concepts are included along with emotion-behavior fusion for nuanced engagement metrics.

2.1.7 DIPSER Dataset: Multimodal Engagement Recognition[6]

Summary: Marquez-Carpintero et al. develop a hardware-integrated, multimodal dataset and recognition system that fuses visual and IMU data (YOLO, MIVIOLC, DeepFace).

Critical analysis: This dataset is a valuable multimodal resource. However, the work does not account for key behavioral features such as posture, gaze, and lacks classroom validation.

Relation to proposed work: We extend multimodal data collection to include posture and gaze, with in-classroom validation.

2.1.8 SLR of Computer Vision in Student Engagement[3]

Summary: Garbal et al. (2025) provide a systematic literature review (meta-analysis of 113 CV-based studies), synthesizing trends, methods, and datasets in the field.

Critical analysis: Comprehensiveness and synthesis are clear strengths; however, the work does not provide original model proposals and reveals serious dataset scarcity.

Relation to proposed work: Dataset expansion and multimodal integration are specifically emphasized to address the cited gaps.

2.1.9 Dynamic Interaction Between Student Behaviour and Environment[4]

Summary: Li and Xue's (2023) meta-analysis (148 effects, 71 studies, 93,189 participants) reveals the relationships between emotional, behavioral, and cognitive factors in student engagement.

Critical analysis: The diverse participant base and multi-factor perspective are strengths, offset by heterogeneity, publication bias, and focus on higher education.

Relation to proposed work: Our project designs multi-factor engagement metrics that explicitly incorporate emotional, behavioral, and cognitive streams.

2.1.10 Detecting Nonverbal Speech and Gaze Behaviours[10]

Summary: This work explores gaze tracking, group interaction detection (rule-based logic), and process mining to study collaboration in classroom video data.

Critical analysis: The study's mapping of group interaction loops is novel. Its limitations are lack of emotion analysis, reliance on rules, and weak classroom generalizability.

Relation to proposed work: The fusion of interaction, gaze, and emotion networks is a key enhancement over this reference.

2.1.11 Vision-Based Gesture Recognition for Engagement Assessment[9]

Summary: Zhang and Wang (2025) propose gesture recognition using hand joint tracking, temporal segmentation, feature fusion, and self-attention—focused on classroom gestures (e.g., hand-raising, pointing).

Critical analysis: Robust gesture recognition with attention-based fusion is a clear strength. Limitations are focus on gestures only (not emotions) and evaluation on gesture-centric datasets exclusively.

Relation to proposed work: We integrate these gesture techniques into our pipeline, fusing them with emotion outputs for richer multimodal engagement scoring.

2.2 Comparative Table of Reviewed Studies

Table 2.1: Comparative analysis of existing studies on AI-based student engagement detection.

Paper	Methodology	Contribution	Limitations
Detecting Student Engagement with CNN and FER [2]	CNN on FER-2013 dataset (48×48 grayscale images with augmentation)	CNN model for engagement detection using static facial expressions	Facial-only analysis, static images, no posture/action data, not real-time
Measuring Student Engagement through Behavioral and Emotional Features [5]	ResNet-50 for features + TCN for temporal analysis on video clips	ResNet + TCN hybrid for engagement detection from classroom videos	Basic emotions only, misses confusion/curiosity, ignores gestures and posture
Improving Student Engagement Detection with ResNet + TCN Hybrid [1]	Transfer learning (ResNet50/VGG16/InceptionV3) with behavioral gating + emotional fusion	Multimodal approach combining behavior and emotions with new engagement metric	Oversimplified cues, ignores fine actions like note-taking or posture changes
Multimodal Engagement Analysis from Facial Videos [9]	ResNet-50 Attention-Net + Affect-Net, multiple classifiers, real classroom data	Facial video-based engagement classification in real classrooms with multi-state labels	Mostly facial expression-based, tested in small controlled lab settings
Student Engagement Detection Using YOLOv4 [7]	YOLOv4 with GAN-augmented dataset (1,276 images), binary classification + feedback	YOLOv4 system for binary engagement with weighted scoring and feedback	Binary classification, lacks multimodal feedback, heavy computational needs
Student Engagement Detection in Classrooms with CV [8]	YOLOv4-based empirical CV system, with GAN data augmentation	Practical classroom demonstration of YOLOv4 for real-time CV monitoring	Binary classification only, limited dataset, no multimodal fusion
DIPSER Dataset: Multimodal Engagement Recognition [6]	Visual + IMU fusion (YOLO, MIVIOLC, DeepFace) on Raspberry Pi clusters	Comprehensive multimodal dataset and system for engagement recognition	Ignores posture, gaze, and head movement, not tested in real classrooms
SLR of Computer Vision in Student Engagement [3]	Meta-analysis of 113 CV-based engagement studies across modalities	Comprehensive synthesis of CV methods, datasets, and engagement trends	No experimental model; highlights lack of public datasets
Dynamic Interaction Between Student Behaviour and Environment [4]	Meta-analysis of 148 effects from 71 studies (93,189 participants)	Identifies emotional, behavioral, and cognitive factors influencing engagement	Heterogeneity, publication bias, limited to higher education
Detecting Nonverbal Speech and Gaze Behaviours [11]	Uses video data for gaze tracking and group interaction detection with rule-based logic; process mining to study collaboration patterns	Introduced seven group interaction types, showed how interaction loops improve group learning, and connected findings with collaborative learning theory	Focuses only on gaze, no emotion detection, no student action analysis, rule-based methods may not generalize, limited classroom applicability
Vision-Based Gesture Recognition for Engagement Assessment [10]	Uses hand joint tracking with deep learning tested on gesture datasets	Provides gesture-based engagement analysis, robust across classroom conditions	Focuses only on gestures, no emotion analysis, not fully multimodal

Chapter 3

Proposed Approach

This chapter details the comprehensive methodology for developing the AI-based student engagement analysis system. Emphasizing multimodal fusion of emotional and behavioral cues, the approach blends computer vision, deep learning, and robust software engineering to deliver both per-student and class-level engagement insights.

3.1 System Overview

The proposed solution adopts a modular, scalable architecture engineered to reliably process classroom video data from ingestion to actionable feedback. The approach is systematically structured around the following core modules:

- **Data Preprocessing:** Handles video file input, frame extraction at regular intervals, and rich data augmentation strategies to boost model robustness.
- **AI Model Pipeline:** Includes two parallel neural networks—one specialized for emotion recognition via facial analysis, and another for action detection (e.g., hand gestures, posture) using object detection and pose estimation.
- **Backend API Services:** Manages all computation beyond model inference, including RESTful API design, secure data storage in both SQL and NoSQL databases, and automated report generation.
- **Frontend Visualization Dashboard:** Provides real-time feedback, video upload, engagement visualization, and detailed reporting to educators through an intuitive web interface.

Figure **??** provides an architectural overview of the entire framework.

3.2 Data Collection and Preprocessing

Custom datasets are constructed to capture the nuanced reality of classroom engagement, including:

- Individual recording of students in diverse classroom-like scenarios to ensure high annotation quality.
- Annotation of primary emotions (happy ,confused/bored, neutral) and key behavioral actions (hand raise, writing notes, mobile usage, looking away, sleeping , yawning).
- Automated preprocessing pipeline for video frame extraction and class-balanced augmentation.

This structured approach ensures the resulting dataset supports robust model training and fair, repeatable evaluation.

3.3 Model Architecture and Fusion

3.3.1 Parallel Model Design

Two specialized neural network pipelines operate in parallel:

- **Emotion Recognition Network:** Convolutional Neural Networks (CNNs) fine-tuned for facial expression analysis.
- **Action Detection Network:** Object detection and pose estimation models for identifying key student behaviors.

3.3.2 Multimodal Fusion and Engagement Scoring

Outputs from the emotion and action branches are fused through a multi-level integration mechanism:

- Weighted engagement scoring combines emotion probabilities with detected behavioral cues.
- Temporal smoothing accounts for engagement variability across video sequences, yielding both per-frame and session-level scores.

3.4 System Implementation

- **Backend:** Developed in Python using FastAPI/Flask, leveraging both relational (MySQL) and non-relational (MongoDB) databases.
- **Frontend:** Built with React.js for responsiveness and ease-of-use, incorporating dynamic visualizations for educators.
- **Deployment:** Docker-based containerization ensures consistent and scalable rollout across cloud or local environments.

3.5 Evaluation Methodology

Performance is assessed through a multi-phase evaluation:

- **Technical Validation:** Assessed via standard metrics (accuracy, precision, recall, F1-score), real-time performance, and computational resource analysis.
- **User Validation:** Involves classroom educators testing the system, providing feedback on practicality, interpretability, and the utility of engagement reports.

3.6 Parameters

These are the parameteres or classes in the dataset :

1. Hand Raise
2. Looking Away
3. Mobile Using
4. Neutral
5. Sleeping
6. Writing Notes
7. Yawning
8. Happy
9. Confusion/Bored

These discrete states reflect diverse forms of student engagement and serve as primary labels for multimodal analysis.

3.7 Graphical Models and System Design

The following figures represent the complete set of UML and system diagrams developed for the proposed AI-Based Student Engagement Analysis system.

All these diagrams collectively demonstrate the design perspective of the system — from user interactions (use case) to structural (class), behavioral (activity and sequence), logical (DFD), and physical (component) viewpoints — ensuring a well-rounded representation of the proposed approach.

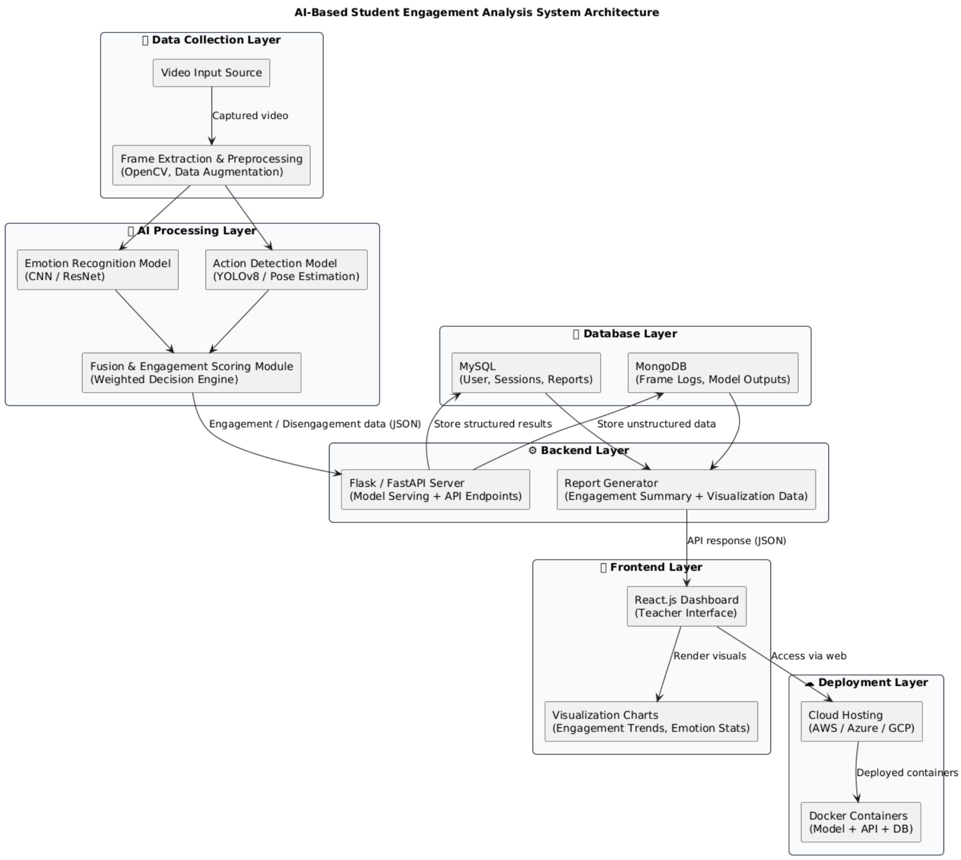

Figure 3.1: System Architecture Diagram illustrating the end-to-end workflow of the proposed multimodal engagement detection framework.

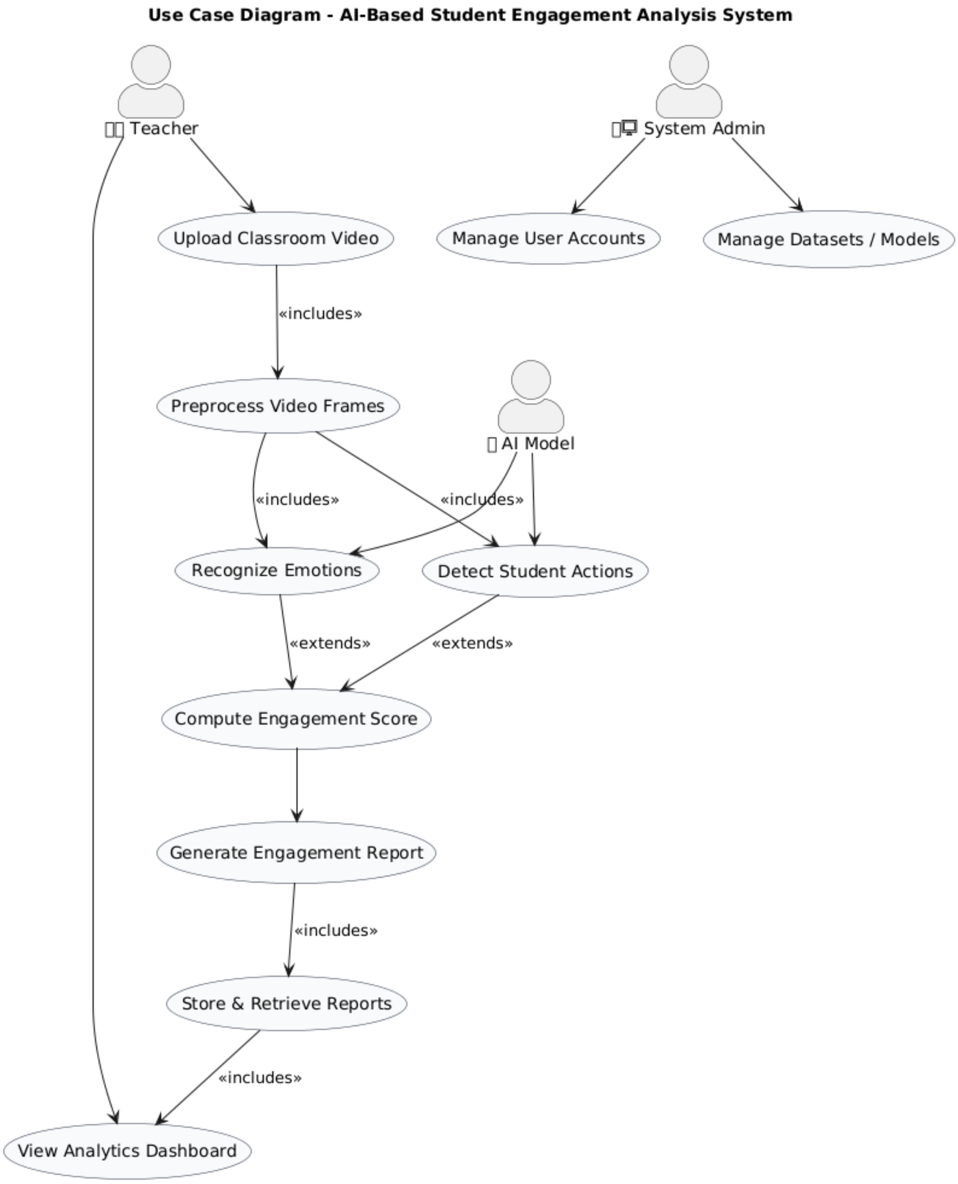

Figure 3.2: Use Case Diagram illustrating the main actors and interactions within the system.

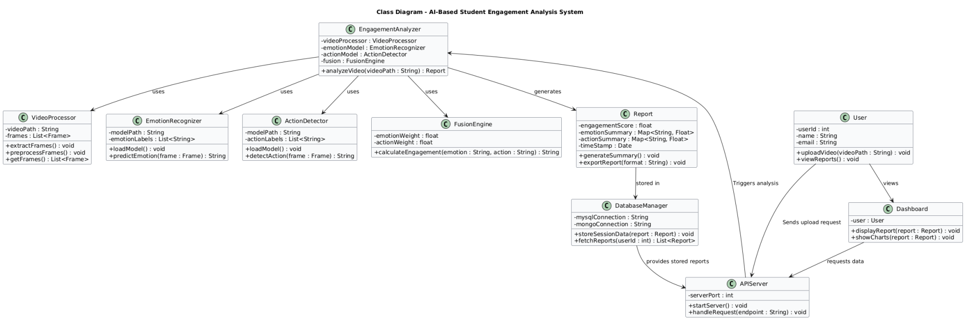

Figure 3.3: Class Diagram depicting the structural relationships among the core classes of the system.

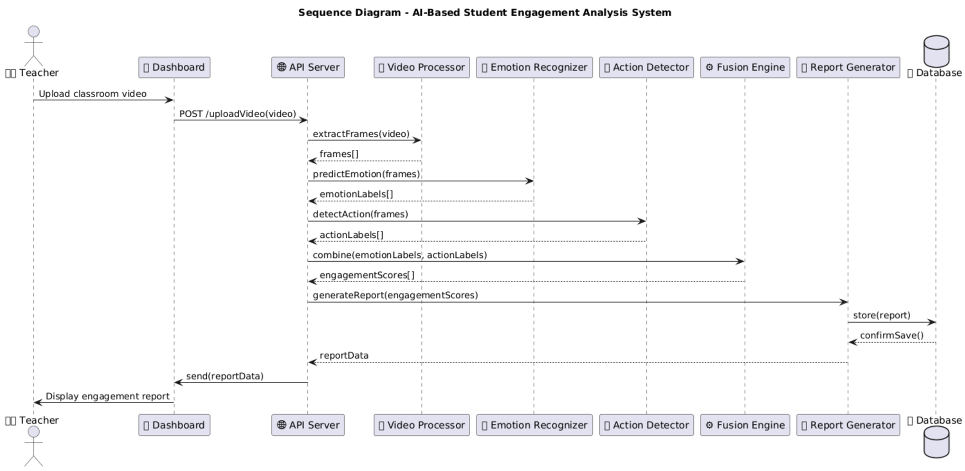

Figure 3.4: Sequence Diagram showing the interaction flow between system components during engagement analysis.

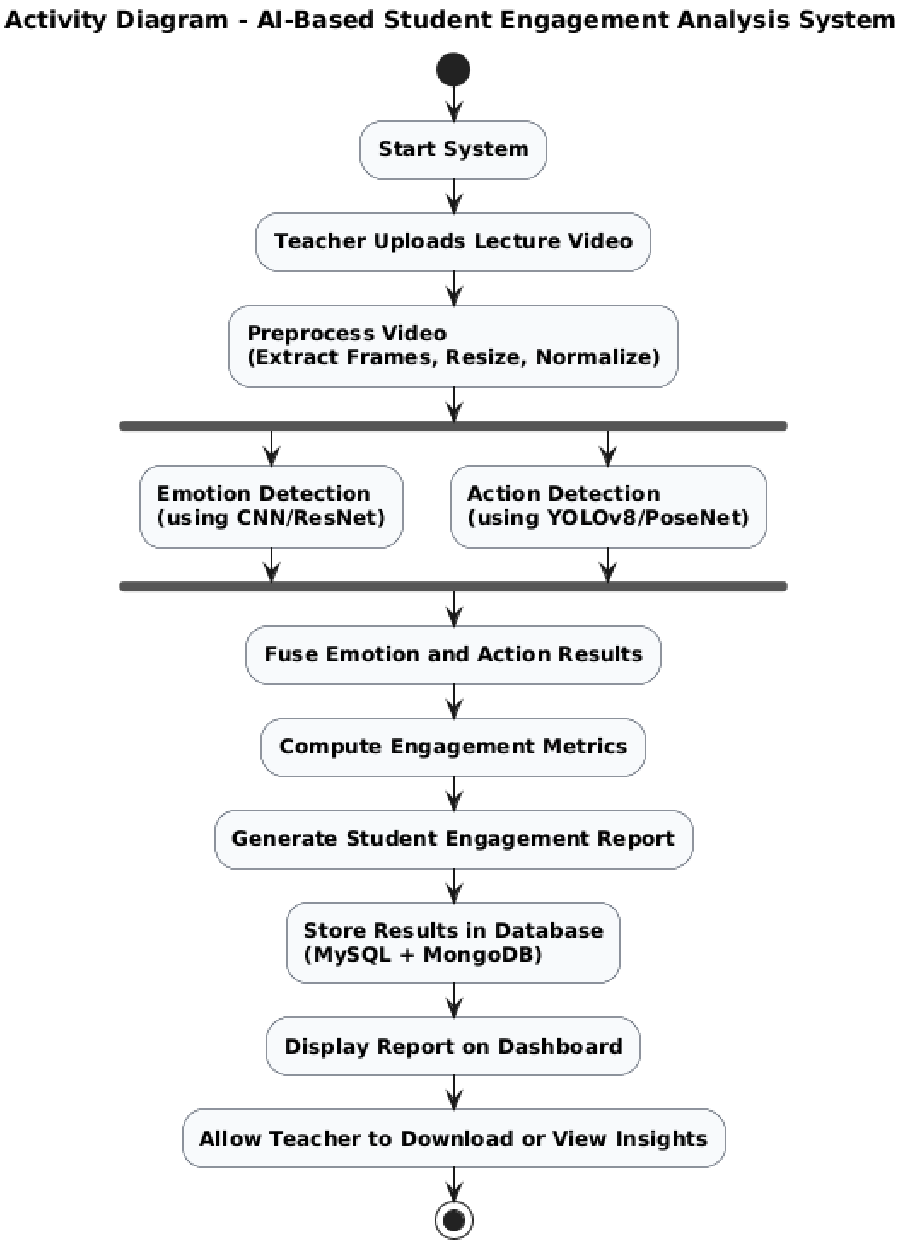

Figure 3.5: Activity Diagram outlining the process flow from video input to engagement report generation.

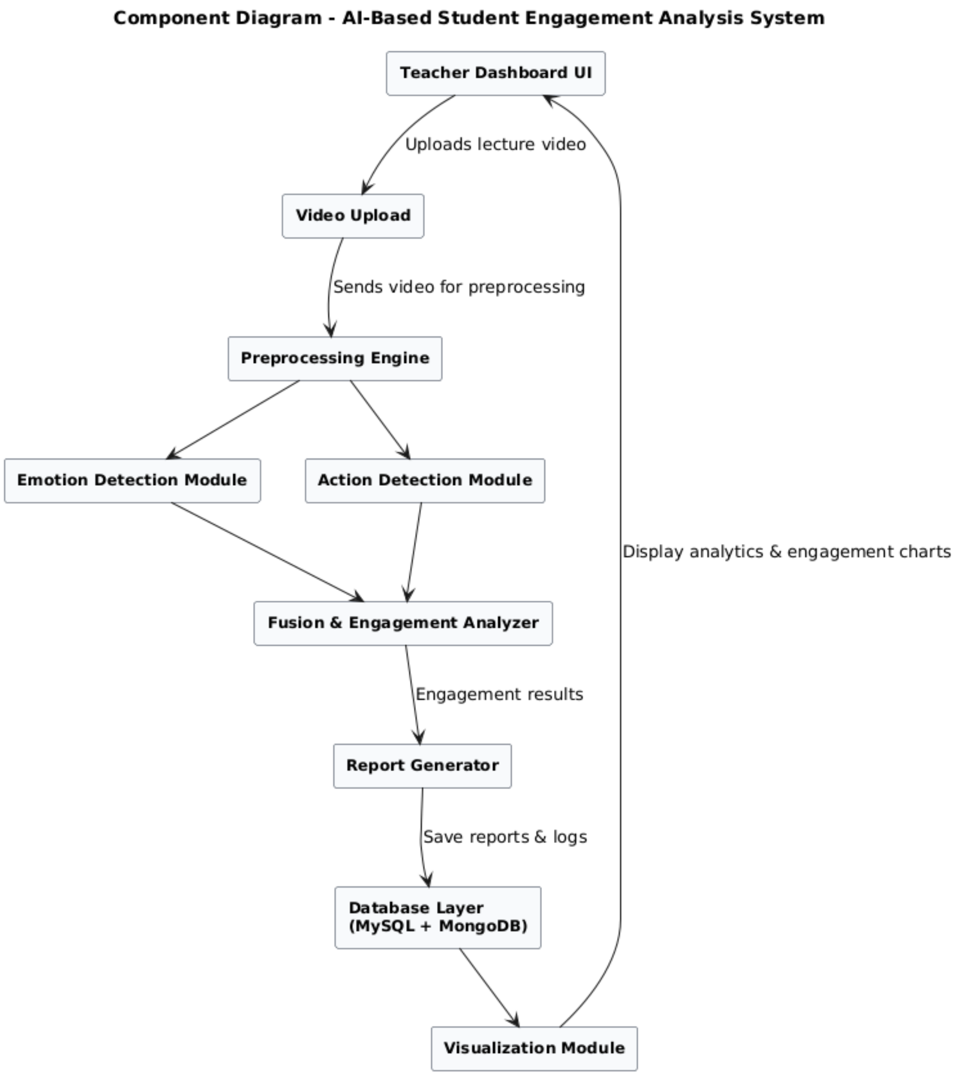

Figure 3.6: Component Diagram illustrating the high-level modular decomposition of the system.

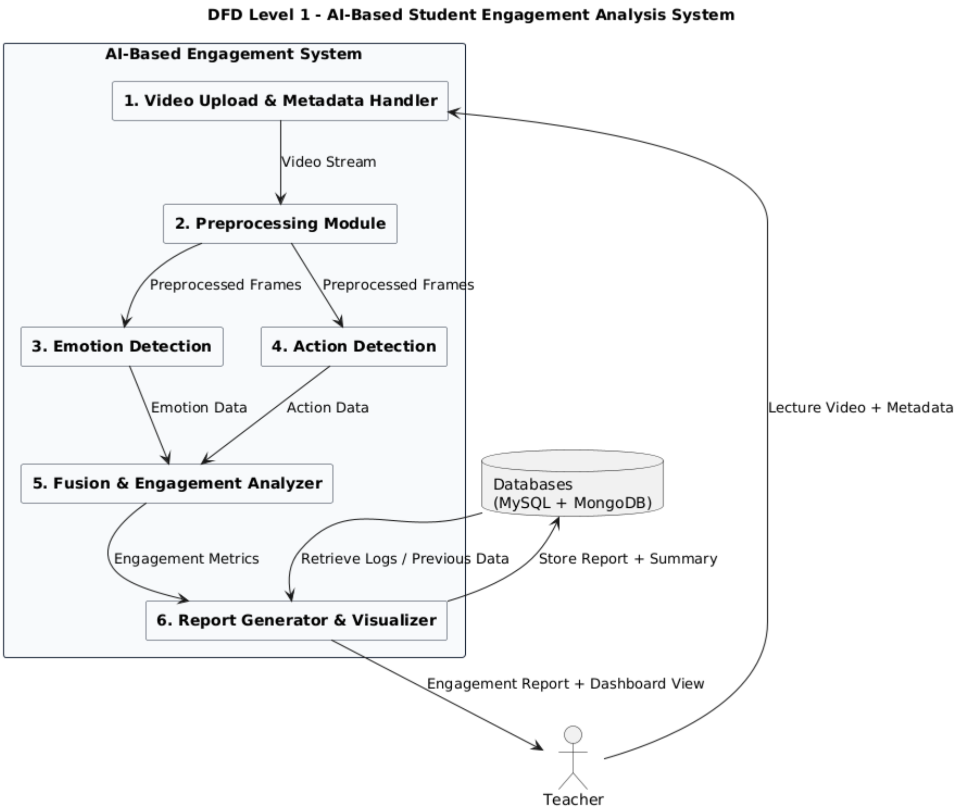

Figure 3.7: Data Flow Diagram (Level 1) representing the logical flow of data among system processes and data stores.

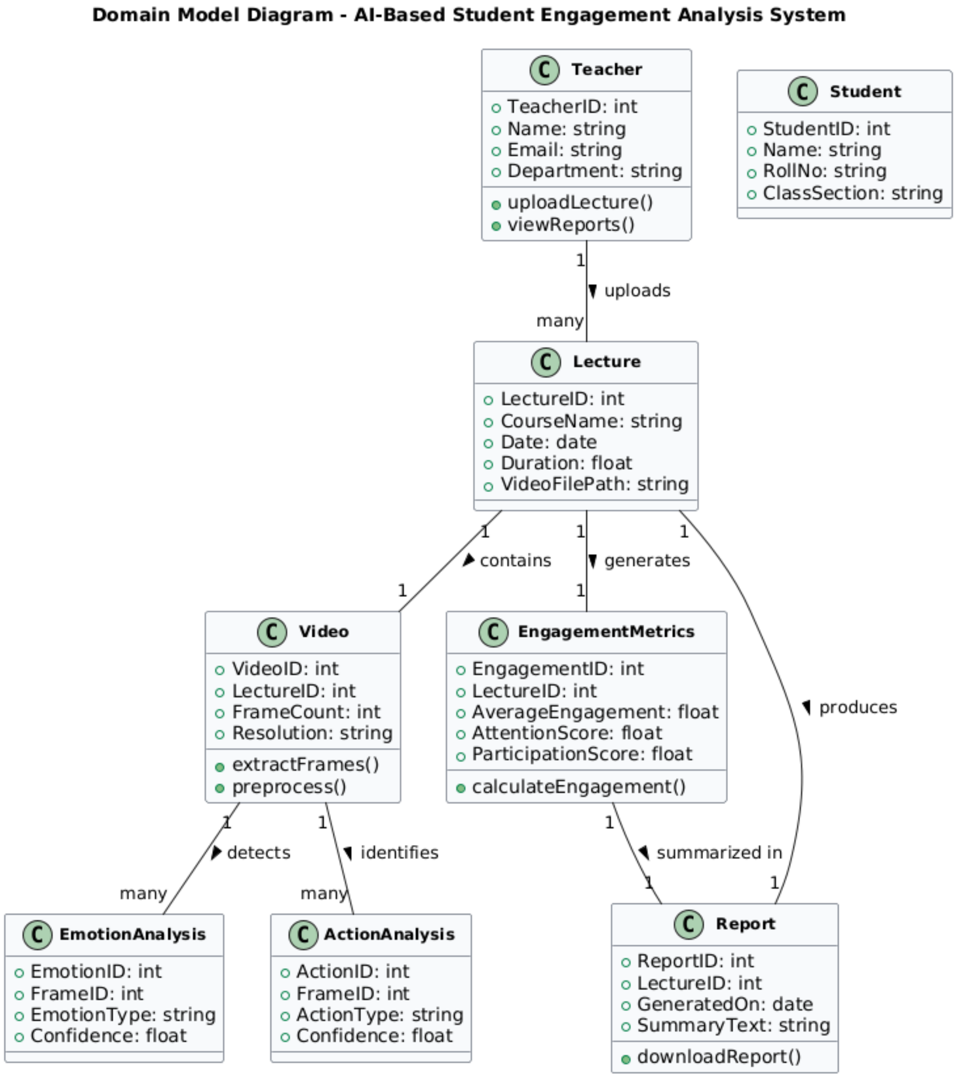

Figure 3.8: Domain Model Diagram representing the key classes and their relationships for engagement detection.

Bibliography

[1] Ali Abedi and Shehroz S. Khan. Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network. *2021 18th Conference on Robots and Vision (CRV)*, pages 151–157, 2021.

[2] Nuha Alruwais and Mohammed Zakariah. Detecting student engagement with convolutional neural networks and facial expression recognition. *Traitement du Signal*, 42(2):229–240, 2025.

[3] Mohamed Garbal, Siham El Janati, and Lahcen El Ghadraoui. Student's engagement detection based on computer vision: Systematic literature review. *Education and Information Technologies*, 30(3):2845–2891, 2025.

[4] Xiaojing Li and En Xue. Dynamic interaction between student learning behaviour and environment: Meta-analysis of student engagement in higher education. *Behaviour Information Technology*, 42(4):471–497, 2023.

[5] Nasir Mahmood, Sohail Masood Bhatti, Hussain Dawood, Manas Ranjan Pradhan, and Haseeb Ahmad. Measuring student engagement through behavioral and emotional features using deep-learning models. *Algorithms*, 17(10):458, 2024.

[6] L. Marquez-Carpintero, S. Suescun-Ferrandiz, C. Lorenzo Álvarez, J. Fernandez Herrero, D. Viejo, R. Roig-Vila, and M. Cazorla. Dipser: A dataset for in-person student engagement recognition in the wild. *arXiv preprint arXiv:2502.20209*, 2025.

[7] A. S. Pillai. Student engagement detection in classrooms through computer vision and deep learning: A novel approach using yolov4. *Sage Science Review of Educational Technology*, 5(1):87–97, 2023.

[8] Aishwarya S. Pillai and N. Radhika. Student engagement detection in classrooms through computer vision and deep learning. In *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, pages 1–6. IEEE, 2022.

[9] Xin Zhang and Li Wang. A vision-based gesture recognition and student engagement assessment model for interactive educational environments. *IEEE Transactions on Learning Technologies*, 2025.

[10] Qi Zhou, Wannapon Suraworachet, and Mutlu Cukurova. Detecting non-verbal speech and gaze behaviours with multimodal data and computer vision to interpret effective collaborative learning interactions. *Computers Education*, 197:104712, 2023.

[11] Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*, 14(2):1012–1027, 2021.