DEEP LEARNING-BASED STUDENT MONITORING IN COMPUTER LABS USING LSTM, HOI, AND FACIAL FEATURE EXTRACTION WITH NYN TRACK DASHBOARD

Ms. N. Pushpa M.E., (Ph.D),

Department of Artificial Intelligence and Data Science

R P Sarathy Institute of Technology, India

pushpa.n@rpsit.ac.in

Mr. A. Ajay

Department of Artificial Intelligence and Data Science

R P Sarathy Institute of Technology, India

ajay09official@gmail.com

Mr. R. R. Kavin

Department of Artificial Intelligence and Data Science

R P Sarathy Institute of Technology, India

kaivnaids@gmail.com

Mr. N. Gunaseelan

Department of Artificial Intelligence and Data Science

R P Sarathy Institute of Technology, India

guna25750@gmail.com

ABSTRACT

Computer laboratory sessions are essential for practical learning among technology students. However, due to manual monitoring practices in laboratories, attendance records often fail to accurately reflect students' actual work and the effective time spent during lab sessions. To validate staff who regularly use and supervise computer laboratories. The survey results revealed that a significant percentage of students reported experiencing distraction during lab periods, while a considerable proportion of staff indicated difficulties in continuously monitoring all students. These challenges negatively impact student productivity and learning habits, leading to abnormal activities such as mobile phone usage, talking, idle sitting, and leaving seats without supervision. To address these issues, this work proposed an automated laboratory monitoring solution that integrates overhead CCTV cameras and front-facing webcams installed in the laboratory environment. Overhead cameras are used to track student actions and seating behavior, while front-facing webcams are utilized to extract facial features and analyze emotional states. The proposed system records the duration of each detected activity using computer vision techniques. The processed data is then analyzed and presented as detailed reports through a web-based ERP platform accessible to students, staff, and departmental authorities.

Keywords: Student Behavior Analysis, CCTV Surveillance, Educational Surveillance, Automated Lab Monitoring

1. INTRODUCTION

1.1 Background

In contemporary society, artificial intelligence has made a remarkable impact on the education system by improving both the teaching and learning processes through various tools and technologies [1]. AI-based solutions help students with personalized learning and also assist teachers in presenting information more effectively and correctly. With the growing use of digital technology in educational institutions, various tools have been developed to enhance student engagement and ensure the quality of education provided to students [2]. Among these, classroom monitoring systems have also gained popularity as they assist institutions in maintaining discipline, enhancing teaching efficiency, and ensuring that learning outcomes are met.

To further improve teaching efficiency and student learning, surveillance systems have been implemented in classrooms to assist teacher monitoring and analyze student activity in real time [3]. These systems help in better understanding of classroom dynamics and also

help in improving the overall learning environment.

In addition, efficient management of institutional resources and academic data regarding students and teachers is also necessary for improving educational outcomes. Appropriate management of student and teacher data helps institutions in monitoring performance, managing activities, and assisting in decision-making processes [4].

1.2 Motivation

Supervision of computer labs in technical education has continuously posed a difficult task. The supervisor or instructor is responsible for managing attendance, monitoring the use of the systems, and ensuring the students are engaged, all in one session. As the number of students in classrooms continues to grow, it will become increasingly inefficient to provide supervision through manual means, and this may lead to less than-ideal learning conditions and inability for instructors to know whether students are engaged in the learning activities or not.

The administration of instructor overhead has been addressed with the introduction of automated systems based on computer vision technology to take the place of manual attendance registers [5]. While these automated systems have helped to improve attendance accuracy, attendance alone is not an adequate indication of whether a student is engaged in the learning process, particularly in computer labs, where students may be physically present but actively disengaged from their learning activities. As the use of technology-assisted learning continues to grow, understanding engagement among students has become critical to increasing students' academic achievement.

Recent studies have shown that using multi-modal machine learning methodologies for the purpose of evaluating student engagement via a mixture of behaviourally and interactionally based data can provide real-time evaluations of student engagement. [6] While these types of systems can demonstrate the importance of engagement-aware monitoring, they are generally studied in isolation from the context of computer lab operations. Furthermore, intelligent educational systems emphasize the need for integrated frameworks that combine monitoring, engagement analysis, and academic management in order to improve the quality of learning at institutions of higher learning while increasing the operational efficiency of the institution [7].

These limitations suggest a need for the development of an AI-based computer lab supervision framework that will integrate automated attendance monitoring with real-time engagement-aware supervision.

1.3 Objectives of proposed solution

The main aim of the proposed research work is to design and develop an AI-powered computer lab supervision system that facilitates continuous monitoring and academic management in technical education settings. The objectives of the proposed research work are as follows:

- To design and develop an automated attendance monitoring system for computer labs using CCTV-based visual data, thereby reducing manual intervention and administrative burden.
- To facilitate real-time monitoring of student engagement during lab sessions through continuous visual analysis.
- To facilitate continuous supervision of computer lab activities, thereby assisting teachers in handling large numbers of students effectively.
- To design a web-based ERP system for storing, managing, and visualising attendance, engagement, and academic data.
- To combine student behavioural knowledge with academic data to facilitate systematic analysis and informed institutional-level decision-making.
- To develop an intelligent alert system that automatically notifies teachers or administrators in real-time when predefined conditions for attendance or engagement are met.
- To ensure real-time processing, scalability, and non-intrusive implementation without affecting the normal functioning of computer labs.

2. LITERATURE SURVEY

2.1 Face Recognition for attendance system

Face recognition technology is widely used for automated attendance systems because it can identify students quickly and accurately. Traditional attendance methods such as manual roll calls and ID cards are time-consuming and may lead to errors. Therefore, many researchers have developed automated attendance systems using computer vision and deep learning techniques.

Dang proposed a smart attendance system based on an improved facial recognition model that combines FaceNet, MobileNetV2, and Single Shot Detection (SSD) for face detection and recognition. The system extracts facial features and compares them with stored facial data to identify students and record attendance automatically. The system achieved an accuracy of approximately 97–99% and supports real-time attendance monitoring [8].

Alruwais and Zakariah proposed a deep learning based framework for student recognition and activity monitoring in online classrooms. The system uses a Convolutional Neural Network (CNN) model to recognize students and analyse their engagement using facial expressions and head pose information. Experimental results showed that the system achieved around 99% accuracy in student recognition [9].

2.2 Emotion Recognition and Engagement score

Emotion recognition plays an important role in understanding human behaviour and improving human–computer interaction systems. Facial expressions provide important information about human emotions, and computer vision techniques are widely used to analyze these expressions automatically. Jonathan et al. studied facial emotion recognition using computer vision techniques and reviewed several algorithms used to detect facial expressions and classify emotions from facial images and videos. The study analyzed methods such as Support Vector Machine (SVM), Adaboost, neural networks, and SURF feature extraction to improve emotion recognition performance in computer vision systems. Similarly, Pandey et al. proposed a facial emotion detection and recognition system using deep learning techniques, particularly Convolutional Neural Networks (CNN). The proposed system performs emotion recognition through three main stages: face detection, feature extraction, and emotion classification. The study also utilized datasets such as the Karolinska Directed Emotional Faces (KDEF) dataset and the Japanese Female Facial Expression (JAFFE) dataset for training emotion recognition models. These studies demonstrate that machine learning and deep learning techniques can effectively recognize emotions from facial expressions and can be applied in applications such as human–computer interaction, behavioral analysis, and intelligent monitoring systems [10], [11]

2.3 Human Object Interaction

Human–object interaction (HOI) recognition is an important task in computer vision that enables machines to understand how humans interact with objects in real-world environments. It has applications in intelligent surveillance systems, smart environments, robotics, and activity monitoring. Ozaki et al. proposed a human–object interaction recognition method designed for intelligent spaces and edge devices. The system combines human pose estimation and object detection techniques to recognize interactions between humans and objects. The method uses YOLOv5 for object detection and MediaPipe for extracting human pose landmarks, and then applies machine learning techniques to classify interaction classes based on features such as human posture, object shape, and the distance between human landmarks and objects. The results show that the system can achieve high recognition accuracy and robustness across different environments and camera conditions. Similarly, Shehata and Abdolrahmani proposed a human–object interaction recognition approach that integrates scene information and multi-task learning to improve action recognition performance. The proposed framework uses skeleton-based action recognition with Graph Convolutional Networks (GCNs) and incorporates scene interaction nodes representing objects in the environment. The system combines GCN and GRU architectures to capture spatial and temporal relationships between human poses and objects. Experimental results demonstrated that integrating scene information significantly improves recognition performance, achieving an accuracy of 99.25% for human–object interaction recognition tasks [12], [13].

2.4 Action Recognition using Skeleton joins

Human Action Recognition (HAR) is an important research area in computer vision that focuses on identifying and classifying human activities from images or video sequences. It has applications in surveillance systems, healthcare monitoring, robotics, and human–computer interaction. Khan and Jung proposed a deep learning–based human activity recognition model that combines Dilated Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks to capture spatial and temporal features from video sequences. The model expands the receptive field using dilated convolution and uses LSTM to learn temporal dependencies between frames, achieving an accuracy of 94.9% on the UCF50 dataset. Similarly, Gao et al. introduced a skeleton-based human action recognition method using OpenPose to extract joint coordinates and an improved ExGist feature descriptor for representing skeleton features. The extracted features are then classified using machine learning classifiers such as SVM, achieving an accuracy of 89.2%. Furthermore, Zhang and Wang proposed a human action detection and recognition method based on interpretable artificial intelligence using YOLOv5 for human detection, AlphaPose for skeleton estimation, and Spatio-Temporal Graph Convolutional Networks (ST-GCN) for extracting spatio-temporal relationships between skeleton joints. The proposed method achieved a recognition accuracy of 92.04% for various human actions such as running, kicking, squatting, and falling. These studies demonstrate that combining deep learning models, skeleton-based methods, and object detection techniques can effectively improve the performance of human activity recognition systems [14]-[16].

3.5 NYN Track Application

Learning analytics systems are increasingly used in educational environments to monitor students' engagement, concentration, and emotional states during online learning. Hasnine et al. proposed a real-time learning analytics dashboard called MOEMO for automatically detecting students' affective states during online classes. The system analyzes lecture videos captured from students' webcams and extracts emotional information using computer vision and deep learning techniques. The framework detects facial expressions, eye gaze direction, and emotional states to determine students' engagement and concentration levels during the lecture. Several algorithms are used in the system, including MTCNN for face detection, Dlib for face recognition, Mini-Xception for emotion detection, and the PnP algorithm for eye gaze estimation. The platform processes the extracted data using Python-based analytics tools such as Pandas, Matplotlib, and Plotly to generate real-time visualizations. The results are displayed on a dashboard that provides instructors with insights such as engagement levels, concentration levels, emotion distribution, and clusters of engaged and disengaged students. The system also generates an after-class report that helps instructors analyze students' learning behavior and intervene when students show low engagement or concentration during online learning sessions [17].

3.6 Research gap

Although significant progress has been made in the areas of face recognition, emotion recognition, human–object interaction recognition, human action recognition, and learning analytics systems, most existing studies address these components independently rather than integrating them into a unified intelligent monitoring framework. Current face recognition-based attendance systems primarily focus on identity verification and automated attendance marking but do not analyse students' behavioural patterns or engagement during learning sessions. Similarly, emotion recognition systems mainly detect facial expressions to classify emotions, but they often lack contextual understanding of students' physical actions and interactions within the environment.

Human–object interaction recognition and skeleton-based action recognition methods have demonstrated high accuracy in recognizing complex activities, yet these approaches are generally applied in surveillance or activity monitoring scenarios and are rarely integrated with educational monitoring systems.

Furthermore, existing learning analytics dashboards mainly rely on emotional or gaze-based indicators to estimate student engagement, which may not fully capture the complexity of student behaviour during learning activities.

Therefore, there is a clear need for an integrated framework that combines multiple computer vision techniques to provide a comprehensive understanding of student activities and engagement.

To address these limitations, this research proposes a unified intelligent system that integrates face recognition for automated attendance tracking, emotion recognition for engagement analysis, human–object interaction recognition, and skeleton-based action recognition techniques within a learning analytics platform.

The proposed system aims to enhance the accuracy and effectiveness of monitoring student participation, behavior, and engagement in educational environments by combining multiple behavioral and visual analysis methods into a single framework.

Author(s)	Method / Model Used	Key Contribution	Dataset / Tools	Accuracy / Outcome	Limitations
[8] Dang	FaceNet + Mobile Net V2 + SSD	Smart attendance system using deep face recognition	Real time face dataset	97–99%	Only attendance, no behavior analysis
[9] Alruwais & Zakariah	CNN	Student recognition + engagement monitoring	Facial expression & head pose	~99%	Limited to online
[10] Jonathan et al.	SVM, Adaboost, NN, SURF	Emotion recognition techniques comparison	Facial datasets	Improved performance	Not unified system
[11] Pandey et al.	CNN	Emotion detection (3- stage)	KDEF, JAFFE	High accuracy	No real time integration
[12] Ozaki et al.	YOLOv5 + MediaPipe	HOI detection	Real world data	High accuracy	Limited scope
[13] Shehata & Abdolrahman i	GCN + GRU	HOI with spatio temporal learning	Skeleton + scene data	99.25 %	High computation
[14] Khan & Jung	Dilated CNN + LSTM	Action recognition	UCF50	94.9%	Dataset specific
[15] Gao et al.	OpenPose + SVM	Skeleton based HAR	Joint dataset	89.2%	Lower accuracy
[16] Zhang & Wang	YOLOv5 + AlphaPose + ST-GCN	Interpretable HAR	Action datasets	92.04 %	Complex pipeline
[17] Hasnine et al.	MTCNN + Dlib + Mini Xception + PnP	Learning analytics dashboard	Webcam + analytics tools	Real time insights	Limited behavioral depth

Table 1: Comparative Analysis of Existing Methods in Student Monitoring Systems

3. PROPOSED SOLUTION

3.1 System Overview

To address the limitations identified in previous studies, this research proposes an integrated intelligent monitoring framework that combines face recognition, emotion recognition, human–object interaction analysis, skeleton-based action recognition, and an ERP-based analytics dashboard. The proposed system is designed to automatically monitor student attendance, engagement levels, and behavioral interactions during learning sessions using computer vision and deep learning techniques.

The system processes video data captured through cameras in the learning environment and performs multiple stages of analysis. First, the system detects and recognizes student faces to automatically record attendance. Next, facial expression analysis is performed to estimate students' emotional states and engagement levels. In addition, human–object interaction and skeleton-based action recognition techniques are used to analyze students' physical behaviors and interactions with surrounding objects. Finally, the extracted information is integrated into an ERP-based learning analytics dashboard that provides instructors with real-time insights into students' participation, engagement, and behavioral patterns.

The integration of multiple computer vision modules enables the system to provide a more comprehensive understanding of student behavior compared to existing single-module approaches. The proposed framework improves monitoring accuracy and supports instructors in making better decisions to enhance the learning experience.

3. 2 System Architecture

Figure 1: Full System Architecture

3.3 Face Recognition & Attendance Module

In the proposed system, the face recognition and attendance module operates on video input captured from the monitoring cameras. Each frame from the video stream is first preprocessed by resizing it to a fixed resolution of **224 × 224 pixels**, ensuring uniform input for subsequent processing.

The preprocessed frames are then passed to the **Buffalo_L**, which performs face detection for each frame and extracts discriminative facial embeddings. These embeddings represent the unique identity of each student in a high dimensional feature space.

The extracted embeddings are compared with the pre-stored embeddings of students available in the class database using similarity measures. When a match is found, the corresponding **Student_ID** is retrieved, and attendance is marked for that student.

The system incorporates a time-based attendance mechanism to classify attendance status. If a student is recognized before a predefined threshold time, the attendance is marked as *Present (On Time)*. If the recognition occurs after the threshold, the status is updated as *Present (Late)*.

To ensure accurate attendance tracking, the system maintains a **present list**, where each recognized Student_ID is stored only once, avoiding duplicate entries. The recognized students are continuously tracked throughout the session, enabling consistent identification for subsequent modules such as action recognition and engagement analysis.

At the end of the lab session, the system performs a validation step by comparing the present list with the predefined class student list. Students whose Student_IDs are not found in the present list are automatically marked as *Absent*. This session based mechanism ensures reliable and automated attendance management without manual intervention.

The final output of this module consists of structured attendance records containing the Student_ID, attendance status (on time, late, or absent), and timestamp. These records are further utilized by other modules and stored in the database for monitoring and analysis.

3.4 Emotion and Engagement Analysis:

The emotion and engagement analysis module processes video input from a webcam to evaluate student attentiveness in real time. Each frame is preprocessed and analyzed using a dual-model approach.

Facial landmarks are extracted using **MediaPipe**, which provides 468 keypoints. These are used to compute the Eye Aspect Ratio (EAR) and Mouth Aspect Ratio (MAR) to determine eye and mouth states. The system uses threshold values of **EAR < 0.15** for eye closure and **MAR > 0.35** for mouth opening.

In parallel, an ensemble-based emotion recognition model combining **EfficientNet** and **ResNet** is used to classify facial emotions.

The extracted features are fused to determine engagement levels. If the eyes remain closed for more than **3 seconds**, the student is classified as *Not Engaged* and a drowsiness alert is generated. Similarly, repeated mouth opening for more than **2 seconds** within a defined time window triggers a talking or yawning alert.

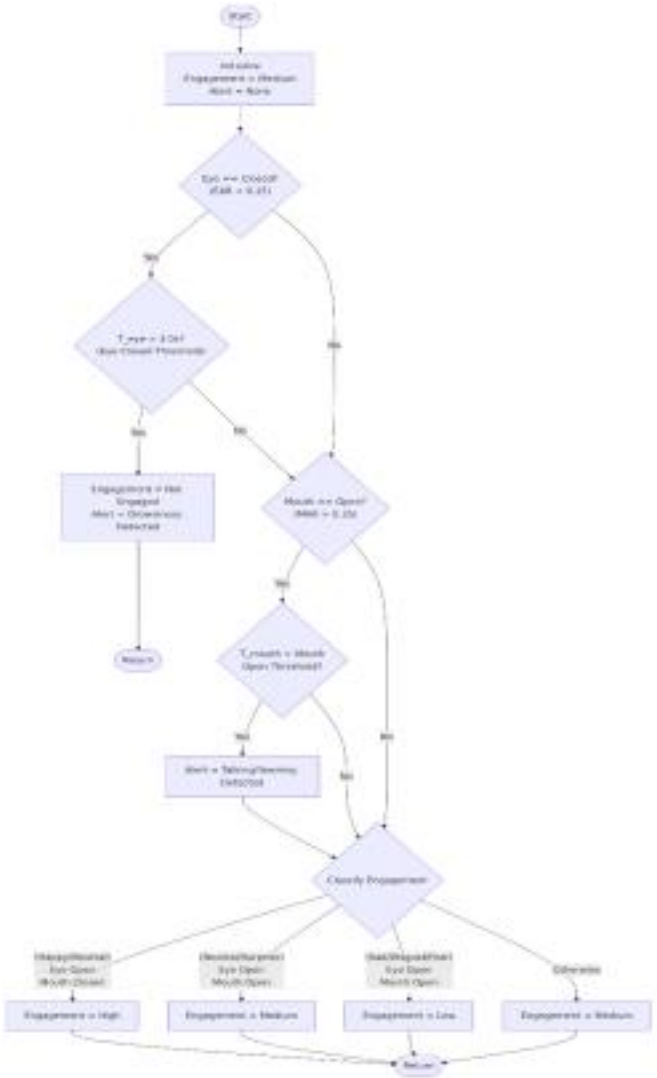

Figure 2: Architecture of the NYN TRACK Dashboard

Based on emotion, eye state, and mouth state, engagement is categorized into *High*, *Medium*, or *Low*. The final output includes Student_ID, emotion, engagement level, and alert status, which are stored for monitoring and analysis.

3.5 Human Object Interaction

The Human Object Interaction (HOI) module detects relevant objects and infers interactions in real time from webcam input. Each frame is resized to **640 × 480 pixels** and processed using a hybrid detection approach.

A **YOLOv8** is used to detect objects such as notebooks, mobile phones, laptops, and keyboards with a confidence threshold of **0.35** and IoU threshold of **0.5**. Additionally, a **Mask R-CNN** model (Detectron2) is employed to detect pens, using a score threshold of **0.8** and a minimum mask area of **800 pixels** to ensure accurate segmentation of small objects.

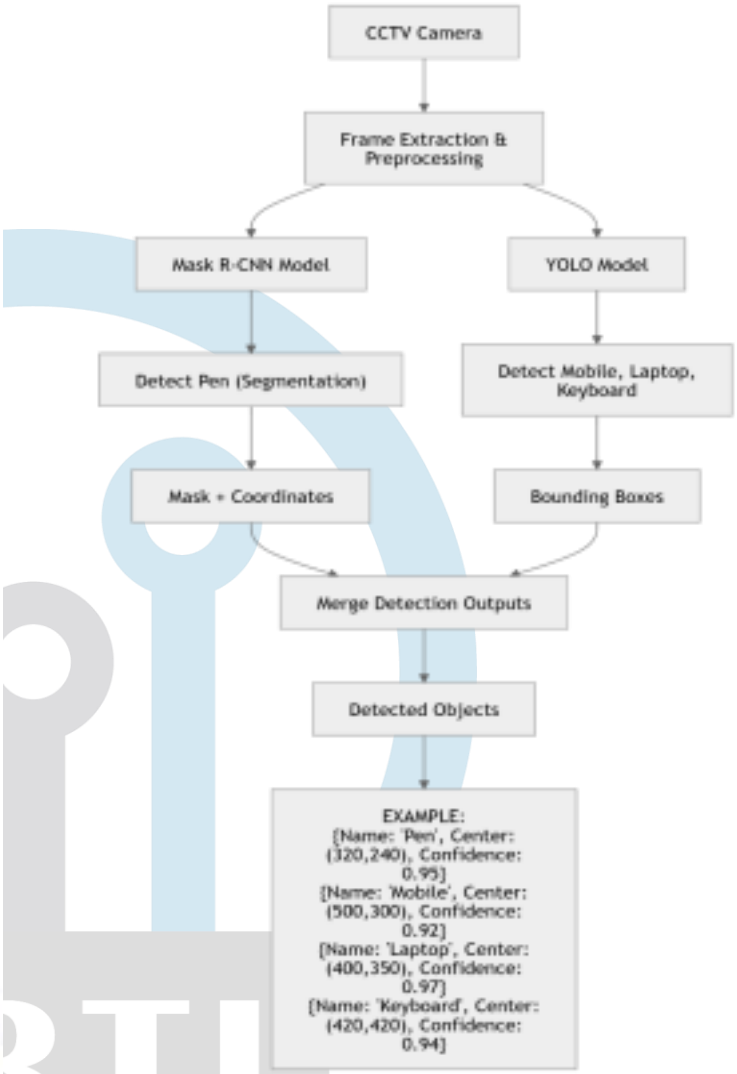

Figure 3: Architecture of HOI

The system filters and retains only relevant objects, namely *pen*, *notebook*, *keyboard*, *and mobile*. For each detected object, the center coordinates (□□□ □□□) are computed from the bounding box, along with the associated confidence score.

The output is represented as a structured list of objects, where each entry contains the object name, center coordinates, and confidence value. This representation simplifies spatial reasoning and facilitates integration with higher-level modules.

Interaction logic is applied using spatial relationships between objects. For instance, if the bounding boxes of a pen and a notebook overlap, the system identifies the activity as *writing*.

3.6 Action Recognition

The Action Recognition module is designed to identify and classify human activities within a classroom environment. Unlike traditional approaches that rely solely on pose estimation or visual features, the proposed system integrates

multiple cues, including skeletal joint information, human-object interactions (HOI), and facial engagement features.

This multi-modal approach improves the robustness and accuracy of activity recognition.The system focuses on recognizing key classroom activities such as working, sitting idle, using a mobile device, talking, sleeping, standing, walking, and writing.

For human pose estimation and tracking, the YOLOv8 large model is utilized. This model detects individuals in each frame and extracts 17 key skeletal joints representing important human body landmarks.

To ensure continuity across frames, each detected individual is assigned a unique **Student_ID**, enabling consistent tracking throughout the video sequence.

This identity preservation is essential for capturing temporal motion patterns and avoiding misclassification caused by identity switching.

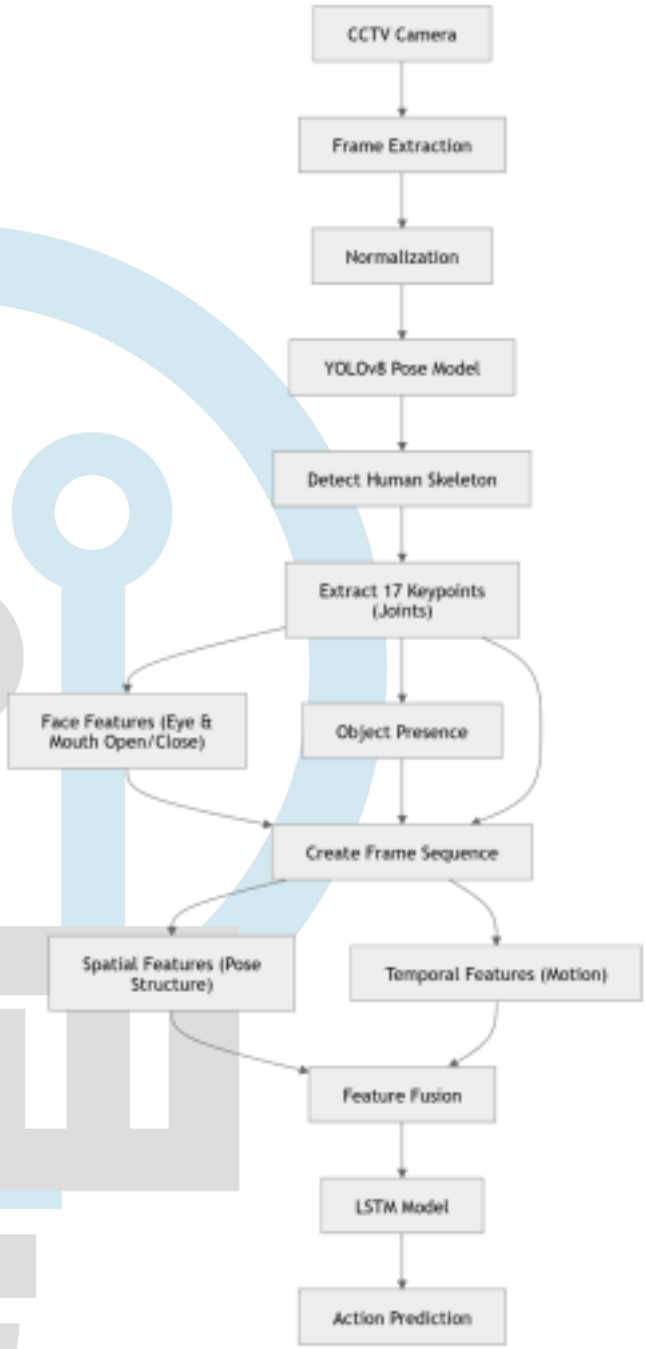

Figure 4: Architecture of Action Recognition model

3.7 NYN TRACK Dashboard

The **NYN TRACK Dashboard** serves as the central interface for managing users and visualizing system outputs through an ERP-based structure. It supports role-based access for *Student*, *Staff*, *Department Admin*, *Lab Admin*, and *Super Admin*.

Users log in using a unique user ID, password, and role. The **Super Admin** manages the system by creating and controlling Department Admin and Lab Admin accounts. Department Admins manage student, staff, and timetable data, while Lab Admins handle lab configuration, including timetables, seating layout, and CCTV integration. Staff members monitor student reports and

schedules, whereas students can view their reports, timetables, and instructions.

All module outputs are integrated using **Student_ID**, enabling unified tracking of attendance, engagement, actions, and object interactions. The dashboard displays real-time data and alerts, providing an efficient and scalable solution for intelligent lab monitoring.

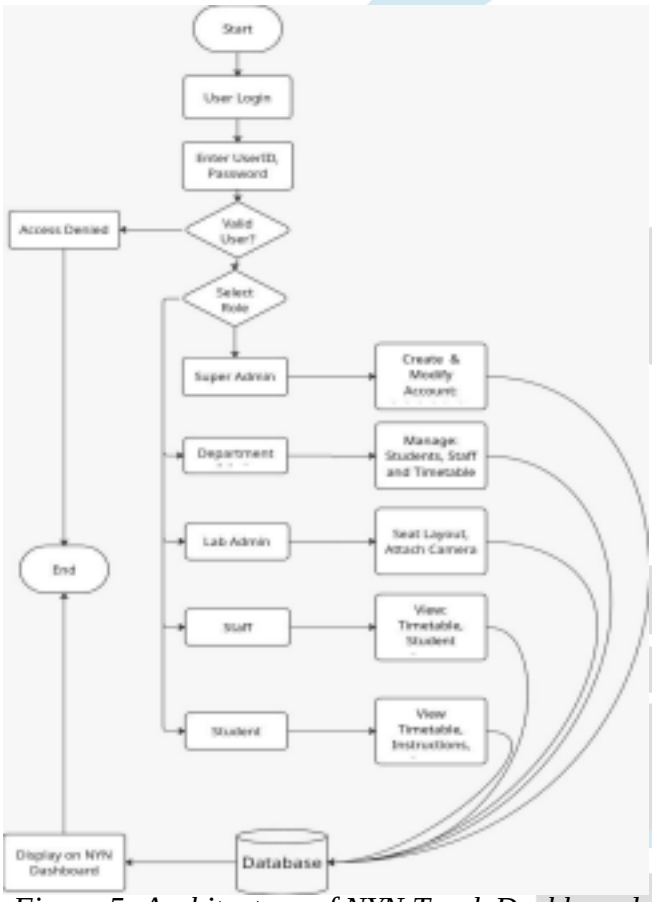

Figure 5: Architecture of NYN Track Dashboard

4. METHODOLOGIES

The proposed system follows a structured methodology to perform real-time monitoring and analysis of student behavior in a laboratory environment. The system processes video input from CCTV cameras and webcams and applies multiple computer vision and deep learning techniques to extract meaningful information.

Initially, the input video streams are converted into individual frames and preprocessed through resizing and normalization to ensure consistency across different modules. The processed frames are then passed through multiple specialized modules, including face recognition, emotion and engagement analysis, human object interaction, and action recognition.

Each module extracts specific features such as facial embeddings, facial landmarks, object coordinates, and skeletal keypoints. These features are further processed using threshold-based and learning-based approaches to derive meaningful outputs such as attendance status, engagement levels, detected actions, and object interactions.

The system employs both spatial and temporal analysis techniques to improve accuracy. Spatial features capture static information such as facial expressions and body posture, while temporal features analyze changes across consecutive frames to identify dynamic behaviors.

Finally, the outputs from all modules are integrated using a unified **Student_ID**, which acts as a primary key to associate all observations with a specific student. The structured outputs are stored in a database and visualized through the NYN TRACK Dashboard for real-time monitoring and analysis.

4.1 Dataset Collection

The proposed system collects real-time video data from two primary sources: CCTV cameras and webcams. CCTV cameras are used to capture the overall laboratory environment, including student activities, body movements, and interactions with surrounding objects.

In contrast, webcams are utilized to acquire detailed facial information required for face recognition and emotion analysis.

The collected data consists of continuous video streams, which are processed into individual frames for further analysis. Each frame contains visual information such as facial features, body posture, and object interactions involving items such as pens, notebooks, mobile phones, and keyboards.

4.2 Dataset Preprocessing

The collected video streams are processed frame by-frame to prepare the data for subsequent analysis. This step ensures that the input data is consistent, efficient, and suitable for different modules of the system.

Frame Resizing

To ensure compatibility with different deep learning models and maintain computational efficiency, each extracted frame is resized to appropriate resolutions based on the requirements of individual modules. The specific frame size configurations used for each module are summarized in **Table**.

S. No	Module	Frame Size
1.	Face Recognition	112 x 112
2.	Emotion & Engagement Analysis	224 x 224
3.	HOI (Human Object Interaction)	640 x 640
4.	Action Recognition	640 x 640

Table 2: Frame Resize for each module

Normalization

Normalization is applied to the input frames to standardize pixel values. This process scales the pixel intensity values to a uniform range, improving model stability and convergence during inference. It also reduces variations caused by lighting conditions and enhances overall performance.

Frame Extraction

The input video streams are converted into individual frames at a rate of **30 frames per second (FPS)**, enabling smooth and continuous real-time analysis. Each frame serves as an independent input to the system, allowing efficient processing across multiple modules.

4.3 Face Recognition and Attendance

The face recognition module is designed to accurately identify students by generating a robust facial representation for each individual. To achieve this, multiple face samples are collected and processed to create a stable embedding for each student. For each student, approximately 100 facial images are captured under varying conditions such as pose, lighting, and facial expressions. Each image is labeled using the corresponding Student_ID, forming a personalized dataset for every student.

Face Detection and Embedding Extraction

Face detection and feature extraction are performed using the **Buffalo_L**. For each input image, the model detects the face region and extracts a high dimensional embedding vector:

$$\text{Embedding} = f(\text{Image})$$

$$L = -\log \frac{e^{\cos(\theta_{y_i} + m)}}{e^{\cos(\theta_{y_i} + m)} + \sum_{j \neq y_i} e^{\cos(\theta_j)}}$$

Embedding Aggregation

To obtain a robust representation for each student, the embeddings from multiple images are averaged:

$$E = \frac{1}{N} \sum_{i=1}^{N} e_i$$

where:

- $N = 100$ (number of images per student)
- E = final aggregated embedding

Figure: Database of embedding extract of each student

Database Storage

The final embedding is stored in the database as a mapping:

$$\text{Student\_ID} \rightarrow E$$

Face Matching

During inference, the embedding of a detected face is computed as:

$$\text{Embedding}_{\text{test}} = f(\text{Image}_{\text{test}})$$

The similarity between the test embedding and stored embeddings is calculated using cosine similarity:

$$\text{Similarity} = \frac{\text{Embedding}_{\text{test}} \cdot E}{\| \text{Embedding}_{\text{test}} \| \, \| E \|}$$

If the similarity exceeds a predefined threshold, the identity is confirmed, and the corresponding **Student_ID** is assigned.

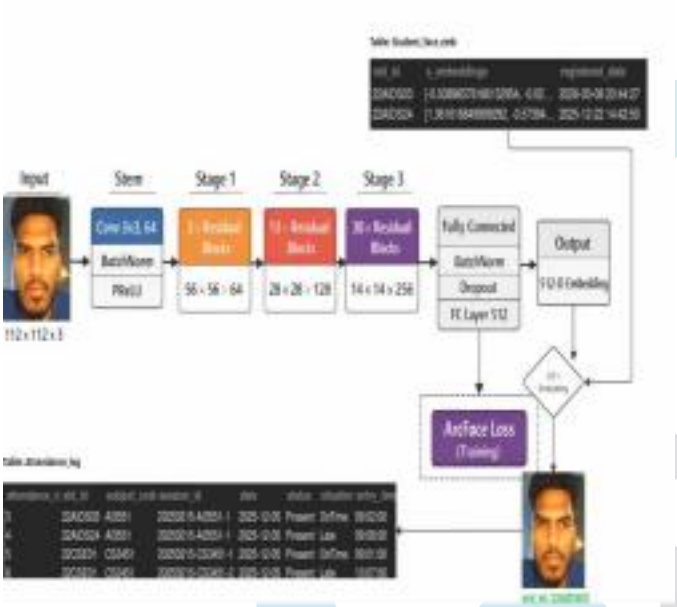

Figure 6: Workflow of face recognition and marking attendance

4.4 Emotion and Engagement Level

The emotion and engagement analysis module is designed to evaluate student attentiveness using facial features and expressions extracted from webcam input. The system combines geometric facial analysis with deep learning-based emotion recognition to determine engagement levels.

The system is trained using the AffectNet Dataset, which is one of the largest publicly available datasets for facial expression recognition. The dataset consists of a total of 29,069 images categorized into multiple emotion classes such as happy, sad, angry, fear, surprise, disgust, neutral, and contempt.

For effective training and evaluation, the dataset is divided into three subsets: 20,302 images for training, 4,357 images for validation, and 4,358 images for testing. This structured split ensures that the model generalizes well to unseen data and avoids overfitting.

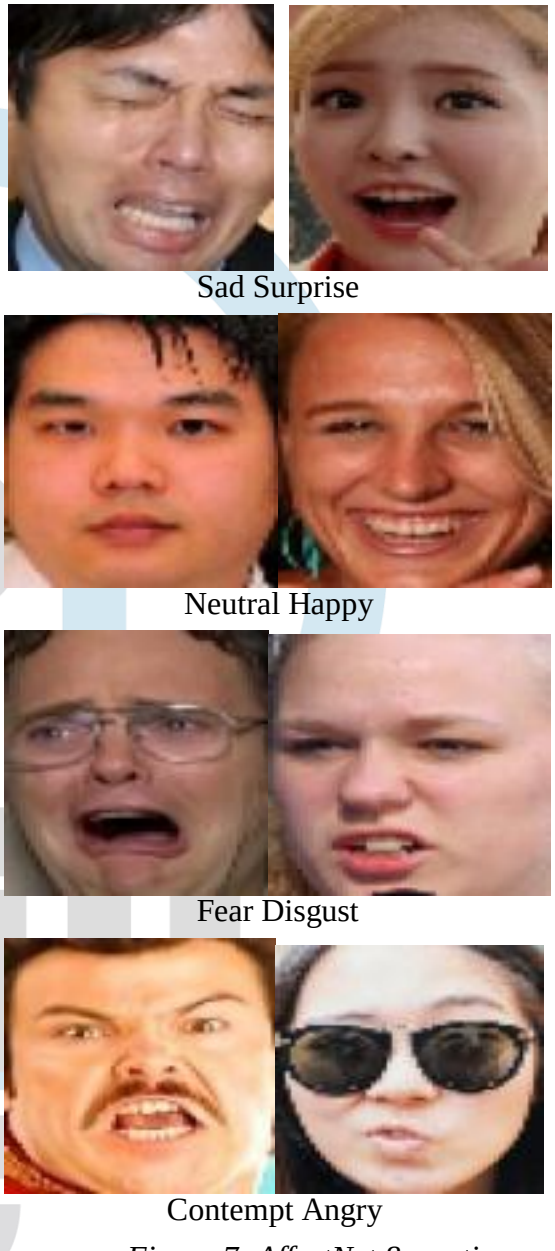

Figure 7: AffectNet 8 emotions

The system utilizes two deep learning models, namely ResNet50 and EfficientNet-B2, for facial emotion recognition. ResNet50 is effective in capturing deep spatial features using residual connections, while EfficientNet-B2 provides optimized performance through compound scaling of network dimensions

Figure 8: Extracting Mouth and Eye Features using MediaPipe

To improve prediction accuracy, an **ensemble learning approach** is adopted. Instead of relying on a single model, the outputs of both models are combined using **weighted averaging**. Each model produces a probability vector for all emotion classes, and the final prediction is obtained by combining these probabilities.

Input

Figure 9: Working architecture of Emotion and Engagement model

Figure 10: ResNet50 Model Accuracy and Training loss

Figure 11: EfficientNet-B2 Model Accuracy and Training loss

Ensemble Formula:

$$P\_final = w\_1 \cdot P\_ResNet50 + w\_2 \cdot P\_EfficientNetB2$$

$$\text{Predicted Class} = \arg\max(P\_final)$$

The outputs of both models are combined using weighted averaging. The final prediction is obtained by selecting the class with the highest probability.

Engagement Level

(i) Eye State Detection

EAR Formula

$$EAR = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2 \cdot \|p_1 - p_4\|}$$

Eye Aspect Ratio (EAR) is used to determine whether the eyes are open or closed. If the EAR value falls below a predefined threshold, the eyes are considered closed.

Figure 12: Ensemble Model confusion matrix for weight (ResNet50(0.45) + EfficientNet-B2(0.55))

(ii) Mouth State Detection

MAR Formula

$$MAR = \frac{\|p_{13} - p_{19}\| + \|p_{14} - p_{18}\| + \|p_{15} - p_{17}\|}{2 \cdot \|p_{12} - p_{16}\|}$$

Mouth Aspect Ratio (MAR) is used to detect whether the mouth is open or closed. A higher MAR value indicates yawning or talking behavior.

Image Alert

After obtaining the final emotion prediction, the system performs engagement analysis by combining emotion, eye state, and mouth state. Facial landmarks are extracted using MediaPipe, and Eye Aspect Ratio (EAR) and Mouth Aspect Ratio (MAR) are computed to determine whether the eyes and mouth are open or closed. Based on these features, engagement levels are classified into high, medium, low, or not engaged categories.

PSEUDO Code:

```
IF eye_state == "Closed" FOR > 3 seconds:
        engagement = "Not Engaged"

ELSE IF emotion in [Happy, Neutral]
        AND eye_state == "Open"
        AND mouth_state == "Closed":
            engagement = "High Engagement"

ELSE IF emotion in [Neutral, Surprise]
AND mouth_state == "Open":
```

engagement = "Medium Engagement"

ELSE IF emotion in [Sad, Disgust, Fear]
AND mouth_state == "Open":
engagement = "Low Engagement"

ELSE:
engagement = "Medium Engagement"

Engagement = f (Emotion, EyeState, MouthState)

To enhance monitoring, an alert mechanism is integrated. Drowsiness is detected when the eyes remain closed continuously beyond a threshold duration:

> 3 seconds

Yawning is detected when the mouth remains open beyond a predefined threshold:

If these conditions occur repeatedly within a time window, the system triggers drowsiness or yawning alerts, enabling real-time attention monitoring.

4.5 Human Object Interaction

Human-Object Interaction (HOI) is a crucial component for understanding human activities by analyzing the relationship between a person and surrounding objects. In this work, HOI is used to infer student activities such as writing, mobile usage, and working on a laptop based on object presence and spatial relationships.

The proposed system integrates two deep learning models: an object detection model (YOLOv8) and an object segmentation model (Mask R-CNN). YOLOv8 is used to detect general objects such as mobile phones, laptops, keyboards, and notebooks, while the custom-trained Mask R-CNN model is used for precise pen segmentation. These models operate simultaneously to extract meaningful object-level features from visual data.

Pen Segmentation

A custom-trained Mask R-CNN model is used for precise pen segmentation. The model is trained on the Poly Pen Data v5 dataset (1565 images, COCO format) using a ResNet-50 + FPN architecture. Images are preprocessed to 700×700 resolution and augmented using horizontal flipping, 90° rotations, random cropping (5–15%), and brightness/exposure adjustments.

The architecture consists of a ResNet-50 backbone with a Feature Pyramid Network (FPN), followed by a Region Proposal Network (RPN) and ROI heads for bounding box prediction and mask generation.

Figure 13: Architecture of Mask R-CNN

Figure 14: Training accuracy and loss curve for Mask R-CNN

During training, the model demonstrated stable convergence, with total loss decreasing from 1.645 to approximately 0.14. The model achieved a bounding box accuracy (AP) of 85.92% and a segmentation accuracy (AP) of 80.19%, indicating strong performance in detecting and segmenting pen objects.

Feature Extraction Output

The system returns object-level features including the object name, center coordinates (Cx, Cy), and confidence score for each detected object.

Figure 15: Pipeline of HOI

The system returns the object name, center coordinates (𝑥, 𝑦), and confidence score for each detected object. These coordinates are then used to analyze human-object interaction by measuring the proximity between the object center and the student's wrist position. If an object's center lies near the wrist, it indicates active interaction. Based on this relationship, the system determines the student's action, such as writing, using a mobile phone, or working on a laptop.

4.6 Action Recognition using skeleton joints

The Action Recognition module is designed to identify and classify human activities in a classroom environment. Unlike traditional approaches that rely solely on pose or visual features, the proposed system integrates multiple cues, including **skeletal joint information, human-object interactions (HOI), and facial engagement features**, to achieve robust and accurate activity recognition.

The module focuses on recognizing specific actions such as **working, sitting idle, using mobile, talking, sleeping, standing, walking, and writing**.

For human pose estimation and tracking, the system utilizes the YOLOv8 large model. This model is employed to detect individuals in each frame and extract **17 key skeletal joints** representing critical body landmarks.

To ensure continuity across frames, each detected individual is assigned a unique **Student_ID**, which enables consistent tracking of the same person throughout the video sequence. This identity preservation is essential for capturing temporal motion patterns and avoiding misclassification due to identity switching.

Figure: 16 skeleton joints of human

The extracted skeletal joints are used to model both **spatial structure** and **temporal dynamics** of human actions. For each frame 𝑡, the human skeleton is represented as $J_t = \{(x_i^t, y_i^t)\}_{i=1}^{17}$, where each joint is defined by its coordinates. To capture the spatial relationship between joints, pairwise distances are computed:

$$d_{ij}^t = \sqrt{(x_i^t - x_j^t)^2 + (y_i^t - y_j^t)^2}$$

These spatial features represent body posture and help distinguish static actions. Temporal dynamics are captured by analyzing joint movements across consecutive frames:

$$\Delta x_i^t = x_i^t - x_i^{t-1}$$

Additionally, joint velocity is defined as: $v_i^t = (x_i^t - x_i^{t-1}, y_i^t - y_i^{t-1})$

Δt

The complete sequence of skeletons is represented as ◆◆= {◆◆1, ◆◆2, ... , ◆◆◆◆}, which serves as input to the model. By combining spatial and temporal features, the system effectively captures both posture and motion, improving action recognition performance.

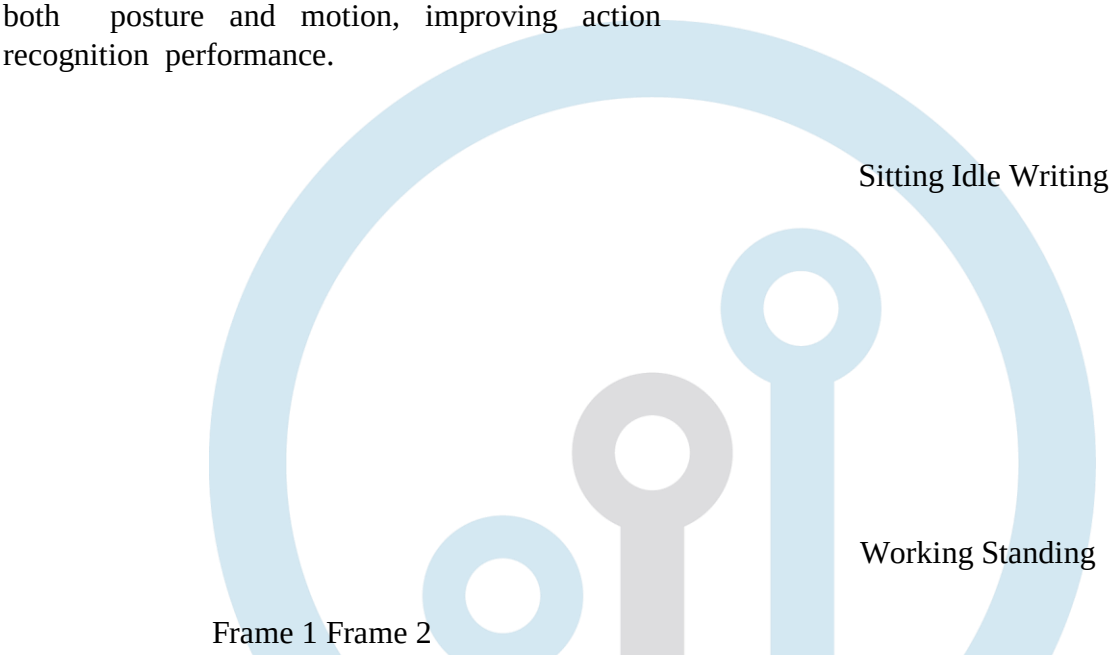

Figure 17: Temporal image

A custom dataset is constructed to train the action recognition model, consisting of six activity classes: **sleeping, mobile usage, working, standing, writing, and walking**. For each class, **100 video clips** are collected, with each clip having a duration of **10 seconds**, recorded at **30 frames per second (FPS)** and a resolution of **1080p**. The dataset is collected under a **real laboratory environment**, capturing variations in human posture and movement in practical scenarios.

Figure 18: Action dataset

4.7 NYN TRACK Dashboard

The **NYN TRACK Dashboard** serves as the central monitoring and visualization module of the proposed AI-based laboratory system. It integrates outputs from multiple intelligent models, including human action recognition, human-object interaction (HOI) detection, and emotion analysis, to provide a unified interface for real-time observation and analysis.

The primary objective of the dashboard is to transform raw model outputs into meaningful insights that can assist administrators and faculty in

understanding student behavior, engagement levels, and activity patterns within the laboratory environment. By presenting structured data

through interactive visualizations, the system enables efficient monitoring, quick decision making, and improved management of academic activities.

Figure 19: NYN Track Login Page

Additionally, the NYN TRACK Dashboard supports a role-based access system, allowing different users such as super administrators, department administrators, lab administrators, staff, and students to access relevant information based on their permissions. This ensures secure and organized handling of data while maintaining system scalability.Overall, the NYN TRACK Dashboard acts as the communication layer between AI models and end users, providing a comprehensive, real-time, and user-friendly platform for smart laboratory monitoring.

5. RESULTS

This section presents the experimental results of the proposed AI-based laboratory monitoring system. The system integrates emotion recognition, human action recognition, and human-object interaction (HOI) detection to analyze student activities in a real-time environment. The performance of each module is evaluated using standard metrics and validated under real laboratory conditions.

5.1 Dataset Description

The emotion recognition model is trained and evaluated using the AffectNet dataset, which contains a large number of facial images categorized into different emotional classes. The dataset consists of 29,069 samples, divided into training (20,302), validation (4,357), and testing (4,358) sets.

In addition to this, a custom dataset is collected in a real laboratory environment to evaluate the performance of action recognition and human object interaction modules. This dataset includes various student activities such as working, using mobile phones, taling, sleeping, standing, walking, and writing under realistic conditions.

Figure 20: Results of Engagement Level

From the experimental results, it is observed that the fusion of ResNet50 and EfficientNet-B2 significantly improves the performance of the emotion recognition system. Among all tested configurations, the weight combination of ***0.45 (ResNet50)*** *and* ***0.55 (EfficientNet-B2)*** *achieves the highest accuracy of* ***73.61%***.

5.3 Evaluation Metrics

The performance of the proposed system is evaluated using standard metrics including accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model predictions, while precision and recall evaluate the quality of positive predictions. The F1-score provides a balance between precision and recall, ensuring reliable performance evaluation.

S. No Model 1 Model 2 W1 W2 Accuracy
1. Resnet50 EffficientNet B2
2. Resnet50 EffficientNet B2
3. Resnet50 EffficientNet B2
4. Resnet50 EffficientNet B2
5. Resnet50 EffficientNet B2
0.4 0.6 72.53% 0.65 0.35 73.01% 0.5 0.5 73.47% 0.6 0.4 73.53% 0.45 0.55 73.61%

5.4 Emotion and Engagement Results

The emotion recognition module is implemented using a hybrid approach that combines ResNet50 and EfficientNet-B2 to leverage their complementary feature extraction capabilities. To determine the optimal configuration, multiple weight combinations are evaluated during the fusion process.

Table 3: Accuracy of Proposed model with Weights

This indicates that a slightly higher contribution from EfficientNet-B2 enhances the model's ability to capture fine-grained facial features, while ResNet50 contributes to strong high-level feature extraction. Balanced and near-balanced weight combinations (such as 0.5–0.5 and 0.6–0.4) also show competitive performance, demonstrating the robustness of the fusion approach.

Overall, the results confirm that multi-model fusion provides better generalization and improved accuracy compared to relying on a single deep learning model for emotion recognition.

5.5 Human Object Interaction (HOI)

Human-Object Interaction (HOI) is used to enhance activity recognition by analyzing the relationship between detected objects and human pose. The proposed system integrates YOLOv8 for object detection and Mask R-CNN for pen segmentation.

Figure 21: Results of HOI

The performance of the pen segmentation model is evaluated using the Poly Pen Data v5 dataset. The model achieves a ***bounding box Average Precision (AP) of 85.92%*** *and a* ***segmentation AP of 80.19%***, indicating accurate detection and segmentation of pen objects in real-world scenarios. Additionally, the training process shows stable convergence, with the loss decreasing from **1.645 to approximately 0.14**.

AP	AP50	AP75	APs	APm	API
85.6	98.66	97.15	0.00	86.50	86.32

Table 4: Bounding Box Detection Performance (%)

AP	AP50	AP75	APs	APm	API
80.19	98.66	94.18	0.00	85.79	80.87

Table 5: Segmentation Performance (%)

The object detection model successfully identifies relevant objects such as mobile phones, laptops, keyboards, and notebooks with high confidence scores in real-time. The system extracts object level features, including object labels, center coordinates (,), and confidence values for each detected instance.

By analyzing the spatial proximity between object centers and wrist keypoints, the system effectively determines human-object interactions. This enables accurate classification of activities such as writing, mobile usage, and working on a laptop.

The integration of object detection and segmentation significantly improves action recognition performance by incorporating contextual information. Overall, the HOI module demonstrates reliable and efficient performance in real-time laboratory environments.

5.6 Action Recognition Results

The performance of the proposed action recognition module is demonstrated through

qualitative results obtained from video sequences collected in a real laboratory environment. The system processes each frame to extract skeletal joint information and predicts the corresponding human activity based on spatial–temporal patterns.

The results show that the model is able to accurately identify actions such as sleeping, mobile usage, working, standing, writing, and walking under real-world conditions. The predicted action labels are overlaid on the video frames along with the corresponding skeletal structure, providing clear visual interpretation of the model's output.

Figure 22: Results of Action Recognition by combining of the HOI, Skeleton and Facial Features

5.7 Face Recognition and Attendance Results

The face recognition module is evaluated based on its ability to accurately identify students and automatically mark attendance in real-time laboratory environments. The system generates a robust facial representation for each student using multiple face samples captured under varying conditions such as pose, lighting, and facial expressions.

The experimental evaluation shows that the use of multiple facial samples per student improves recognition stability and reduces false identifications. By aggregating embeddings from approximately **100 images per student**, the system achieves consistent and reliable identification performance.

Figure 23: Attendance marked

During real-time testing, the system successfully detects and recognizes student faces and assigns the correct identity based on similarity matching. The cosine similarity-based matching approach ensures accurate identification by comparing the test embedding with stored embeddings in the database.

The attendance marking process is automatically triggered once a student is recognized. The system records the corresponding Student_ID along with a timestamp, enabling efficient and contactless attendance management. The system is also capable of handling multiple students simultaneously within the camera frame.

The overall performance demonstrates high recognition accuracy under normal laboratory conditions. However, slight performance degradation is observed in challenging scenarios such as low lighting, partial occlusion (e.g., masks), and extreme head poses. Despite these limitations, the system maintains stable and reliable operation in real-time environments.

5.8 NYN TRACK Dashboard Results

The NYN TRACK Dashboard serves as the centralized interface for monitoring and analyzing student behavior based on outputs from multiple modules, including face recognition, emotion analysis, action recognition, and HOI detection. The dashboard provides a structured and user friendly visualization of real-time and recorded data, enabling efficient observation and decision making.

The system displays key information such as **student identity, attendance status, detected actions, and engagement levels**. Each student is uniquely identified using *student_id*, ensuring consistent tracking across sessions. The dashboard dynamically updates based on live or processed video input, reflecting the current state of each individual.

Figure 24: Student Profile

Figure 25: Staff Profile

Figure 26: Alert & Instruction Logging

Figure 27: Overall Performance of Entire class for specific subject

Figure 28: Timetable for staff

Overall Results

The overall performance of the proposed AI-based laboratory monitoring system is evaluated by integrating emotion recognition, action recognition, human-object interaction (HOI), and face recognition with automated attendance effective and reliable performance in real-time laboratory environments.

The emotion recognition module achieves a maximum accuracy of **73.61%** using the fusion of ResNet50 and EfficientNet-B2, highlighting the effectiveness of multi-model integration. The action recognition module successfully identifies multiple student activities by combining skeleton based features with contextual object information, improving the distinction between similar actions.

The HOI module further enhances system performance by incorporating object-level context. The pen segmentation model achieves a ***bounding box Average Precision of 85.92%*** *and a* ***segmentation Average Precision of 80.19%***, enabling accurate detection of fine-grained interactions such as writing. The object detection model effectively identifies relevant objects in real time, contributing to improved activity inference.

The face recognition module provides reliable identification of students using aggregated facial embeddings, enabling accurate and automated attendance marking. The system operates efficiently in real-time, handling multiple students simultaneously and updating attendance records with minimal latency.

Overall, the integration of multiple deep learning modules improves the robustness, accuracy, and practical applicability of the system. Despite minor challenges under extreme conditions such as low lighting and occlusions, the system maintains stable

performance and demonstrates its suitability for real-world laboratory monitoring applications.

6. CONCLUSION

This paper presents an AI-based laboratory monitoring system that integrates emotion recognition, human action recognition, human object interaction (HOI), and face recognition with automated attendance marking. The proposed system leverages multiple deep learning models, including ResNet50, EfficientNet-B2, YOLOv8, and Mask R-CNN, to analyze student behavior in real-time laboratory environments.

The experimental results demonstrate that the system achieves reliable performance across multiple tasks. The emotion recognition module attains an accuracy of **73.61%** using a multi-model fusion approach, while the HOI module enhances action recognition by incorporating contextual object information. The face recognition module enables accurate identification of students and supports automated attendance marking, reducing manual effort.

The integration of multiple modules improves the overall robustness and effectiveness of the system, allowing it to handle real-time video streams and multiple users simultaneously. The system proves to be practical and scalable for real-world applications in educational environments, particularly for intelligent monitoring and activity analysis.

In conclusion, the proposed system provides an efficient and automated solution for laboratory monitoring, combining advanced computer vision techniques with real-time processing capabilities. It demonstrates strong potential for improving student engagement analysis, attendance management, and overall lab supervision.

7. FUTURE WORKS

The proposed system demonstrates strong performance in laboratory monitoring; however, several enhancements can be explored in future work to further improve its applicability and scalability. One of the primary directions is the deployment of the system across multiple educational institutions, enabling large-scale monitoring and centralized management of student activities.

In addition, the development of a dedicated mobile application can enhance accessibility by allowing administrators and staff to monitor activities, receive alerts, and view reports in real time from handheld devices. This would improve usability and provide greater flexibility in system interaction.

Furthermore, the system can be extended to support working professionals in industrial environments. By adapting the existing framework, the system can be utilized for employee monitoring, productivity analysis, and safety compliance in workplaces, thereby broadening its real-world applications beyond academic settings.

Overall, these future enhancements aim to make the system more scalable, accessible, and adaptable to diverse environments.

8. REFERENCE

[1] Nur Fitria, Tira. (2021). Artificial Intelligence (AI) In Education: Using AI Tools for Teaching and Learning Process.

[2] Oloyede, Adetokunbo & Nureni, Yekini & Olawale, Onadokun & Akinleye, Akinyele. (2017). Networking CCTV Cameras & Passive Infra-Red Sensors for E-classroom Monitoring System: Proactive Approach to Quality Assurance in Education System. International Journal of Advanced Networking and Applications. 8. 3213-3219.

[3] Fu, Wentao & Jiang, Hui. (2024). Computer vision recognition in the teaching classroom: A Review. EAI Endorsed Transactions on AI and Robotics. 10.4108/airo.4079.

[4] Shelke, Kunal & Khan, Yusuf & Kapse, Dr. (2025). Student Management System: A WebBased Solution for Academic Administration. International Journal of Scientific Research in Science and Technology. 12. 50-54. 10.32628/IJSRST251222719.

[5] Kotramma T S, Azizkhan F Pathan, Vinayaka G S, Varshitha A M, Vikram K, 2023, Automatic Attendance System Using Machine Learning,

INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 11, Issue 05 (ICEI – 2023),

[6] Sharma, Prabin & Joshi, Shubham & Gautam, Subash & Filipe, Vítor & Reis, Manuel. (2019). Student Engagement Detection Using Emotion Analysis, Eye Tracking and Head Movement with Machine Learning. 10.48550/arXiv.1909.12913.

[7] D. Kayande and S. Kukreja, "Design of an integrated multi-modal machine learning framework for real-time student engagement evaluation and learning outcome optimizations," MethodsX, vol. 15, p. 103588, 2025, doi: 10.1016/j.mex.2025.103588.

[8] Dang, Thai-Viet. (2023). Smart Attendance System based on Improved Facial Recognition. 4. 46-53. 10.18196/jrc.v4i1.16808.

[9] Alruwais, Nuha & Zakariah, Mohammed. (2024). Student Recognition and Activity Monitoring in E-Classes Using Deep Learning in Higher Education. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3354981.

[10] Jonathan, Jonathan & Lim, Andreas & Paoline, & Zahra, Amalia. (2018). Facial Emotion Recognition Using Computer Vision. 46-50. 10.1109/INAPR.2018.8626999.

[11] Pandey, Amit & Gupta, Aman & Shyam, Radhey. (2022). FACIAL EMOTION DETECTION AND RECOGNITION. 7. 176-179. 10.33564/IJEAST.2022.v07i01.027.

[12] Ozaki, H., Tran, D. T., & Lee, J. H. (2024). Effective human–object interaction recognition for edge devices in intelligent space. SICE Journal of Control, Measurement, and System Integration, 17(1),1–9. https://doi.org/10.1080/18824889.2023.2292353

[13] Shehata, Hesham & Abdolrahmani, Mohammad. (2025). Improvement of Human Object Interaction Action Recognition Using Scene Information and Multi-Task Learning Approach. 10.48550/arXiv.2509.09067.

[14] Khan, B.A.; Jung, J.-W.Deep Learning-Based Human Activity Recognition Using Dilated CNN and LSTM on Video Sequences of Various Actions Dataset. Appl. Sci. 2025, 15,12173. https://doi.org/10.3390/app152212173

[15] Yi, Gao & Wu, Haitao & Wu, Xinmeng & Li, Zilin & Zhao, Xiaofan. (2023). Human action recognition based on skeleton features. Computer Science and Information Systems. 20. 537-550. 10.2298/CSIS220131067G.

[16] Zhang, Heng & Wang, Fa. (2025). Research on Human Action Detection and Recognition Methods Based on Interpretable Artificial Intelligence. Journal of Combinatorial Mathematics and Combinatorial Computing. 127a. 1143-1158. 10.61091/jcmcc127a-066.

[17] Hasnine, M.N.; Nguyen, H.T.; Tran, T.T.T.; Bui, H.T.T.; Akçapınar, G.; Ueda, H. A Real-Time

Learning Analytics Dashboard forAutomatic Detection of Online Learners' Affective States. Sensors 2023, 23, 4243. https://doi.org/10.3390/s23094243