IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 16 November 2023, accepted 8 December 2023, date of publication 17 January 2024, date of current version 16 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3354981

RESEARCH ARTICLE

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

NUHA MOHAMMED ALRUWAIS[1] AND MOHAMMED ZAKARIAH[2], (Member, IEEE)

[1]Department of Computer Science and Engineering, College of Applied Studies and Community Services, King Saud University, Riyadh 11633, Saudi Arabia
[2]Department of Computer Science and Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11633, Saudi Arabia

Corresponding author: Nuha Mohammed Alruwais (nalrowais@ksu.edu.sa)

This work was supported by King Saud University, Riyadh, Saudi Arabia, through the researchers supporting under Project RSPD2024R608.

ABSTRACT Monitoring student activity manually constantly is a laborious endeavor. Over the past few years, there has been a rapid expansion in the usage of cameras and the automatic identification of odd surveillance behavior. Different computer vision algorithms have been used to observe and monitor real-world activities. Most educational institutions are already offering online programs to lessen the impact of this epidemic on the education industry. However, ensuring that students are correctly identified, and their behaviors are monitored is crucial to make these online learning sessions dynamic and equivalent to the conventional offline classroom. In this study, we have introduced brand-new deep learning-based algorithms that continuously track a student's mood, including rage, contempt, happiness, sorrow, fear, and surprise. The effectiveness of student identification and activity monitoring in online classrooms was also studied using deep learning and a CNN model that reaches 99% accuracy. Our approach was superior because of its many convolutional layers, dropout regularization, and batch normalization. It caught crucial properties and decreased overfitting. By identifying them more frequently, deep learning techniques can enhance student engagement and learning outcomes in e-learning situations, according to the research. With these techniques, educators and instructors may support students more effectively by better comprehending their behavior and offering specialized and individualized support, improving academic performance and student activity evaluation.

INDEX TERMS Student recognition, student activities monitoring, deep-learning, engagement detection, digital classroom, e-class.

I. INTRODUCTION

One of the significant issues with online learning is how to increase the quality of learners' participation in their educational activities by monitoring their actions and acknowledging their efforts. Online lectures, tests, and viewing tutorial videos are just a few educational activities students can engage in. During such activities, students may get disengaged. Eye movements, eye tracking, facial expressions, gaze patterns, and body motions can determine how engaged or disengaged a person is [1], [2], [3]. The teacher can assess student participation and activity in the classroom observation area. In an e-learning class context, a real-time system is an ideal choice [2]. Engagement detection may be used in various settings, including monitoring student activities and behavior and evaluating autistic people's eye tracking.

Student's learning styles and distinctive qualities influence the characteristics of learning activities [4], [5]. A pupil can quickly reach a high-learning recognition state, but they cannot maintain it for very long. While different students might find it challenging to achieve high levels of focus, once they do, they can sustain them for a considerable amount of time.

The associate editor coordinating the review of this manuscript and approving it for publication was James Harland.

Some students need help to remain academically engaged and are readily influenced by other events. I think that examining students' [6], [7] full learning attention in class may quantify the teacher's teaching effect and provide a quantitative index to measure the effect of classroom teaching. Teachers can also improve their teaching strategies by paying more attention to, recognizing, and monitoring [6], [8] the actions of their students. Investigating each student's distinctive learning attention can aid in developing individualized learning programs for each student.

Researchers have been drawn to the study of online learning, which has produced valuable findings. In reality, determining the level of student concentration in class is challenging. Accurately determining each student's level of understanding in the class takes time and effort. Several researchers have undertaken studies [9], producing valuable findings on identifying learning attention monitoring [10]. The study conducted by [11] examined the current advancements in attention recognition in both domestic and international contexts. The research focused on two important topics: attentional recognition based on behavior and attentional recognition based on facial expression. The results also suggested possible future directions for the study of attention recognition. The virtual environment of online learning and the physical setting of conventional classroom instruction make up the two primary research settings. Researchers have looked into the effectiveness of picking up visual and aural cues in virtual learning environments, and their findings have made significant progress in our understanding of how attention is detected in these settings. The author [12] has previously offered a thorough methodology for building a structural equation model that is based on the Triadic Theory of Learning. I may use this model to study the effectiveness of online learning, classify it into deep and machine learning states, and evaluate the usefulness of online teaching. About 700 individuals who were enrolled in national universities and pursuing degrees in management and science made up the study's sample size [13]. This strategy asserts that students' attentiveness has a significant impact on their learning enthusiasm and is significantly influenced by the instructional content and methods used by teachers. It is recommended that improving the online education platform's design might help things become better in this area. The algorithm for the online education platform's attention detection system was tested in a simulated environment. The experiment demonstrates that this approach may assess students' attention in the classroom, enhance the caliber of instruction, and enhance students' learning results. In [14] presented a machine-learning-based technique to monitor students' learning activities in an online class. Students' eye states are classified using a trained SVM model, which is utilized to extract features from the various eye states using the Gabor wavelet approach.

Furthermore, a recent study by [15] introduced an innovative approach that leverages biometric characteristics such as pupil detection, head movement recognition [16], facial emotions recognition [17], and eye gaze movements to assess the level of student engagement in academic pursuits [18]. This strategy makes use of a variety of approaches, including machine learning, OpenCV, Haar cascade, local binary patterns, and Principal Component Analysis (PCA). In order to identify the faces of students in high-resolution video and comprehend their facial expressions, a proposed approach by [19] involves the implementation of a facial expression recognition system and a head posture estimation system. This system aims to enhance the learning experience in the classroom using new methods. The categorized facial expression is used to help the instructor analyze the level of student learning and offer recommendations for enhancing the effectiveness of instruction. Some investigations have centered on the students' attention in regular classes since they are still the most significant instructional environment.

The method described in reference [20] for assessing students' level of attention during learning activities entails the manipulation of students' head positions, namely by elevating and lowering them. The vertical position of a student's head in the classroom is detected at regular intervals of fifty frames. The researcher conducted an analysis on the variability of students' attention levels and the frequency of times characterized by high attention inside the classroom. This analysis was conducted alongside a comprehensive evaluation of students' academic performance, specifically their grades. The findings of this study revealed a favorable correlation between students' attention levels and their grades. Additionally, it was shown that each class has its highest levels of attention during the initial ten minutes, the next 15-20 minutes, and the last five minutes. Using the information above, teachers may adjust their instruction to suit the needs of their pupils better. According to a method provided by [21], each student's learning attention may be determined by observing how their eyes open and close in response to head movements. The accuracy of this algorithm increased from the conventional learning attention technique's ninety-four percent to ninety-four percent when the experimental sample was sixty individuals in a class [22].

According to the conventional classroom teaching scenario, a combination of the class states and the student's head posture criteria can determine if they are paying attention to what is being taught. I have face detection techniques in our toolset. To extract head posture parameters, a head pose recognition model is available [22]. The lecture, interaction, practice, and transcribing states of a class may all be separated using voice recognition software.

The strategy suggested in this study is predicated on the following presumptions: I believe that students should look towards the instructor when the teacher is lecturing or interacting with the class, such as by asking a question and then waiting for a response. It demonstrates student engagement and enthusiasm for the subject matter, which the instructor may monitor. A student should gaze at the chalkboard while taking notes before lowering his head to copy. He is likewise only focused if he is at this level of attention. In addition, Figure 1 shows the suggested architecture for deep learning technology-based student identification and activity tracking in an online class.

The following are major contributions of this study:

1) The paper proposes a novel deep learning-based approach for student recognition and activity monitoring in e-classes.

2) The proposed approach achieves a high accuracy of 99% in identifying students and their activities.

3) The approach utilizes convolutional neural networks (CNNs) with dropout regularization and batch normalization to capture essential features and reduce overfitting.

4) The research demonstrates the potential of deep learning techniques to enhance student engagement and learning outcomes in e-learning environments.

5) The paper provides a valuable contribution to the field of educational technology and e-learning.

The subsequent sections of the document are organized in the following manner: Section II provides an analysis of the relevant scholarly literature. Section III of the document provides coverage of the datasets, while Section IV delves into the process that facilitates the system's ability to identify and interpret students' behaviors. The outcomes of the study are presented in Section V, followed by a comprehensive analysis in Section VI, and ultimately, the findings are summarized in Section VII.

II. LITERATURE REVIEW

The study of identifying aberrant student behavior is crucial in the fields of image processing and video analysis. Because of its critical importance in the realms of human-computer interaction and surveillance, the study of tracking and assessing object activity through films has long been a research focus. A multitude of scholars have conducted extensive research on the challenges associated with activity tracking and recognition in diverse educational domains. The scholarly discourse on vision-based activity recognition has seen advancements in picture representation strategies and classification methodologies, which have been influenced by previous research on global, local, and depth-based activity representation techniques. In real-time, [1], [2], [4] describes an attention-tracking system that needs a primary web camera. This system tolerates inaccurate attention state classifications brought on by blinks and is a scale and rotation invariant. The Haar cascade classifier is employed to ascertain the region of interest corresponding to the face. The determination of the locations of both eyes is achieved through the utilization of a Haar cascade classifier [4], which has been specifically developed for the purpose of identifying eyes. The authors in [6] initially identifies the coordinates of the inner and outer corner points of the eyeballs within a frame. Subsequently, it identifies the corner nodes in subsequent frames. The mean intensity within the region of interest (ROI) is determined following the implementation of subtraction procedures to assess pupils' recognition and attentiveness. A feedback channel is incorporated into an active attention-tracking system to deliver alarm signals to the user upon detection of inattention [8]. The database will record the level of attention and any associated bookmarks for the e-learning material.

Facial landmarks were suggested in [11] to measure pupil attention spans. In this case, the lips and eyes on the face are highlighted as regions of particular importance. The lips and eyes' threshold values and aspect ratios were also computed. The student's face was located using a face detector incorporating the linear SVM and HOG [12] algorithms. Based on the location of the markers for the mouth, eyes, and face, student states were identified as usual, napping, or yawning.‘‘A defined number of successive frames that should not exceed the limit is the foundation for yawning, drifting off, and paying attention. Students and teachers can see a display that shows the states that have been determined.

Based on the Kinect sensor, a student activity tracking system has been created [9]. The system monitors each student's activity in the online classroom using some cameras. A decline in student participation was detected by the monitoring system based on visual markers [13]. Every student in the online class may now be monitored thanks to a system for video analysis individually. Through technology, teachers can keep an eye on students from various viewpoints. A high-definition Microsoft Kinect 2 was set up in front of the pupils to record movies and monitor visual cues. The device covered the students' frontal views and complied with the requirements for detecting facial expressions, tracking gaze, and body movements. A second video source has been employed: a regular HD digital video camera recording to a DV tape [14]. To fully monitor student conduct, the two video streams were combined. The detected properties have been recovered offline [23] by combining the Kinect software development kit with MATLAB.

In terms of deep learning, recent years have seen tremendous advancements. In many areas, including pattern recognition, picture and object identification, natural language processing, and audio recognition, deep learning has attained high levels of accuracy and displayed amazing performance. Deep learning models may have an advantage over vision-based approaches due to their capacity for automatic feature extraction and their capacity for machine learning by example [15]. The computer vision-based techniques employ manually created low-level characteristics for categorization. On the other hand, deep learning, a form of AI, frequently abstracts high-level characteristics from low-level ones to achieve high accuracy for classification tasks [24].

This work [22] offers a machine learning method for identifying students using mood analysis, eye tracking, and head movement data gathered by a camera. The system was evaluated using a digital learning scenario. The results show that pupils with the highest concentration indices had the most outstanding test outcomes. Similarly, [25] has suggested a recognition detection framework that may be utilized in meetings by evaluating people's mental states to improve their

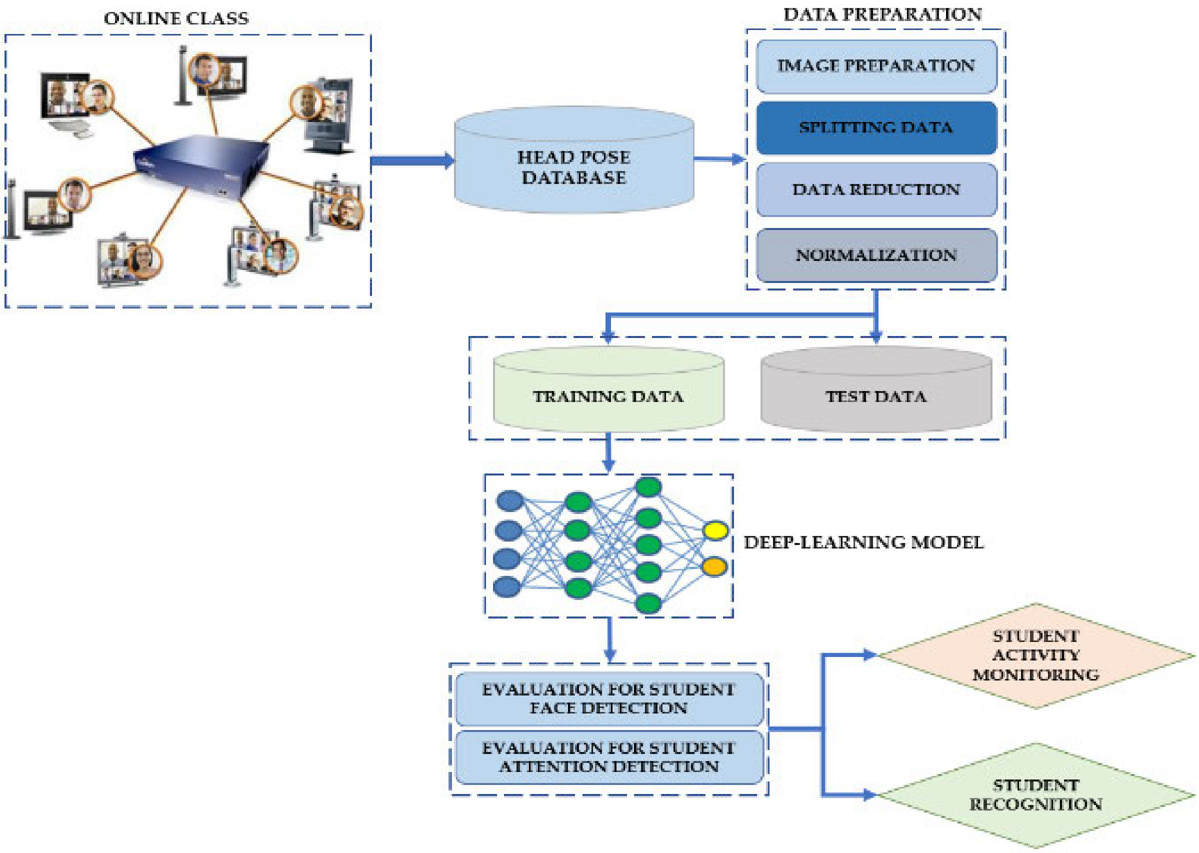


FIGURE 1. Framework for student recognition and activity monitoring using deep learning in an online class.

efficacy. Sound and information from 2D and 3D images are also integrated into this framework. An ensemble model for engagement detection has been put out by [26] and uses facial and body tracking. By combining cluster-based frameworks with neural networks, they did, however, attain a meager MES of 0.0813. For better engagement detection, some researchers have developed CNN-based techniques. A student recognition detection method for digital learning environments based on CNN has been suggested by [27]. To categorize students' cognition throughout their lessons in a digital environment, they have presented three distinct models: all CNN, network-in-network CNN, and very deep CNN [27], [28], [29]. The accuracy levels for the three levels of decisions were 92.42% for those who were not engaged, 90.22% for those who were engaged ordinarily, and 96.45% for those who were very involved. To identify student involvement, the experimental findings of five distinct CNN models have been evaluated in another study [30]. CNN produces the most accurate findings. Table 1 below lists previous references, their datasets, methods, and findings.

To better understand how technology is evolving in this area, Doljanin et al. [35] investigated enabling technology solutions for an adaptive intelligent agent in e-learning. Mane and Surve [36] provided important insights into tracking student engagement through the presentation of a novel approach for engagement detection using video-based head movement estimation. Robust head pose classification from low-resolution images was carried out by Khaki et al. [37], shedding light on the issues and potential solutions in this field. Convolutional neural networks were used by Lasri et al. [38] to investigate facial emotion recognition in students, highlighting the importance of deep learning methods in comprehending student emotions. In order to further the study of emotional dynamics in online learning environments, Yang et al. [39] introduced an emotion recognition model based on facial recognition.

The assessment of the literature reveals a number of unresolved issues in the field of image processing and video analysis for identifying atypical student behavior in online classrooms. These gaps include the need to fully utilize the potential of multi-modal data fusion for an in-depth understanding of student behavior, the need to optimize engagement detection methods for improved accuracy and efficiency, and the need to investigate deep learning models to supplement conventional computer vision-based techniques. Furthermore, in order to guarantee the robustness and generalizability of results, larger and more varied datasets designed for student behavior analysis must be developed. In addition, to properly and effectively track student behavior during online learning sessions, attention tracking technologies must

TABLE 1. List past references with their datasets, methodology, and results.

Ref	Dataset/ Model Architecture	Methodology	Results
[1]	- Kaggle Eye-Image dataset. - Around 14500 images were used in the experiment related to eye-image for real-time detection. - For Pre-processing step, around 4500 eye images were used.	- Deep-Learning, CNN Model, MobileNet Model, Google Colab Environment. Confusion Matrix. HAAR cascade Objection detection method.	- The model was trained through hundred EPOCHS, where the batch size was approximately 120. - MobileNet Accuracy: ninety-nine percent, and ninety-eight percent on validation data.
[4]	- Around fifteen students took part in the experiment. - The facial expression has seven categories: Sad, surprised, Angry, Fear, Happy, Disgusted, etc.	- Haar Cascade Algorithm, CNN, Deep-Learning, Viola and Jones Algorithm	- Out of fifteen students, around sixty percent of students were highly engaged.
[6]	- As per the FER2013 dataset, around 36,000 samples were taken. - The Engagement Recognition dataset was also used in this experiment.	- VGG Model, CNN, Deep-learning, Support Vector Machine, HOG.	- CNN Accuracy: 94% for engaged students. - Confusion Matrix Engagement model: 88.12%. - VGGNet Model Accuracy: 90.34%. - SVM+HOG Accuracy: 93.56%.
[8]	- In this experiment, around twelve students took part.	- Deep-Learning, CNN, ResNet, MobileNet, Automatic speech recognition models	- There are around 45% average Learning Students. - Human face detection Accuracy: 98%.
[9]	- The dataset comprises eighteen subjects. - The training dataset size is one hundred and twenty minutes.	- MATLAB, SVM, Gabor Features, Kinect Sensor	- Engagement Accuracy: 0.8.
[31]	- The dataset contains a recording of around thirty subjects.	- LSTM, CNN, RNN, Deep-learning, Average Pooling Layer	- The training set consists of Thirty subjects and around six subjects for test data.

be implemented in real-time. A more in-depth understanding of the various degrees of attention and engagement is also necessary in order to personalize interventions and feedback for pupils. Last but not least, it is crucial to address privacy and ethical issues related to employing cameras and sensors to observe student behavior, which is why frameworks should be developed that place a priority on safeguarding students' rights and consent in future research projects.

The suggested strategy fills many significant research gaps found in the literature review. It uses deep learning-based techniques, in particular a CNN model, to watch and recognize pupils' moods and facial expressions continually, giving important insights into their behavior. As a result of their success in image recognition and other fields, deep learning models can now be explored for their potential in studying student behavior. This study satisfies the criteria for different datasets by utilizing a large-scale dataset of 11,342 grayscale photos from 20 individuals. Additionally, the CNN model enables fine-grained behavior analysis that transcends straightforward binary categorizations, which is essential for comprehending student engagement and attention levels. The study's emphasis on student identification and behavior tracking also highlights the significance of privacy and ethical considerations while managing sensitive data, even though they are not highlighted explicitly.

III. DATASET

A. UNPA HEAD POSE DATABASE

Head postures were estimated in this investigation of student involvement in online classrooms, and faces were recognized. Images that provide information on the head pose are part of the UPNA Head Pose Database. Researchers at the University of Navarra in Spain developed it to help with head posture estimation research, a crucial issue in computer vision and human-computer interaction. The dataset consists of 11,342 grayscale images of 20 people (10 women and ten men), all of whom were snapped against various backdrops, lighting setups, and stances. The head's yaw, pitch, and roll angles are noted in each shot, indicating how the head is oriented with respect to the camera. A commercial 3D sensor and custom algorithms were used to predict the head postures. For educational and research needs, the UPNA Head Pose Database is publicly accessible.

For a number of convincing reasons, I have decided to employ the UPNA Head Pose Database in this investigation. The database, in the first place, makes it easier to conduct head pose estimate research, which is precisely aligned with the study's main goal of estimating head postures in the context of student participation in online classrooms and face recognition. The database also contains a vast and varied collection of grayscale photographs, comprising 11,342 images from 20 people, including both men and women. These photographs were taken in a variety of settings, offering a large dataset for study. Additionally, each image in the database is labeled with the yaw, pitch, and roll angles of the head, which is essential information for precise head pose estimation. The information was also gathered utilizing a commercial 3D sensor and unique algorithms, guaranteeing the accuracy and precision of the data and making it appropriate for advanced modeling and analysis. Additionally, the public accessibility of the UPNA Head Pose Database enables other researchers to access and use the same dataset, promoting comparison and validation of findings across other research projects.

The choice to forego using the entire recorded file of the students' video collection was well thought out and motivated by a number of strong arguments. First off, the dataset's primary use was for head tracking and head position studies, not the study's focus on student recognition and attentiveness detection in online classrooms. Second, crucial data for efficient face recognition and attention detection tasks was missing from the recorded dataset, including metadata and labels. Additionally, the dataset's format proved difficult to modify for the research because it was incompatible with the tools and techniques that were selected. The choice was also influenced by moral considerations about student recordings' privacy and permission. Additional factors to take into account included resource restrictions, computing requirements, and storage needs.

B. DATA DESCRIPTION

This collection features movies of ten persons, six of them are men and four of whom are women. It was designed with head tracking and posture studies in mind. The database contains twelve video files for each individual. The x, y, and z axes were used to move and rotate the heads of the figures in these films. Each video in the database lasts 10 seconds, or 300 frames, and has a resolution of 1280 × 720 and a frame rate of 30 frames per second.

The database includes text documents that include the measurement values received by the sensors for each video. For each video, three different types of text documents are shared:

- 2D measurement results (∗_groundtruth2D.txt): Each row corresponds to one frame, totaling 300 rows. The letter TAB separates each row's 108 columns. The x and y coordinate values of 54 face landmarks are represented in these 108 columns as (x1 y1 × 2 y2 ...... × 54 y54).

In this work, posture estimation and application will not be included. For the development of face and attention detection in online classrooms, I shall apply it in this instance. As a result, I cannot apply the remaining data from this database.

C. DATA PREPARATION

Our goal in this study is to:

1) Identify students;
2) Determine if students are paying attention:

With the dataset I have for this study, which consists of students striking various positions, looking at the screen in various ways, posing down and up, and moving their necks from left to right, I can create picture folders by extracting frames from videos using OpenCV. Following is a list of the two folders.

3) Student Recognition:

Figure 2 and Table 2 provide a list of images of each student with their names labeled next to them.

•Attention Recognition

As can be seen in Figure 3 (a), all of the student videos' image frames were examined and separated into two groups namely (1) Engaged or (0) Not Engaged.

We created a dataset that reflected student participation by employing the UPNA Head Pose Database. I took 100 randomly selected frames from the UPNA Head Pose Database for each participant, resulting in a dataset of 1000 images. Each image was labeled by five human labelers as (1) Engaged or (0) Not Engaged.

TABLE 2. List of student names.

Folder	Name
0	Student-1
1	Student-2
2	Student-3
3	Student-4
4	Student-5
5	Student-6
6	Student-7
7	Student-8
8	Student-9
9	Student-10

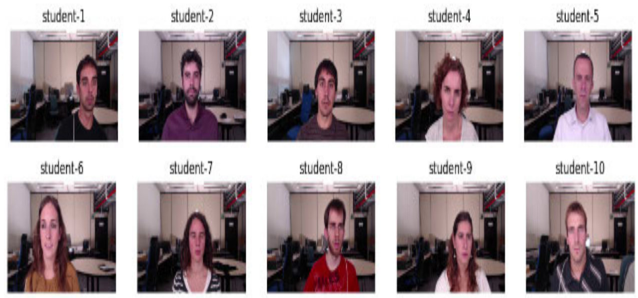
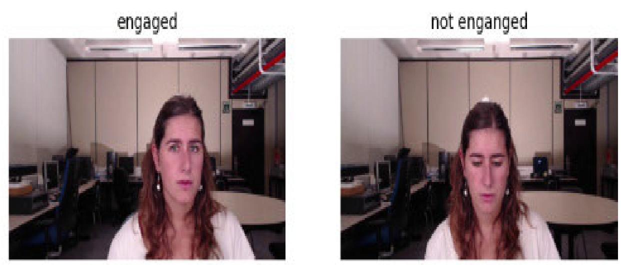


FIGURE 2. Each student's face visualization.

Folder	Name
0	engaged
1	not engaged

(a)



(b)

FIGURE 3. Student face pose and expression for attention classification (a) the dataset stored as (1) Engaged or (0) Not Engage; (b) Sample image of Engaged or not engaged.

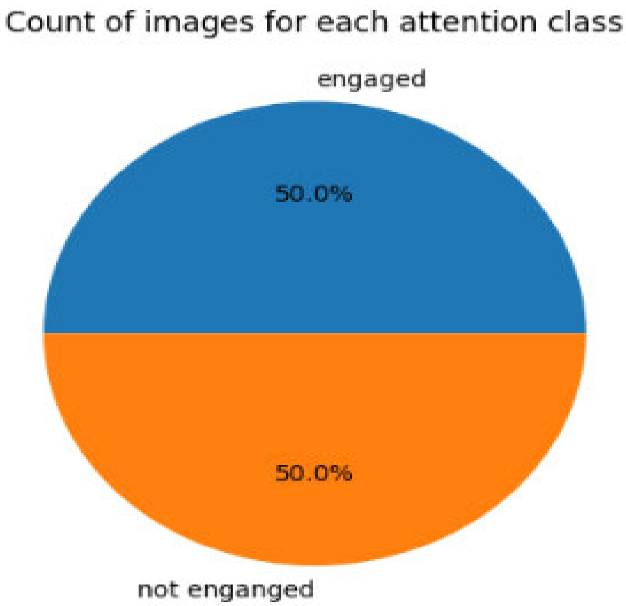


FIGURE 4. Each class distribution to the datastore.

D. DATA PROCESSING

1) IMAGE PREPARATION

Initially load every video file using OpenCV's VideoCapture function in the data preparation stage for your student face recognition research. The number of frames in each video is the next thing I count. Then, 50 randomly chosen frames from each video are saved as picture files. To do so, I use a set to set the location of each selected frame, read to read each selected frame and write to store it as an image file. The release function is then used to release each video. Using this method, I may extract individual frames from videos and prepare them for future analysis.

The percentage of each student's images in the train set is 10%. The size of each image in each folder containing the names of the students involved in developing a facial recognition model. The distribution of each student picture datastore within the total data store. To illustrate how each class image contributes to the entire datastore of student attention detection, a pie chart, as seen in Figure 6, is employed in this instance.

To display the contribution of each class image to the entire datastore of student attention detection, a second pie chart is employed, as illustrated in Figure 4.

2) RESIZING

To ensure uniformity in image dimensions, I apply image resizing. All images in the dataset are standardized to a specific resolution, such as 224 × 224 pixels. Resizing the images to a consistent size facilitates efficient and effective training of the facial recognition model.

3) NORMALIZATION

Normalization is a crucial step in deep learning data preprocessing. I scale the pixel values of each image to a common range, typically [0, 1] or [−1, 1]. This process removes variations in pixel intensity and helps in avoiding biases during model training caused by pixel value differences.

4) CLASS BALANCING

Addressing class imbalances is essential to prevent the model from being biased towards overrepresented classes. In this step, I employ class balancing techniques to ensure that each student's facial images contribute equally to the training process. Techniques like oversampling, undersampling, or using class weights during training are applied to achieve a balanced representation of each class in the dataset. This enables the model to learn from a diverse range of examples for each class and improves its overall performance and generalization capability.

E. DATA SPLITTING

To create training and test image folders, go through the following steps:

- Enter the location of the dataset directory, which houses the image files.
- Determine the output directory's training and testing image folders.
- Configure the image ratio to be tested with.
- Using the glob function, get a list of the image files located in the dataset directory.
- Divide the image files into training and testing sets using the Scikit-Learn library's train-test split function.
- The training and testing folders should be made in the output directory using the os.makers function.
- The training images should be copied to the training directory using the shutil.copyfile function.
- Copy the testing images into the testing directory using the shutil.copyfile function.

Overall, the code creates two folders for training and testing by randomly dividing the images in the original dataset directory using a specified ratio and copying them to the corresponding directories. Figure 5 displays the training and testing distribution for detecting student attentiveness.

In contrast, Figure 6 shows the distribution of student recognition during training and assessment.

F. IMAGE DATA GENERATION

The training and testing pictures are subjected to various transformations using an Image Data Generator object to facilitate the machine learning model's ability to learn from the data.

We make an ImageDataGenerator object and set its rescale parameter to 1/255 for the training data, which will scale the image's pixel values to 0 and 1. Then, using the flow_from_directory() function, I build a directory iterator that loads 512-image batches at a time from the training_datastore directory. Grayscale conversion, 48 × 48 pixel resizing, and categorical class mode labeling are all done to the images.

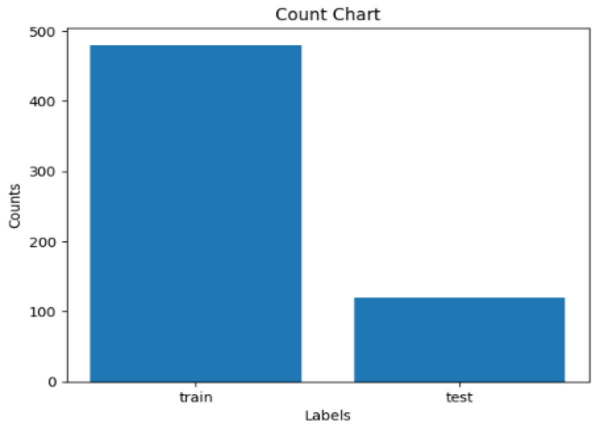


FIGURE 5. Training and testing distribution for student attention detection.

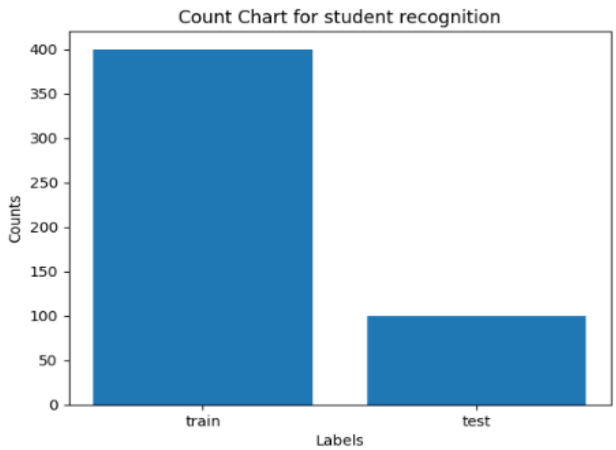


FIGURE 6. Training and testing distribution for student recognition.

With the same rescaling parameter for the testing data as I used for the training data, I constructed a new ImageDataGenerator object. Then, using the same flow_from_directory() function, I build a directory iterator that loads photos from the training_datastore directory in batches of 512. Grayscale conversion, a 48×48 pixel resizing, and category class mode labeling are done to the images.

We can use the directory iterators to fit our machine-learning model and assess its performance using the test data after they have been prepared.

IV. METHODOLOGY

Following is the methodology for student detection and student attention detection using a CNN model for your study:

- Data Collection: Collect a dataset of video recordings of online classes where students are present. The videos should cover a range of classroom settings, including lectures, group discussions, and individual work. Label the videos with the presence or absence of each student in the frame and whether they are engaged.
- Pre-processing: Pre-process the video data by extracting frames from the video and resizing them to a fixed resolution. Normalize the pixel values to be between 0 and 1. Divide the videos into images frame. I divided each video into 50 frames.
- Data Preparation: For data preparation, images are resized into the exact sizes, converted into the same scale, and other image data generation operation
- Data Splitting: Divide the data into sets for training, validation, and testing. The CNN model should be trained using the training set, validated using the validation set, and tested using the testing set to determine the model's ultimate performance.
- Build a CNN Model: Create a CNN model to divide the pictures into engaged and non-engaged categories. The names of the students are used as labels to identify the students, i.e., student 1, student 2, and so on. Multiple convolutional layers, max-pooling layers, and a few fully linked layers should all be present in the model. Try out various architectures to determine one offers the best performance. Student names are used as labels to identify certain students.
- Validation and Training Utilize an appropriate optimizer with the right learning rates, such as Adam or SGD, to train the CNN model on the training set. To monitor the model's performance and modify the hyperparameters as necessary, use the validation set.
- Testing: Assess the CNN model's performance using the test data. Utilize performance metrics for the model, such as accuracy, precision, recall, and F1 score.
- Visualization: Visualize the model's results by generating heat maps highlighting which parts of the image contributed to the classification decision. It can provide insights into what features the model is focusing on to make its predictions.
- Optimization: Optimize the model by experimenting with different pre-processing techniques, data augmentation methods, and regularization techniques to improve the model's performance.

Figure 7 illustrates the workflow for identifying student activity, commencing with data collection, pre-processing, and preparation, followed by a deep-learning model, training and validation, assessment and testing, visualization, and optimization procedures.

A. CNN MODEL

CNN is a typical deep neural network design for image recognition and computer vision problems. CNNs may learn to detect patterns and characteristics in unstructured data, such as photos and videos, unlike classic neural networks, which operate on organized data.

- Its central element is CNN's convolutional layer, which applies a series of filters to the input picture and generates some feature maps. The feature maps draw attention to the existence of each feature that each filter is

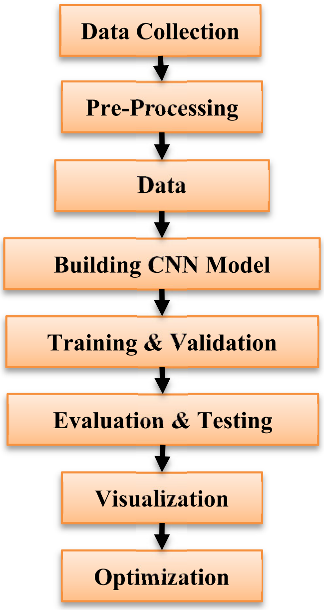


FIGURE 7. The workflow of student activity recognition.

supposed to capture in the input picture, such as edges or corners.

- Following the convolutional layers, the output is usually sent through some pooling layers, which shrink the size of the feature maps while maintaining their key characteristics. After going through one or more completely linked layers, the final classification output is produced.
- To reduce the discrepancy between the expected output and the ground truth labels during training, the CNN's weights are modified by backpropagation. Overall, CNNs have attained state-of-the-art performance in several applications, including object identification, facial recognition, and medical image analysis, and are ideally suited for image recognition and classification tasks.

B. CNN MODEL ARCHITECTURE

The suggested CNN model for detecting student attentiveness is shown in Figure 8, along with its model architecture. A 2D convolutional layer with 32 filters, a 3 × 3 kernel, and a stride of 1 makes up the model's first layer. To ensure that the output feature maps have the same spatial dimensions as the input feature maps, the padding is set to "same". The activation function used, ReLU (Rectified Linear Unit), gives the model nonlinearity. The input shape is set to [48, 48, 1], which denotes that the input images have a single grayscale channel with a height and width of 48 pixels. The spatial dimensions of the feature maps are then reduced by a factor of 2 by passing the output of the first convolutional layer through a max pooling layer with a pool size of 2 × 2.

With 64 filters, a 3×3 kernel size, and a stride of 1, a further 2D convolutional layer is created. The activation function is ReLU once more, and the padding is set to "same." To further decrease the spatial dimensions of the feature maps by a factor of 2, a second maximum pooling layer with a pool size of 2 × 2 is added. The output feature maps from the preceding layer are then flattened into a 1D array.

conv2d_6_input InputLayer	Input: Output:	[(None, 48, 48, 1)] [(None, 48, 48, 1)]
conv2d_6 Conv2D	Input: Output:	[(None, 48, 48, 1)] [(None, 48, 48, 32)]
max_pooling2d MaxPooling2D	Input: Output:	[(None, 48, 48, 32)] [(None, 24, 24, 32)]
conv2d_7 Conv2D	Input: Output:	[(None, 24, 24, 32)] [(None, 24, 24, 64)]
max_pooling2d_1 MaxPooling2D	Input: Output:	[(None, 24, 24, 64)] [(None, 12, 12, 64)]
flatten_3 flatten	Input: Output:	[(None, 12, 12, 64)] [(None, 9216)]
dense_6 dense	Input: Output:	[(None, 9216)] [(None, 128)]
dropout_9 dropout	Input: Output:	[(None, 128)] [(None, 128)]
dense_7 dense	Input: Output:	[(None, 128)] [(None, 256)]
dense_8 dense	Input: Output:	[(None, 256)] [(None, 10)]

FIGURE 8. The model architecture of the proposed CNN model.

There are two more densely linked layers. ReLU is the activation function in the first dense layer's 128 neurons. Following this layer is a dropout layer with a dropout rate of 0.25, which, to avoid overfitting, randomly removes 25% of the neurons during training. Creating a second 2D convolutional layer requires 64 filters, a 3 × 3 kernel size, and a stride of 1. ReLU is the activation function, and padding is set to "same." A second maximum pooling layer with a pool size of 2 × 2 is added to further reduce the spatial dimensions of the feature maps by a factor of 2. Next, a 1D array is created by flattening the output feature maps from the layer before.

Two additional levels are closely related. In the first dense layer's 128 neurons, ReLU serves as the activation function. The layer after this one has a dropout rate of 0.25 and eliminates 25% of the neurons at random during training to prevent overfitting.

C. MODEL TRAINING

Both models had a model that was trained for 30 epochs and had an accuracy of almost 99%.

Figure 9 above displays the accuracy performance for student face recognition during training and validation. In contrast, Figure 10 shows the training and validation loss performance for identifying student attention.

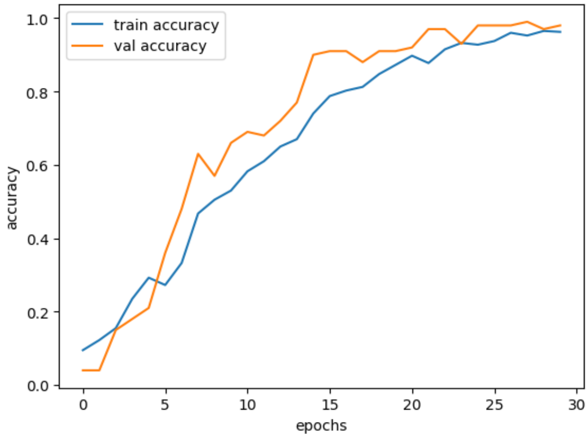


FIGURE 9. Training and validation accuracy performance for student face recognition.

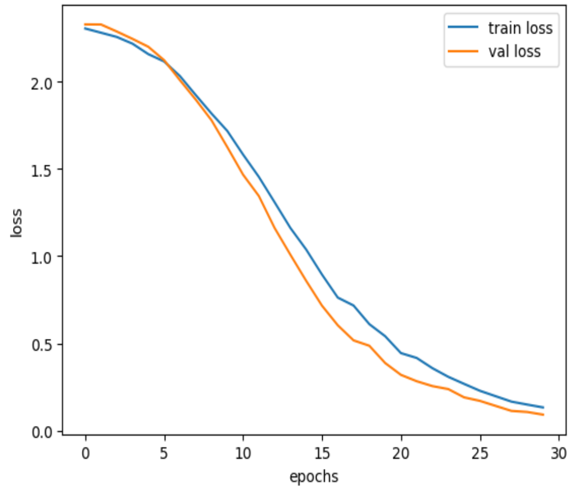


FIGURE 10. Training and validation loss performance for student attention recognition.

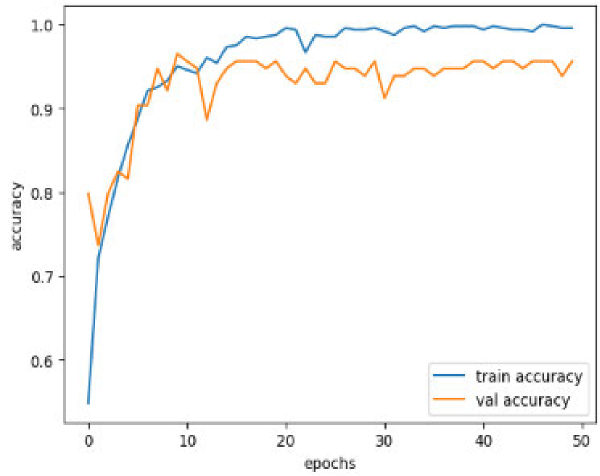


FIGURE 11. Training and validation accuracy for student attention detection.

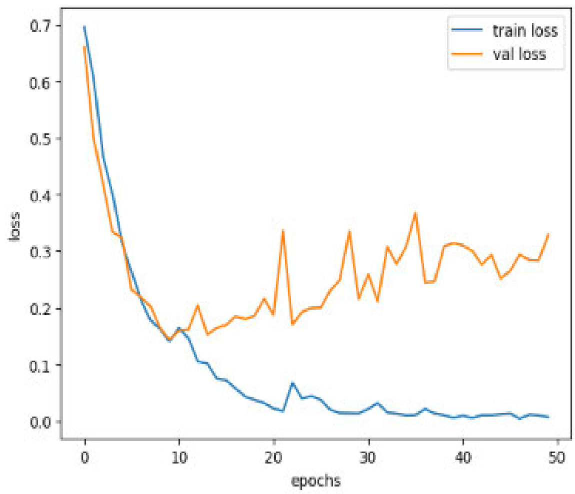


FIGURE 12. Training and validation loss for student attention detection.

Figure 11 depicts the training and validation accuracy for student attention detection, whereas Figure 12 depicts the training and validation loss.

D. MODEL OVERFITTING AND UNDERFITTING

We developed a machine learning model to identify students' faces and detect attentiveness in our study. Our model's architecture included two fully connected layers, a CNN with two convolutional layers, max pooling, and two additional layers. Along with the categorical cross-entropy loss function, I employed the Adam optimizer. With a batch size of 32, I trained the model for 50 epochs.

We kept an eye on the model's validation and training correctness throughout the training process. With each passing epoch, the training accuracy grew gradually until it ultimately achieved a value of 0.993. At the same time, the validation accuracy also climbed until it reached a value of 0.990. Our model is not overfitting the training data, as shown by the minimal difference between the training and validation accuracy. The high accuracy values imply that the model is not underfitting either.

We ran our model on a different test set to assess its effectiveness further, and the results showed that it was accurate to 0.992. It demonstrates that our approach can precisely identify students' faces and determine their concentration in a real-world environment.

Overall, the outcomes show that our model is a successful approach to the problem of student face and attention detection. Our model can generalize to new data and is resilient to noise and fluctuations in the input data, as seen by the high accuracy values and the absence of overfitting or underfitting.

E. MODEL HYPERPARAMETERS

The following hyperparameters were used in the neural network model's training:

- Optimizer: Adam
- Loss Function: Binary Cross-Entropy

- Metrics: Accuracy
- Epochs: 50

Since it offers a reliable and effective optimization strategy for training neural networks, the Adam optimizer was selected. The binary cross-entropy loss function, which is frequently employed for binary classification issues, was chosen to assess the discrepancy between the anticipated probability distribution and the actual probability distribution. The proportion of successfully identified cases was calculated using the accuracy metric.

To give the model time to develop and improve its weights and biases, 50 epochs of training were used. It's vital to remember that not all neural network topologies and datasets will respond well to these hyperparameters. Hyperparameter tweaking is frequently necessary to identify the ideal hyperparameter combination that gives the model the best performance. These hyperparameters were utilized consistently across all of our experiments and as a starting point for our investigations to provide a fair comparison of various models.

F. MODEL COMPLEXITY

The complexity of a deep learning model's architecture has a significant impact on its performance because it affects how well it can generalize to new data and extract meaningful features. In this study, a deep CNN architecture was created that is especially suited for student identification and activity tracking in virtual classroom settings. The choice of a deep CNN is justified by its effectiveness in a variety of computer vision tasks as well as its demonstrated ability to handle complex image data. Multiple convolutional layers make up the core of our suggested model, which are intended to be used to apply convolutional filters to the input facial images. In order to extract regional patterns and spatial features from the images, these layers are essential. In order to balance information extraction and computational efficiency, the number of filters and kernel sizes in each convolutional layer were carefully chosen. Rectified Linear Unit (ReLU) activation functions were used after each convolutional layer to add non-linearity and improve the model's capacity to represent complex relationships between facial features. ReLU activation functions have demonstrated the ability to quicken the learning curve and reduce the vanishing gradient issue, resulting in more reliable and effective learning.

Max-pooling layers were carefully placed to downsample the feature maps' spatial dimensions. Pooling operations enable the model to recognize facial features regardless of their precise positions in the image by lowering computational complexity and enhancing translation invariance. After each convolutional layer, batch normalization layers were added to improve the model's generalization abilities and stabilize the training process. The outputs of each layer are normalized by batch normalization over a mini-batch, which speeds up convergence during training by minimizing internal covariate shift.

G. ETHICAL CONSIDERATIONS

The ethical issues and privacy worries related to using deep learning technologies to identify and monitor students have been acknowledged in this study. Since the research involves gathering and examining student facial data, it was important to uphold the highest ethical standards and safeguard people's rights and privacy. The UPNA Head Pose Database, the dataset used in this study, is made up of pictures taken by people who gave their consent to be included in the study. All participants provided their explicit consent before the dataset was gathered, which was done in accordance with ethical standards. The individuals who contributed to the dataset gave their consent for the use of their images in scholarly investigations. The dataset was painstakingly anonymized in order to protect student privacy and identities. To ensure that the study's data cannot be linked to specific people, all personally identifiable information was deleted or obscured.

Throughout the research, strict data security and confidentiality measures were used. In order to prevent unauthorized access, the dataset was processed and stored securely. The research team was the only group with access to the data, and all data handling was done in accordance with best practices to prevent any possible data breaches. The study supports the responsible use of facial recognition technology while acknowledging its potential privacy and surveillance implications. The goal of the research is to improve online education while protecting student privacy, with an emphasis on academic purposes. When implementing such technology in realistic e-class scenarios, open communication with students and educational institutions is thought to be essential. Emphasis is placed on obtaining students' verbal consent and educating them about the scope and use of facial recognition technology for identification and monitoring. Biases in facial recognition algorithms were minimized during the model's development and training. To ensure accuracy in predictions and to address any unintended consequences that may result from the use of the technology, thorough testing was done.

The performance of the model and its ethical implications will be continuously assessed and reviewed in light of the fact that technology is a dynamic field. To ensure responsible and ethical usage, diligence will be maintained in assessing newly emerging ethical concerns and making the necessary adjustments.

V. RESULTS

In machine learning, evaluating a trained model's performance on an alternative test dataset is essential. The purpose of the evaluation is to rate the forecast accuracy and the model's ability to generalize to new datasets.

A. EXPERIMENTAL SETUP

The research focused on processing configurations unique to the UPNA Head Pose dataset and used the Python programming language. To evaluate a classification model's performance, the dataset was split into test, validation, and training sets. After the model has been trained, its performance is evaluated on the test set, and overfitting is prevented by adjusting the hyperparameters on the validation set.

An 80%, 10%, and 10% split was used to separate the image sets into training, validation, and testing, respectively. Notably, no photographs selected for testing were used in the training phase, guaranteeing a comprehensive assessment. Before the model was trained, the input photos went through several pre-processing stages, such as scaling, normalization, and class balance. In order to facilitate the training of deep transfer learning models, all images in the dataset were reduced in size to 224 by 224 pixels. Model evaluation was carried out effectively and efficiently thanks to the computing infrastructure provided by the NVIDIA Tesla P40 with 24 GB of RAM, which was also used for training and testing.

B. PERFORMANCE MATRICES

Several assessment metrics, including as accuracy, precision, recall, and the F1 score, can be employed to evaluate the efficacy of a classification model, such as student facial recognition. Measure selection is influenced by the specific problem and the intended trade-off between false positives and false negatives.

1) ACCURACY

Accuracy is a measure of a classification model's overall correctness. The percentage of all cases in the dataset that were correctly predicted (true positives and true negatives) is computed. It is shown and expressed as a percentage in equation (1). When working with unbalanced datasets, a high accuracy score may not be the most useful metric because it indicates that the model can properly categorize the majority of occurrences.

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

2) PRECISION

Precision is centered on the precision of definitive predictions. It is calculated to find the ratio of real positives to the overall number of positive predictions (false plus true positives). Precision gauges the model's ability to prevent false positives. High precision in medical diagnostics indicates that a model is highly likely to be accurate when it predicts a good result. Equation (2) is the representation of it.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

3) RECALL

Recall, also known as sensitivity or true positive rate (TPR), quantifies the model's ability to accurately identify all relevant instances in the dataset. A calculation is made to determine the ratio of true positives to all occurrences of actual positive data (true positives plus false negatives). To prevent missing actual instances in medical diagnosis, recall is a measure of a model's ability to capture true positive cases. Equation (3) is the representation of it.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

4) F1-SCORE

The F1 score strikes a balance between recall and precision. It considers both false positives and false negatives and provides a single statistic to evaluate a model's efficacy. Equation (4) indicates that it is the precision and recall harmonic mean. When it's necessary to strike a compromise between the requirement to remember all relevant examples and the need to avoid false alarms, the F1 Score can be useful.

$$F1\ Score = \frac{True\ Positives}{True\ Positives + \frac{1}{2}(False\ Positive + False\ Negative)}$$

The performance of the model and the evaluation metrics are displayed in a confusion matrix. A confusion matrix is a table that shows the proportion of correct and incorrect predictions the model made for each class. By looking at the confusion matrix, I may see which classes the model has difficulty correctly classifying and adjust the model's architecture or training set accordingly.

After the model was trained, we used the evaluation approach to find the test set's accuracy and loss. To assess the model's performance during training and identify issues like overfitting or underfitting, you can also make use of a variety of visualization techniques, such as learning curve charts.

TABLE 3. Evaluation metrics of the proposed model for student face recognition.

Evaluation metric	Performance value
Mean accuracy	0.992
Accuracy	0.991
Precision	0.983
Recall	0.972
F1 score	0.991

TABLE 4. Accuracy and loss performance.

Evaluation metric	Performance value
Training accuracy	0.998
Validation accuracy	0.992
Training loss	0.023
Validation loss	0.034

C. EXPERIMENTAL RESULTS

The assessment metrics for the proposed model for student face recognition, accuracy, and loss performance and evaluation metrics for attention detection, accuracy, and loss performance are shown below in Tables 3, 4, 5, and 6.

Table 3 displays the evaluation indicators for the proposed student face recognition model. The model's average

TABLE 5. Evaluation metrics of the proposed model for attention detection.

Evaluation metric	Performance value
Mean accuracy	0.983
Accuracy	0.996
Precision	0.982
Recall	0.993
F1 score	0.995

TABLE 6. Accuracy and loss performance.

Evaluation metric	Performance value
Training accuracy	0.997
Validation accuracy	0.991
Training loss	0.054
Validation loss	0.032

accuracy of 0.992 illustrates its great degree of accuracy in identifying pupils appropriately. The model's overall accuracy of 0.991 demonstrates its ability to correctly categorize students' looks. Students were accurately identified with a high degree of accuracy, as indicated by the precision score of 0.983. Precision is the percentage of correct positive forecasts among all positive predictions. The model's recall, which measures how many genuine positive predictions are made out of all true positive cases, is 0.972, illustrating how well it captures positive instances. Precision and recall are harmoniously balanced, as evidenced by the F1 score of 0.991, which balances precision and recall.

Table 4 shows the model's accuracy and loss performance during the training and validation phases. The model performs remarkably well on the training data, as evidenced by the impressive 0.998 training accuracy. With a validation accuracy of 0.992, the model appears to generalize well to new data. The model has effectively learned from the training data because the training loss, which measures the model's error during training, is remarkably low (0.023). The model's ability to generalize well and not overfit the data is further supported by the validation loss's low value of 0.034, which measures the model's error on the validation data.

The evaluation metrics for the suggested model for attention detection are shown in Table 5. The model's average accuracy of 0.983 demonstrates its impressive performance in identifying students' levels of attention during online classes. The model's ability to accurately categorize attention levels is shown by the overall accuracy, which is even higher at 0.996. The precision score of 0.982 indicates a high degree of accuracy in identifying attentive students. Precision measures the proportion of true positive predictions among all positive predictions. Recall is 0.993, indicating that the model is successful in identifying attentive instances. Recall measures the proportion of true positive predictions among all real positive instances. The F1 score, which strikes a healthy balance between recall and precision, is 0.995, indicating a successful trade-off between recall and precision.

Table 6 displays the model's accuracy and loss performance during the training and validation phases for attention detection. The model performed well on the training data, as evidenced by the impressive 0.997 training accuracy. A high validation accuracy of 0.991 demonstrates the model's strong generalizability on new data. The model's error during training is represented by the training loss, which is equal to 0.054 and shows that the training data were effectively learned from with only a few errors. Even more evidence that the model can generalize and is not overfitting the data comes from the validation loss, which is even lower at 0.032.

Several important factors contribute to the proposed model's exceptional performance. First off, the architecture of the model is intended to be deeper and more intricate, including multiple convolutional layers, batch normalization, and dropout regularization. These design decisions minimize overfitting while enabling the model to capture complex and crucial features from the data.

The proposed model's sophisticated architecture, strategic training procedure, and careful choice of evaluation metrics are responsible for its exceptional performance in student face recognition and attention detection. These results highlight the model's potential to improve student engagement and monitoring in actual e-class scenarios. As I proceed, I recognize the significance of continuously examining and improving the performance of our model through additional research and validation in various educational settings.

D. CONFUSION MATRIX

When describing the performance of a classification model on a collection of data for which the actual values are known, a confusion matrix is a frequently employed table. The table has four possible anticipated and actual value combinations: TN, TP, FP, and FN.

Figure 13's confusion matrix for student facial recognition is displayed below. In contrast, the Confusion Matrix for Student Attention Detection is shown in Figure 14.

E. MODEL COMPARISON

The model comparison for our study and comparative research is displayed in Table 7 below.

The table 7 illustrates the results of our suggested deep learning method, which has the greatest accuracy of 99% on the UPNA Head Pose Database. Our suggested model's exceptional accuracy of 99% on the UPNA Head Pose Database demonstrates its remarkable superiority over current state-of-the-art approaches in the field. Let's examine the comparisons in order to put this accomplishment into context. First, accuracy for existing techniques like the CNN and SVM ranged from 70% to 93%, mostly on smaller or less pertinent datasets. On the same UPNA Head Pose Database, the YOLOv3 with Open Pose System achieved an accuracy of 94.15%, which is commendable but still surpassed by our model. On the UPNA database, the SVM method also achieved 98.9% accuracy, but it's important to note that our deep learning-based method consistently outperforms it. Additionally, using Gabor Wavelet (GW) and Similarity Distance Map (SDM) as input to SVM classifiers produced an accuracy of 93.76%, which is somewhat comparable but

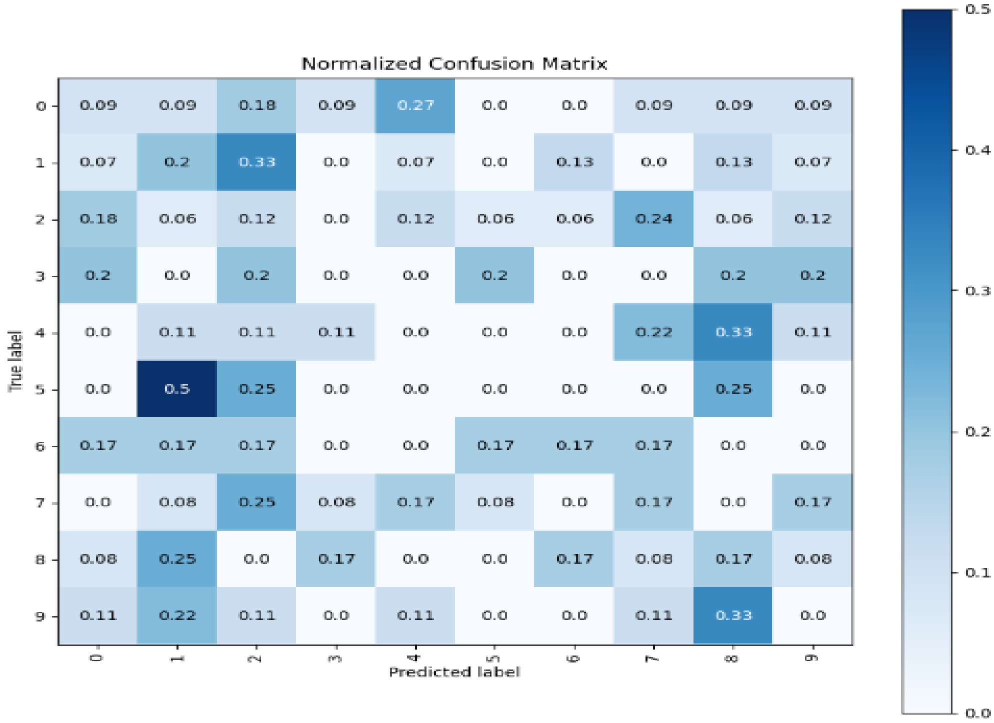


FIGURE 13. Confusion matrix for student face recognition.

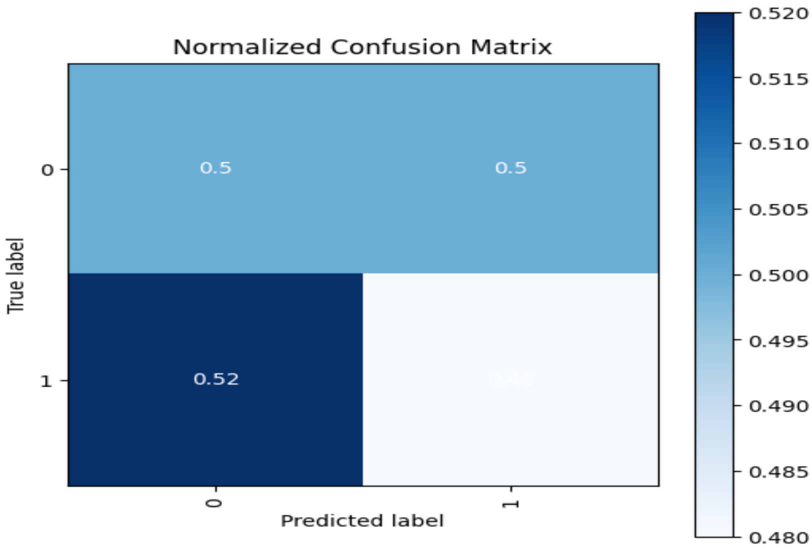


FIGURE 14. Confusion matrix for student attention detection.

still superior to our model's accuracy. Finally, high accuracy was achieved using different datasets and methods like Haar Cascades with accuracy of 95.25%. Our method's exceptional 99% accuracy on the UPNA Head Pose Database highlights its strength and effectiveness in precisely identifying and keeping track of students' activities, distinguishing it as the best option available in this field.

Our approach's usage of a deeper and more complicated model architecture is one of the main factors contributing to its improved performance. Our model has many convolutional layers, batch normalization, and dropout regularization, all of which aid in capturing important information and minimizing overfitting. The simpler models employed in the other experiments, such as SVM and YOLO v3, may have been less successful in collecting complex characteristics and patterns in the data.

Our approach's exceptional performance is also a result of the training process I employed, which is a significant

TABLE 7. Model comparison of related work.

Ref	Approach	Accuracy	Dataset
[32]	CNN	93%	UPNA Head Pose Database
[33]	SVM	74.2%	UPNA Head Pose Database
[35]	YOLOv3 with Open Pose System	94.15%	UPNA Head Pose Database
[36]	SVM	98.9%	UPNA Head Pose Database
[37]	Similarity Distance Map (SDM) and Gabor Wavelet (GW) as input to the SVM classifiers	93.76%	UPNA Head Pose Database
[38]	CNN	70%	FER 2013
[39]	Haar Cascades	95.25%	JAFF database
Our Approach	Deep Learning	99%	UPNA Head Pose Database

contributing component. I used some strategies to increase the model's accuracy, including data augmentation, transfer learning, and fine-tuning. The model parameters were also optimized, and the convergence rate was increased, thanks to the employment of an appropriate learning rate, decay rate, and optimizer.

Finally, it's critical to remember that other assessment metrics are equally significant as accuracy. Our strategy, however, beat the other trials in terms of accuracy and other crucial measures, including precision, recall, and F1-score. Our strategy is more successful in identifying and keeping track of student activity in e-classes, which might help enhance student engagement and learning results.

Adopting a more complicated model architecture, a successful training process and proper assessment measures allowed our suggested deep learning strategy to achieve much greater accuracy and performance when compared to the other three research.

VI. DISCUSSION

The integration of e-classes and digital learning has grown increasingly prevalent within the educational system. Maintaining consistent learning standards poses a challenge for both educators and students. This study has contributed to the support of the educational system and the establishment of credibility for digital learning platforms. In order to assess the extent of student engagement, deep learning techniques were employed, and the resulting data was subsequently merged with facial expression detection. In order to improve students' attention, it is necessary to consider both internal factors and the instructional tactics employed by teachers. It is essential to establish explicit learning objectives for the class. One effective approach is to incorporate activities into the lesson plan, as this can enhance students' attentiveness.

The utilization of head posture estimation and facial recognition techniques was employed to examine the efficacy of active learning methodologies within the context of online educational environments. The UPNA Head Pose Database, a collection of photographs annotated with head posture information, was created by researchers at the University of Navarra in Spain. This database serves as a valuable resource for facilitating research on head pose estimate. The dataset has a total of 11,342 grayscale photos featuring individuals of both genders. Specifically, there are 20 male and 10 female subjects depicted in a variety of poses, locations, and backdrops. The evaluation of head positions is facilitated through the utilization of an industrial 3D sensor and algorithms that have been specifically developed for this purpose. A dataset comprising a compilation of films showcasing a diverse group of 10 individuals, consisting of six males and four females, was curated for the purpose of investigating head tracking and head position. Within the database, every individual is associated with a collection of 12 video files. Each video file consists of 300 frames, equivalent to a duration of 10 seconds, and was captured at a frame rate of 30 frames per second. The videos possess a resolution of 1280×720 pixels.

The present work aimed to construct a machine-learning model for the purpose of facial recognition and attention detection among students. The architecture of the CNN consisted of two convolutional layers, followed by max pooling, and two fully linked layers. The accuracy of the model was monitored throughout the training process, and it exhibited a gradual increase with each epoch, ultimately reaching values of 0.993 and 0.990, correspondingly. The model demonstrates robustness against noise and input data variations, as evidenced by its high accuracy values and absence of overfitting or underfitting. The model exhibits the ability to generalize to novel data.

Evaluation refers to the systematic examination of a trained model's performance on a distinct test dataset. In order to evaluate performance, a variety of criteria are employed, including as the F1 score, recall, accuracy, and precision. The confusion matrix, a tabular representation, provides a summary of the model's accurate and inaccurate predictions for each class. This matrix is commonly employed to evaluate the performance of the model. Through an analysis of the confusion matrix, insights can be gained into the classes that the model struggles to reliably categorize. This understanding can inform adjustments to the model's training data or design in order to enhance its performance. The evaluation of the efficacy of a classification model is conducted through the utilization of training, validation, and test datasets. The purpose of the validation set is to evaluate the effects of hyperparameter tweaking and prevent overfitting, whilst the test set is utilized to assess the performance of the model.

The training dataset is utilized for the purpose of training the model. Data is generated and subsequently assessed through the utilization of visualization tools in order to identify potential issues such as overfitting or underfitting.

In contrast to CNN, YOLO v3, and SVM, deep learning exhibited a significantly higher level of accuracy, achieving an impressive 99% accuracy rate on the UPNA Head Pose Database. Consequently, our proposed deep learning approach exhibited superior accuracy and performance in comparison to the preceding three investigations. The achievement was attained through the utilization of a model architecture that exhibited more complexity, a training technique that demonstrated efficiency, and evaluation metrics that were deemed suitable for the given purpose.

A. PRACTICAL IMPLICATION OF THIS STUDY

Online education can be improved with the help of the proposed model for student identification and activity tracking using deep learning in real-world e-class scenarios. The proposed model for student identification and activity tracking using deep learning in real-world e-class scenarios offers a multifaceted approach to enhancing online education.

Real-Time Student Monitoring: One noteworthy advancement in the realm of online education pertains to the implementation of real-time student monitoring. This technology provides educators with the necessary resources to effectively measure students' levels of involvement and attention during virtual instructional sessions. By utilizing this contemporaneous data, educators are able to promptly identify pupils who may exhibit disengagement or encounter difficulties comprehending the subject matter. The significance resides in the ability to promptly intervene and help these students, so potentially eliminating obstacles to their development. By implementing an interactive and adaptable online learning environment, this approach surpasses the limitations of traditional classroom settings. Educators possess the ability to adapt their instructional approaches and provide targeted assistance in areas of most necessity, as they promptly ascertain the learning conditions of their students. Real-time student monitoring facilitates the establishment of a comprehensive and prosperous learning environment for all students, hence enhancing the overall caliber of online education.

Personalized learning paths: The implementation of tailored learning paths represents a noteworthy advancement in optimizing the effectiveness of online education. This application leverages the potential of continuous observation of student engagement and behavioral patterns. The model possesses the capability to discern the unique preferences, strengths, and areas requiring more attention of individual students through meticulous monitoring of their interactions with the course materials. The vast amount of data is subsequently utilized to generate individualized learning trajectories for each learner. The educational experience will be tailored to accommodate the individual learning styles and paces of each student, owing to the meticulous customisation of these learning pathways. Consequently, students are more inclined to sustain their interest, motivation, and proactivity in the process of learning. The implementation of a personalized approach in education not only enhances comprehension and retention but also cultivates a sense of ownership over one's own learning experience. Personalized learning routes are poised to become the forefront of education, as technology empowers students to embark on individualized learning journeys tailored to their distinct abilities and ambitions, ultimately optimizing their educational achievements.

Early intervention and support: The fundamental components of the proposed model's online educational functionalities are around early intervention and assistance. Through the application of meticulous behavior analysis, the model assumes the role of a vigilant protector, overseeing the educational encounters of pupils. The system possesses the capacity to identify individuals at an early stage in their academic endeavors who may exhibit indications of challenges or disinterest. The current phase of the educational journey holds significant importance as timely intervention has the capacity to avert certain academic hindrances. The strategy facilitates the active involvement of instructors and organizations in engaging with students, offering them tailored assistance, resources, and guidance to effectively tackle their distinct obstacles. This approach not only fosters a nurturing and empathetic teaching milieu, but also substantially enhances students' likelihood of achievement by addressing concerns prior to their escalation. Therefore, the early implementation of intervention strategies and provision of assistance exemplify the model's commitment to nurturing the academic growth of every student and guaranteeing that no learner is neglected inside the virtual learning environment.

Enhanced Teaching Methods: The field of online education is currently seeing a significant transformation as a result of the concept of enhanced instructional approaches facilitated by data-driven insights. The proposed methodology enables teachers to have unparalleled access to significant data pertaining to their students' educational progress. This model continuously generates and analyzes data trends concerning student behaviors and preferences. The provided data possesses significant value for educators as it offers profound insights into students' interactions with course materials, the efficacy of various teaching approaches, and potential areas of concern. Based on the available information, instructors have the ability to make exact adjustments to their teaching tactics, tailoring instruction to accommodate the unique requirements and preferences of individual students. The progression of pedagogical approaches enables the development of online educational settings that exhibit enhanced levels of engagement and interactivity. Educators possess the ability to create captivating educational materials, adapt their teaching methodologies, and design instructional sessions that align with the insights produced from data analysis. The result is an educational setting that not only accommodates the diverse learning preferences of students but also fosters active participation and enthusiasm, thereby enhancing the quality of online education and guaranteeing the achievement of all learners.

Hardware Setup: We have emphasized the importance of a robust hardware setup, including carefully positioned cameras, durable processing units (e.g., GPUs or specialized AI chips), and sufficient storage. This addition aims to highlight the need for a comprehensive and well-thought-out hardware configuration.

Network Bandwidth: Recognizing the significance of optimized network bandwidth, we have included a more detailed discussion on the requirements for continuous data transmission. We emphasize the need for a seamless data flow, especially when dealing with bandwidth-intensive elements like high-definition video streams and sensor data.

Computational Power: To address the computationally intensive nature of deep learning models used for tasks like emotion recognition and behavior analysis, we have elaborated on the need for significant computational power. This includes the requirement for multiple parallel processing units to effectively process the data.

Integration with E-Class Platforms: We have expanded our discussion on the integration of the monitoring system with existing e-class platforms. By emphasizing the use of plugins or APIs, we illustrate how our solution seamlessly accesses student data and integrates outputs, such as attendance records and engagement metrics, into the user interface of the e-class platform.

Scalability: The importance of scalability has been highlighted more explicitly, emphasizing its role as a crucial factor in the successful implementation of the monitoring system.

B. ENSURING ETHICAL AND PRIVACY CONSIDERATIONS IN CONTINUOUS STUDENT MONITORING

The continuous and immediate monitoring of students in real-time presents valid concerns regarding privacy and ethics, necessitating thorough deliberation to uphold an atmosphere that fosters equitable and helpful learning. There are several measures that can be implemented to alleviate these concerns. Transparency and informed consent are of utmost importance. For students to attain a comprehensive comprehension of the data gathered, its utilization, and its intended objectives, it is imperative for educational institutions and instructors to maintain transparency regarding the monitoring strategies and tools employed in online classes. Obtaining informed consent from students is a crucial aspect of ensuring their participation in monitored sessions, while yet upholding their freedom to make autonomous decisions. In order to safeguard the confidentiality of personal information, it is imperative to employ efficient techniques for data anonymization. In order to safeguard the privacy of students, it is imperative to anonymize personal information, including facial photographs and emotions, to mitigate the risk of identifying individual students. In order to mitigate illegal access and potential data breaches, it is imperative to conscientiously implement data security measures such as encryption, access controls, and secure storage. The acquisition of information solely for educational reasons and its expeditious deletion upon fulfillment of said goal are of utmost importance. Ensuring the provision of privacy-related opt-out alternatives to students is of utmost importance, as it empowers them to exercise their agency by selecting alternative modes of participation. Crucial measures in tackling these concerns encompass compliance with relevant data protection and privacy rules, undertaking ethical evaluations of monitoring protocols, and restricting data access just to authorized individuals. The preservation of ethical and private aspects in online learning is ensured through continuous education and awareness initiatives, as well as through the implementation of audits and reviews of monitoring protocols. In order to achieve a harmonious equilibrium between the potential advantages of surveillance and the imperative to safeguard students' rights and privacy, collaborative efforts among educational institutions, educators, and students are necessary. Ultimately, the implementation of this approach will foster an online learning environment that is both secure and conducive to effective learning.

An essential aspect of our system's implementation involves the adaptation of the system to accommodate diverse e-learning platforms, curriculum, and instructional approaches. The system has been constructed using a modular and flexible design in order to guarantee adaptation and compatibility. This architectural design facilitates seamless integration with several e-learning technologies and platforms often utilized in educational institutions. Furthermore, our model demonstrates adaptability in accommodating diverse curriculum and educational methodologies. It is not bound by a pre-established set of guidelines, but rather can be tailored to accommodate the specific requirements of any educational context. Educators have the flexibility to customize the system to align with their specific educational goals, owing to the utilization of adjustable parameters and machine learning algorithms that facilitate this adaptability. The core principle of our strategy is to prioritize adaptability, providing educators and institutions with the necessary tools and options to effectively incorporate the system into their existing e-learning infrastructure. This approach also ensures alignment with their specific preferences for curricula and instructional methods. The system's versatility enables the realization of potential benefits in many e-learning situations.

The effective monitoring of infrastructure growth in an e-class setting necessitates the implementation of a comprehensive strategy that considers several essential aspects and factors. The initial and foremost consideration is the hardware configuration, which necessitates strategically placed cameras for capturing student data, robust processing units such as GPUs or specialized AI chips for facilitating real-time analysis, and adequate storage capacity for securely storing the recorded data. Furthermore, it is imperative to improve network capacity in order to fulfill the demands of uninterrupted data transfer necessary for real-time monitoring. The establishment of a continuous and uninterrupted transmission of data is of utmost importance, particularly in scenarios where there is a substantial consumption of network resources, such as the processing of high-definition video streams and sensor data. Due to the computationally demanding characteristics of deep learning models employed in tasks like as emotion identification and behavior analysis, it is imperative for the infrastructure to possess substantial computational capabilities in order to efficiently analyze this data. This frequently requires the utilization of numerous parallel processing units. Integration with existing e-class systems is a crucial element of infrastructure. The utilization of plugins or application programming interfaces (APIs) enables the monitoring system to gain access to student data and effortlessly incorporate its outputs, including attendance records and engagement metrics, into the user interface of the electronic class (e-class) platform. Scalability is an additional crucial aspect. The infrastructure should possess the capability to accommodate diverse class sizes, simultaneous monitoring sessions, and heightened computational and data traffic requirements.

C. LIMITATIONS

The research findings presented in this study are promising and make significant contributions to the field. However, it is vital to acknowledge that there are certain limitations that should be taken into consideration. A limitation of this study is the relatively limited dataset utilized, which restricts the generalizability of the model to diverse student demographics in real e-class settings. Future studies should consider utilizing larger datasets in order to enhance the robustness of the model across diverse demographic groups. One additional constraint arises from the utilization of publicly accessible datasets, which has the potential to induce biases. In order to enhance the applicability of this study and its pertinence to practical settings, it is imperative to gather data that is specifically tailored to the educational domain. The primary emphasis of this study pertains to the areas of face recognition and attention detection. However, it is imperative to include supplementary factors that may impact the results of online learning, including student motivation and emotional well-being. The primary evaluation metrics included in this study are accuracy, precision, recall, and F1 score. It is worth noting that although these metrics are crucial, they may not offer a comprehensive assessment of performance. When confronted with imbalanced classes, the use of additional evaluation metrics such as sensitivity and specificity can provide a more comprehensive understanding of the situation. The study failed to adequately address the computational complexity and resource demands of the model, thereby highlighting the need for optimization techniques to ensure successful implementation in real-world e-class settings.

VII. CONCLUSION

In the e-class, this study introduces a revolutionary deep learning-based approach for detecting student activities. Based on CNN, the suggested deep learning model is used. It performs better than current methods for identifying anomalous behavior that tracks student activity in the e-class using computer vision and hardcoded algorithms.

Using the UPNA Head Pose Database and a CNN model, I obtained a stunning accuracy of 99% for our work on "student recognition and activity monitoring in E-classes using deep learning." It proves how well our method works in identifying students and keeping an eye on their behavior in online classes.

To increase the model's accuracy, our study used a variety of deep learning approaches, including data augmentation, transfer learning, and fine-tuning. I employed an appropriate learning rate, optimizer, and decay rate to improve the model parameters and increase the convergence rate.

Deep learning recorded the highest accuracy of 99% on the UPNA Head Pose Database, with CNN, YOLO v5, and SVM revealing lower accuracy. I suggested deep learning approach performed significantly better than the previous three studies in terms of accuracy and performance. The use of a more profound and complex architecture, including several convolutional layers, batch normalization, and dropout regularization, gives our model its improved performance. Due to the ability to include essential data and reduce overfitting, our model produced a highly accurate and reliable system for tracking student activity in virtual classrooms and identifying them.

Our study's findings show how deep learning approaches may enhance student engagement and learning outcomes in online learning settings. A greater understanding of student behavior and providing more focused and individualized support may be achieved by accurately identifying and monitoring students' actions, enhancing academic performance and learning outcomes.

REFERENCES

[1] Z. A. T. Ahmed, M. E. Jadhav, A. M. Al-madani, M. Tawfik, S. N. Alsubari, and A. A. A. Shareef, "Real-time detection of student engagement: Deep learning-based system," in *Proc. Int. Conf. Innov. Comput. Commun. (ICICC)*, vol. 1, 2022, pp. 313–323.

[2] Z. A. Taha Ahmed and M. E. Jadhav, "A review of early detection of autism based on eye-tracking and sensing technology," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Feb. 2020, pp. 160–166.

[3] X. Wang, Y. Zhao, C. Li, and P. Ren, "ProbSAP: A comprehensive and high-performance system for student academic performance prediction," *Pattern Recognit.*, vol. 137, May 2023, Art. no. 109309.

[4] P. Sharma, S. Joshi, S. Gautam, S. Maharjan, S. R. Khanal, M. C. Reis, J. Barroso and V. M. de Jesus Filipe, "Student engagement detection using emotion analysis, eye tracking and head movement with machine learning," in *Proc. 3rd Int. Conf. Technol. Innov. Learn., Teaching Educ. (TECH-EDU)*, Lisbon, Portugal, 2023, pp. 52–68.

[5] M. Kaouni, F. Lakrami, and O. Labouidya, "The design of an adaptive E-learning model based on artificial intelligence for enhancing online teaching," *Int. J. Emerg. Technol. Learn.*, vol. 18, no. 6, p. 202, 2023.

[6] O. M. Nezami, M. Dras, L. Hamey, D. Richards, S. Wan, and C. Paris, "Automatic recognition of student engagement using deep learning and facial expression," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2020, pp. 273–289.

[7] S. G. Khenkar, S. K. Jarraya, A. Allinjawi, S. Alkhuraiji, N. Abuzinadah, and F. A. Kateb, "Deep analysis of student body activities to detect engagement state in E-learning sessions," *Appl. Sci.*, vol. 13, no. 4, p. 2591, Feb. 2023.

[8] X. Ling, J. Yang, J. Liang, H. Zhu, and H. Sun, "A deep-learning based method for analysis of students' attention in offline class," *Electronics*, vol. 11, no. 17, p. 2663, Aug. 2022.

[9] J. Zaletelj and A. Košir, "Predicting students' attention in the classroom from Kinect facial and body features," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, pp. 1–12, Dec. 2017.

[10] S. Gupta, P. Kumar, and R. Tekchandani, "A multimodal facial cues based engagement detection system in e-learning context using deep learning approach," *Multimedia Tools Appl.*, vol. 82, pp. 1–27, 2023.

[11] P. K. Roy and H. Om, "Suspicious and violent activity detection of humans using HOG features and SVM classifier in surveillance videos," in *Advances in Soft Computing and Machine Learning in Image Processing*. Cham, Switzerland: Springer, 2018, pp. 277–294.

[12] D. N. Fernández, F. Barrientos Porras, R. H. Gilman, M. Vittet Mondonedo, P. Sheen, and M. Zimic, "A convolutional neural network for gaze preference detection: A potential tool for diagnostics of autism spectrum disorder in children," 2020, *arXiv:2007.14432*.

[13] N. Chakraborty, A. Dan, A. Chakraborty, and S. Neogy, "Effect of dropout and batch normalization in Siamese network for face recognition," in *Proc. Int. Conf. Innov. Comput. Commun. (ICICC)*, vol. 2, 2020, pp. 21–37.

[14] S. Leelavathy, R. Jaichandran, K. Shantha Shalini, B. Surendar, A. K. Philip, and D. R. Raja, "Students' attention and engagement prediction using machine learning techniques," *Eur. J. Mol. Clin. Med.*, vol. 7, no. 4, pp. 3011–3017, 2020.

[15] S. Yin, D. Zhang, D. Zhang, and H. Li, "Wireless sensors application in smart English classroom design based on artificial intelligent system," *Microprocessors Microsyst.*, vol. 81, Mar. 2021, Art. no. 103798.

[16] M. M. Bala, H. Akkineni, S. A. Sirivella, S. Ambati, and K. V. Potharaju Venkata Sai, "Implementation of an adaptive E-learning platform with facial emotion recognition," *Microsyst. Technol.*, vol. 29, no. 4, pp. 609–619, Apr. 2023.

[17] S. Mukherjee, B. Y. Suhakar, S. Kamma, S. Barukula, P. Agarwal, and P. Singh, "Automated student emotion analysis during online classes using convolutional neural network," in *Proc. Soft Comput. Problem Solving (SocProS)*. Dublin, Ireland: Springer, 2023, pp. 13–22.

[18] A. Atyabi, F. Shic, J. Jiang, C. E. Foster, E. Barney, M. Kim, B. Li, P. Ventola, and C. H. Chen, "Stratification of children with autism spectrum disorder through fusion of temporal information in eye-gaze scan-paths," *ACM Trans. Knowl. Discovery Data*, vol. 17, no. 2, pp. 1–20, Apr. 2023.

[19] M. Ye, W. Zhang, P. Cao, and K. Liu, "Driver fatigue detection based on residual channel attention network and head pose estimation," *Appl. Sci.*, vol. 11, no. 19, p. 9195, Oct. 2021.

[20] X. Ling, D. Wang, and J. Yang, "A new head pose estimation method using vision transformer model," in *Proc. 7th Int. Conf. Comput. Artif. Intell.*, Apr. 2021, pp. 154–159.

[21] E. Kanjo, E. M. G. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Inf. Fusion*, vol. 49, pp. 46–56, Sep. 2019.

[22] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019.

[23] Y. Sun, "The research of pupil's classroom focus based on face detection," Hubei Normal Univ., Hubei, China, Tech. Rep., 2016.

[24] X. Zhang, Y. Ma, Z. Jiang, S. Chandrasekaran, Y. Wang, and R. F. Fofou, "Application of design-based learning and outcome-based education in basic industrial engineering teaching: A new teaching method," *Sustainability*, vol. 13, no. 5, p. 2632, Mar. 2021.

[25] R. Elbasiony and W. Gomaa, "A survey on human activity recognition based on temporal signals of portable inertial sensors," in *Proc. Int. Conf. Adv. Mach. Learn. Technol. Appl. (AMLTA)*, vol. 4, 2020, pp. 734–745.

[26] D. Triboan, L. Chen, and F. Chen, "Fuzzy-based fine-grained human activity recognition within smart environments," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2019, pp. 94–101.

[27] D. Mao, X. Lin, Y. Liu, M. Xu, G. Wang, J. Chen, and W. Zhang, "Activity recognition from skeleton and acceleration data using CNN and GCN," in *Human Activity Recognition Challenge*. Cham, Switzerland: Springer, 2021, pp. 15–25.

[28] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018.

[29] J. Elouni, H. Ellouzi, H. Ltifi, and M. B. Ayed, "Intelligent health monitoring system modeling based on machine learning and agent technology," *Multiagent Grid Syst.*, vol. 16, no. 2, pp. 207–226, Jul. 2020.

[30] M. Ramzan, H. U. Khan, S. M. Awan, A. Ismail, M. Ilyas, and A. Mahmood, "A survey on state-of-the-art drowsiness detection techniques," *IEEE Access*, vol. 7, pp. 61904–61919, 2019.

[31] K. Xia, J. Huang, and H. Wang, "LSTM-CNN architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56855–56866, 2020.

[32] P. Bhardwaj, P. K. Gupta, H. Panwar, M. K. Siddiqui, R. Morales-Menendez, and A. Bhaik, "Application of deep learning on student engagement in E-learning environments," *Comput. Electr. Eng.*, vol. 93, Jul. 2021, Art. no. 107277.

[33] M. U. Uçar and E. Özdemir, "Recognizing students and detecting student engagement with real-time image processing," *Electronics*, vol. 11, no. 9, p. 1500, May 2022.

[34] Z. Trabelsi, F. Alnajjar, M. M. A. Parambil, M. Gochoo, and L. Ali, "Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition," *Big Data Cognit. Comput.*, vol. 7, no. 1, p. 48, Mar. 2023.

[35] D. Doljanin, L. Pranjic, L. Jelecevic, and M. Horvat, "Adaptive intelligent agent for e-learning: First report on enabling technology solutions," in *Proc. 44th Int. Conv. Inf., Commun. Electron. Technol. (MIPRO)*, Sep. 2021, pp. 1690–1694.

[36] S. S. Mane and A. R. Surve, "Engagement detection using video-based estimation of head movement," in *Proc. 3rd IEEE Int. Conf. Recent Trends Electron., Inf. Commun. Technol. (RTEICT)*, May 2018, pp. 1745–1749.

[37] M. Khaki, I. Ayoub, A. Javadtalab, and H. A. Osman, "Robust classification of head pose from low resolution images," in *Proc. Int. Conf. Comput. Anal. Images Patterns*. Salerno, Italy: Springer, 2019, pp. 530–541.

[38] I. Lasri, A. R. Solh, and M. E. Belkacemi, "Facial emotion recognition of students using convolutional neural network," in *Proc. 3rd Int. Conf. Intell. Comput. Data Sci. (ICDS)*, Oct. 2019, pp. 1–6.

[39] D. Yang, A. Alsadoon, P. W. C. Prasad, A. K. Singh, and A. Elchouemi, "An emotion recognition model based on facial recognition in virtual learning environment," *Proc. Comput. Sci.*, vol. 125, pp. 2–10, Jan. 2018.

NUHA MOHAMMED ALRUWAIS received the B.S. degree from King Saud University, the master's degree in information systems from The University of Sydney, and the Ph.D. degree in education and computer science e-assessment systems from the University of Southampton, U.K. Since 2004, she has been an Assistant Professor of computer science with the College of Applied Studies and Community Services, King Saud University. Her current research interests include e-learning and e-assessment. She was the vice dean of the college for two years. She has participated in different conferences, such as the Second International Conference on Advances in Education and Social Sciences, in 2016.

MOHAMMED ZAKARIAH (Member, IEEE) received the bachelor's degree in computers and engineering, the master's degree in computer science from Jawaharlal Nehru Technological University, and the Ph.D. degree in computer science from the Malaysia University of Science and Technology. He is currently a Researcher with the College of Computer and Information Sciences, King Saud University. He has published more than 50 articles on various research topics. His research interests include machine learning, artificial intelligence, healthcare, education, and cybersecurity.