INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Design and Development of a System for Monitoring Student Attention and Concentration during Learning Using CNN Model and Face Landmark Detection

Syamsul Arifin [a,*], Aulia Siti Aisjah [a], Azzezza Nurul Fatima [a], Haniah Mahmudah [a,b]

[a] *Engineering Physics Department, Institut Teknologi Sepuluh Nopember, ITS Campus, Raya ITS, Surabaya, Indonesia*
[b] *Electrical Engineering Department, Politeknik Elektronika Negeri Surabaya, PENS Campus, Raya ITS, Surabaya, Indonesia*
*Corresponding author: *syamsul.arifin@its.ac.id*

***Abstract*—Mobile learning media has been wide and provides a tendency for lecturers to identify students' concentration levels in online classes. To bring the class into active learning, efforts are needed from lecturers and educational institutions to return students' concentration to the ongoing learning process. In this paper, a monitoring and alarm system is designed to increase student concentration and combines two elements of statistical analysis to validate CNN models that recognize face emotions in real time while learning. The research was carried out by recording face data using a camera, extracting digital features, and analyzing facial features. The results of the analysis are used as data input for the decision-making system regarding the level of concentration. The concentration level will be used to activate alarms and send them via chat so that students can focus on learning. The system is created by merging facial expression recognition (FER) and decision-making with a convolutional neural network. The system uses a face landmark via camera V2 and a Raspberry Pi 4 performed with the Haar-Cascade classifier, extracting facial features. Face detection via camera is performed using the Haar-Cascade classifier, which extracts facial features. The results of CNN model face detection with landmark features showed good results, with weighted average performance of precision, recall, and F1-score close to 0.99. According to the implementation results, the average number of facial expressions identified in drowsy and neutral states. The device can alert lecturers to how frequently drowsy detects students within a 10-minute interval.**

***Keywords*—Mobile learning; attention; monitoring; CNN model; face landmark.**

Manuscript received 10 Jul. 2024; revised 22 Aug. 2024; accepted 5 Sep. 2024. Date of publication 31 Jan. 2025.

I. INTRODUCTION

Mobile learning applications are increasing in popularity and interest in the modern educational system, and they have become a regular phenomenon. Mobile learning is a new trend and one of the most cutting-edge technologies for improving learning and teaching outcomes. As a result, it is essential to provide students with real-time access to learning materials, activities, and university services via mobile devices [1]. Learning media evolved during the 2000s, with many strategies implemented by lecturers and educational institutions. Online learning resources are used to boost learners' focus levels while studying [2]. M-learning has been recognized as a creative technique for improving students' learning experiences both inside and outside of the classroom, as learners are already accustomed to such gadgets [3]. The implementation of mobile learning resources using several Android apps and ubiquitous learning theories. The design system utilizes the database, and the implementation of several platform methods on the front end of the mobile learning device has improved the quality of mobile learning [4]. Hybrid learning is becoming a culture, with direct or virtual implementation and a focus on interaction in learning. Nuclear engineering education is classified as requiring science with a grasp of knowledge and technology, and security tries to strengthen these competencies through hybrid learning [5].

As ICT technology develops, webcams can be used to enhance synchronous methods and learning. Several methodologies, including importance-performance map analysis, were employed to assess the significance and performance of each element in mobile learning. To measure mobile learning performance, one hybrid analysis approach combines SEM and artificial neural network (ANN)-based deep learning. Furthermore, it uses facial expressions to

assess the performance of mobile learning [6]. The lecturer or teacher can observe whether the student is able to remain concentrated on the activity or redirect attention to others. Facial expressions can be classified into numerous forms, including normal and aberrant, such as sleepy learners [7].

Face recognition during learning is a challenging subject due to the multiple placements of images of faces [8]. The camera could capture the face from the front, side, or at an angle, obscuring some facial characteristics like the eyes or nose [9]. Another aspect is the presence of structural components like beards, moustaches, or glasses in facial images. Structural components vary in shape, color, and size. Other elements that can influence face identification accuracy include light level, occlusion, and facial expressions. Illumination is the change in light distribution caused by the skin's reflectance qualities and internal camera settings, which may produce shadows on certain areas of the face. Examples of smiling, laughing, angry, sad, surprised, and scared facial expressions must be identified [10]. One way to distinguish these diverse expressions is through classification strategies using artificial neural networks (ANN), CNN models, and some combinations of deep CNN models [11],[12].

Online learning creates an atmosphere and learning environment that each learner can customize. When students connect with online learning media via a monitor system in front of them, they will see a little screen. It is clear from this example that light, learning space comfort, and sound all have an impact on the psychological learning environment [13].

Students' expressions and emotions reveal their level of attention. Students' expressions will assist lecturers in understanding how things are general in the online learning environment, encouraging more class involvement and improving the quality of the learning process. It would be more beneficial for lecturers to be aware of students' lack of attention during online class meetings if students and lecturers had access to a monitoring, warning, or alert system. This technique will boost learning effectiveness for the lecturer, and it will make the students a part of the lecturer's consideration throughout class time.

It is vital to consider the consequences of online classes on job stress, burnout, and performance, as well as the moderating role of emotional intelligence in these effects. However, few studies have investigated the association between emotional intelligence and online education [14], [15]. Researchers [14] was developed MobiFace using a gadget designed for universal facial identification. The dataset is a large-scale face with ten million images of 100,000 celebrities. However, this data includes a substantial number of noisy photos and inaccurate ID labels. To receive cleaned training data for every subject in the database. MobiFace was developed by researchers using a gadget designed for universal facial identification. The dataset is a large-scale face with ten million images of 100,000 celebrities. However, this data includes a substantial number of noisy photos and inaccurate ID labels. To receive cleaned training data for every subject in the database. Research [15] on facial expressions indicates happiness, sadness, surprise, and anxiety using local binary patterns (LBP) and the CNN model. None of that research has examined how facial expressions affect learning, especially mobile learning. There has been no research employing facial emotional recognition to detect the influence of mobile learning applications.

Based on studies [1] and [6] on mobile learning, researcher [1] covers statistical analysis of mobile learning models, whereas researcher [6] integrates statistical analysis with ANN. Both of these studies have been implemented in real time. To reduce the gap, this study blends statistical analysis to determine learning with real-time facial expression recognition using the automatic alarm system. The system is built by combining facial expression recognition (FER) and decision-making with a convolutional neural network via camera V2 and a Raspberry Pi 4 using the Haar-Cascade classifier to extract facial features.

This study is a complicated method that combines two elements of statistical analysis to validate CNN models that recognize face emotions in real time while learning. This is a statistical analysis research contribution to the examination of CNN models that recognize face emotions, allowing CNN models to be applied to online learning while also adding statistical analysis to confirm facial emotion recognition.

The topic of study is how to create a monitoring system for student face detection during learning and link it with an automatic alarm system to remind students to return to a focused position in the learning process. This research specifically contributes to:

a. This study presents an important contribution by integrating statistical analysis and CNN models to create a way for reminding students to concentrate their attention during learning sessions.
b. Implementation of a combining facial expression recognition (FER) and decision-making with a CNN model as artificial intelligence system in online and distance learning processes.
c. Applied statistical analysis to validate CNN models for facial recognition.
d. Develop a robot-based system to alert students both online and offline using camera V2 and a Raspberry Pi 4 using the Haar-Cascade classifier to extract facial features.
e. Develop a real-time system for student face recognition and integrate it with technological gadgets.

This paper delivers the discussion of this journal material in the following order: introduction, methods, results, and discussion, followed by conclusions and recommendations for future research.

II. Materials And Method

This section explains the materials and method of monitoring and alert system for mobile learning, which is organized as follows:

A. The Set-Up System

Some preliminary research on the students' involvement and willingness/intention, as well as the preparations required for ITS lecturers to maintain the quality of learning, processes, and results. The usage of mobile phones, laptops, and other media to access learning resources on the ITS-myclassroom platform is being studied and researched with the goal of expanding lecturer competence and improving facilities and infrastructure in online media [15]. The learning process, with lecturers and students performing the primary

roles, as well as widely available learning resources and an atmosphere that supports the learning process, has been attempted to form the intention of students "having the will" to learn independently [6]. This research is being performed by developing a monitoring and alert system that will deliver an "early warning" every 10 minutes during the online learning process. The system design is shown in Fig. 1.

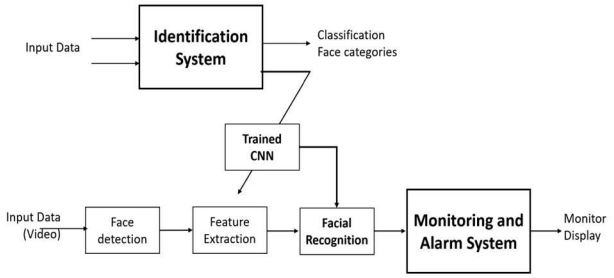

Fig. 1 Schematic: design of a monitoring and alarm system for the learning process

There are two basic systems: the identification system and the monitoring and alert system. The identification method is intended to recognize faces and then determine whether the face is learning "neutral" and "drowsy" condition.

B. Identification System Design

The identification system classifies student conditions during study. The classification of student conditions is based on a facial image identification decision. The face image will select two decisions: "neutral," and " drowsy." The system relies on two types of data: (i) survey data and (ii) face data in the zoom frame.

During the system's training phase, data was collected from individuals who were required to switch on the camera and remain present throughout the learning experience. Online lectures, often known as m-learning, in Indonesia are typically 100 minutes long and worth two credits. Participants are requested to focus on the monitor screen for roughly 100 minutes. During the training, participants' faces were identified and recorded, with the records utilized to activate the monitoring and alarm system. Each face is recorded 64 times [7]. Using n individuals, 64 x n facial data will be collected.

Face detection using the Haar-Cascade classifier generates face images; this method differs from Hasnine's concept [7]. The face detection results are kept in the "face_extract" folder as images that have been cropped to include the face area. The crop causes feature extraction. The investigation focused on the nose, eyes, brows, and mouth area. The areas used are points in the range [18, 68], and the feature area is cut from the image and saved in the "feature_extract" folder [16]. The extract results are compared to 449 drowsy image data and 662 neutral image data collected by previous studies [9]. Facial feature extraction is the process of obtaining certain traits from the face. Facial characteristics are extracted using 68-point landmarks [17].

A CNN-Multi Layer Perceptron (MLP) technique created exclusively for two-dimensional image recognition is used to identify faces [18]. The array's shape is determined by the image's resolution and size. The CNN model architecture has dimensions of (480 × 480) pixels x 3, where 3 represents the RGB magnitude [19]. Data pre-processing and data augmentation help to improve the quality of CNNs for identification. Data preprocessing is the process of cleansing and preparing data for use [20]. Data augmentation is a strategy for increasing or expanding the amount of training datasets. The augmentation technique employs rotation_range, which rotates the image randomly by the specified degree, as well as width_shift_range and height_shift_range, which shift the width and height, respectively [20]. Shear_range provides a random shear transformation, zoom_range applies a random zoom transformation, and horizontal_flip has a value. True seeks to flip the image horizontally, while fill_mode fills the newly formed pixels after rotating or moving [21].

The CNN design utilizes a separable convolution followed by a point convolution in which each input path is filtered independently. This is to reduce processing time, which also reduces costs [22]. There are three sets of convolutions that extract characteristics from the face image and then perform image mapping operations, returning the mapping result as output as shown in Fig. 2. The kernel (filter) is the matrix that performs the convolution process on the incoming data. The 64x64 input layer will be split into numerous convolution layers, including layers 1, 2, and 3, as well as the Fully Connected (FC) layer [23].

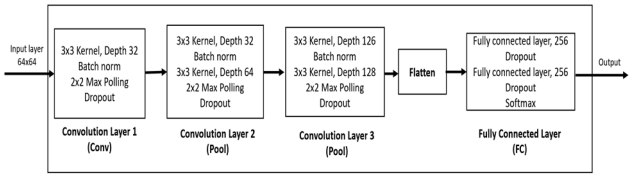

Fig. 2 Block schematic of the CNN structure

The input layer receives a grayscale image of 64x64 pixels ('input_shape = (64,64,1)'). The number 1 denotes a grayscale image, while the number 3 implies an RGB color image. This CNN model consists of three layers: Convolution Layer 1 (CL1), CL2 and CL3, also known as Pooling 1 and Pooling 2, and a Fully Connected layer (FC) with the Softmax activation function. Between CL3 and FC, there is a flattening layer that reduces the multidimensional tensor to a one-dimensional vector. The convolution and pooling layers provide a multidimensional feature map. The flatten layer transforms it into a long vector, which is then connected to the final classification layer, or FC layer [24].

The CNN network uses the ReLU activation function, which adds nonlinearity to the model. This feature aids the NN in learning about difficult data. Within the CNN model, a normalization approach is used between layers. Batch norm accelerates NN training and enables higher learning rates, which facilitates learning. Batch norms also reduce susceptibility to network initial operations and prevent overfitting [25].

Fig. 2 displays the first set of convolutions, which has 32 3 x 3 filters; the second set, which has 64 3 x 3 filters; and the third set, which has 128 3 x 3 filters. The maximum pooling size employed is 2 x 2, which means taking the maximum value of 2 x 2 and halving the feature size. This standard differs from the one proposed by Cakir [18]. The second layer, dropout, is employed to avoid overfitting [26]. The flatten layer converts the convolution layer's matrix output into a vector, allowing the data to be entered into the dense layer. The dense layer is the

FC that connects each neuron to the previous layer. The activation function employed is SoftMax, which converts output values into probabilities for each class, allowing the model to deliver the most accurate class prediction.

The amount of data required for face detection, facial feature extraction, and labeling operations is divided into three parts: training, validation, and test data, with a size ratio of 70:15:15 [23]. The model is trained on some example data. The overall performance of CNN is expressed as accuracy expressed as equation 1, sensitivity in the form of F1-score expressed as equation 2, accuracy represented as equation 3, macro average expressed as equation 4, and weight average expressed as equation 5, in detail expressed in the form.

1) Precision [25]: Precision measures the model's positive prediction accuracy. It is the ratio of the right positive predictions to the total number of positive forecasts. The higher the precision, the fewer false positives appear in the model findings. The accuracy equation is represented as the following equation 1 [27]:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \quad (1)$$

2) Recall: often referred to as sensitivity [25], is the same parameter as precision. Recall measures the model's ability to find all true positive examples. A high recall indicates that the model can capture most of the true positive examples.

3) F1-score [28]. The F1-score is the harmonic mean of precision and recall. It provides a balanced measure of precision and recall, which is useful when there is a class imbalance. A high F1-score indicates a model that is good at balancing precision and recall.

$$\text{F1} - \text{score} = 2\text{x}\ \frac{\text{PrecisionxRecall}}{\text{Precission+Recall}} \quad (2)$$

4) Support [28]: is the actual number of occurrences of each class in the test data. Support correlates with the context of how often each class occurs.

5) Accuracy [28]: which represents the proportion of correct predictions to total predictions, is calculated using the following equation. A high accuracy value suggests less inaccuracy.

$$\text{Accuracy} = \frac{\text{True Positive+True Negatives}}{\text{Total samples}} \quad (3)$$

6) Macro average (Macro Avg): it is the average precision, recall, and F1-score determined with equal weight for each class. The macro average is used to understand the overall performance of the model, regardless of class distribution.

$$\text{Macro avg} = \frac{\text{Precs}_1+\text{Precs}_2+\cdots+\text{Precs}_n}{\text{n}} \quad (4)$$

7) Weighted average (Weight Avg): it is the average of precision, recall, and F1-score derived by assigning weight according to the number of instances in each class using the following equation: It is used to examine class imbalances based on the frequency of each class.

$$\text{Weight avg} = \frac{\sum(\text{Support}_\text{i}\ x\ \text{Metric}_i)}{\sum \text{Support}_i} \quad (5)$$

C. Survey Data

The second input is survey results for statistical analysis. Survey data is utilized to validate the results of facial recognition. The survey was gathered by administering a questionnaire to all student participants following the lecture. The survey contains three categories of questions. Group 1 is the level of mastery or understanding of the topic throughout the lecture process; Group II is a question about "activity during the learning process," and Group III is the level of "engagement." The three data groupings indicate whether a person is attentive, indifferent, or "sleepy" [29].

The survey results were tested in a class of engineering students at ITS Surabaya. The data was further evaluated based on the students' responses after they completed the online learning. There were 29 students attending the training. 28 out of 29 pupils responded to the questions. For question group 1, 20 out of 28 students answered, "paying attention" for 100 minutes (2 credits), with the remaining students responding "not paying attention." The survey data consisted of questions about the course material, as indicated in Table 1. For each question, students selected one of the most relevant responses.

TABLE I
THE QUESTION FOR RESPONDENT IN HYBRID LEARNING

Group of Question	Q	Description
Knowledge	1-9	9 (nine) Questions are linked to learning material.
Engagement (P1)	10	Q1. How interesting was the lesson that was just implemented?
	11	Q2. How interested are you in the learning that has just been carried out?
	12	Q3. How interesting was it for you to take the lesson?
	13	Q4. How bored are you while studying?
	14*	Q5. How much do you want to do something else?
Challenge (P2)	15	Q1. How challenging is the newly acquired knowledge?
	16	Q2. I feel challenged during learning.
	17*	Q3. How difficult is it for you to concentrate?

*Mark *: questions are inverted or reversed in statistical analysis.*

D. Monitoring and Alarm System Design

The design of the monitoring and alarm system is shown in Fig. 3. CNN models that have been trained with various facial features are used to identify face recording results during video streams in online lectures [24]. The electronic devices used have customized specifications and inexpensive prices. This is to procure overall equipment for each computer installed in the classroom, when hybrid learning is conducted.

The Raspberry Pi specifications describe it as a tiny computer with a camera module and an external monitor. The Raspberry Pi 4 runs the 64-bit Raspbian OS Bullseye and uses Python as its programming language. The speaker module is used to generate sound as part of the alarm system. Python and OpenCV software are utilized to activate speakers. OpenCV is an open-source computer vision and machine learning software library [30]. This library contains almost 2500 optimized algorithms. This technique may be used to identify objects, classify human movement in video, detect and distinguish faces, and track object motion.

Facial expression detection is a technology built into the Raspberry Pi 4 that detects data by using a model developed on test data called Face Expression Recognition (FER). Face

detection with the Haar-Cascade classifier, which can be obtained by shrinking the face to the size of this FER, will begin with face detection using 640 x 640 and cut on the face area that the data will pass before proceeding to the next procedure, namely facial feature extraction. The Raspberry Pi is set up as a little computer with a camera module and an external monitor to display videos and images captured during the initial setup. Raspberry Pi 4 runs the 64-bit Raspbian operating system Bullseye and the Python programming language, which is used in this research.

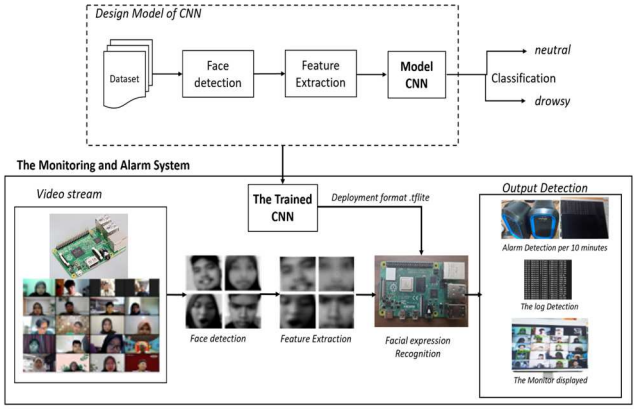

Fig. 3 Schematic of an alarm system for facial expression recognition

The alarm system is designed to generate a sound signal by activating sensors connected to the speaker. The alarm works by activating the speaker every 10 minutes. The 10-minute time frame is used to refocus the learner and engage them in learning activities [31]. The speaker module produces sound to activate the alarm system. The system's software includes the Python programming language and Open-Source Computer Vision (OpenCV). OpenCV is an open-source software library for computer vision and machine learning. This library contains almost 2500 optimized algorithms. These methods may be used to identify objects, classify human movement in video, detect and recognize faces, and track object motion.

E. Implementation of Monitoring and Alarm Systems

The developed monitoring and alert system is used in online classes taught by students using m-learning in numerous engineering study programs at ITS.

Fig. 4 Configuration of monitoring and alarm system tools.

The Raspberry Pi Camera V2 is put in front of a monitor screen that shows an online class, which is one of the m-learning activities. The system will immediately carry out its functions as intended, as shown in the monitoring system overview flowchart in Fig. 3. The initial test of the designed system is performed by assembling the system depicted in Fig. 4. The system was tested multiple times to guarantee that the software and hardware the alarm system were properly.

III. Results and Discussion

This chapter explains several sections, including the results of the sample data analysis, facial expression detection, model validation, and alarm system hardware. Detailed explanation is as follows:

A. Statistical Analysis.

By presenting the items in Table 1 to the test group, the Cronbach's alpha value as a metric of data reliability from the questionnaire was 0.809, with a 10% level of significance. This score is more than 0.7, indicating that the data obtained is reasonably reliable [32]. Fig. 5 shows the video display of the test students. The 32 students who unlocked the webcam revealed 31 faces. Shortly after the trial period for participants, survey data was obtained by presenting a form in Gdrive with 17 brief questions, which were divided into three sections: P1 and P2, as shown in Table 1. The responses will be used to qualitatively validate the CNN results.

Fig. 5 The video stream of students on the zoom platform

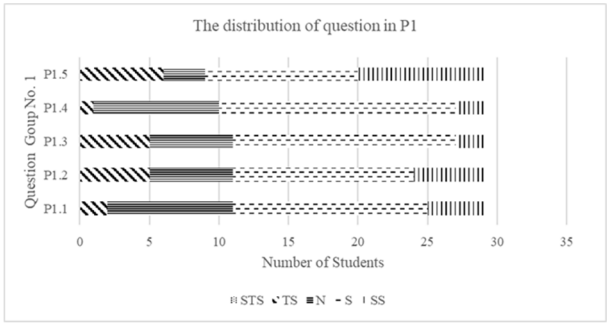

Fig. 6 The number of students who answered Question Group No. 1 (P1) (Engagement)

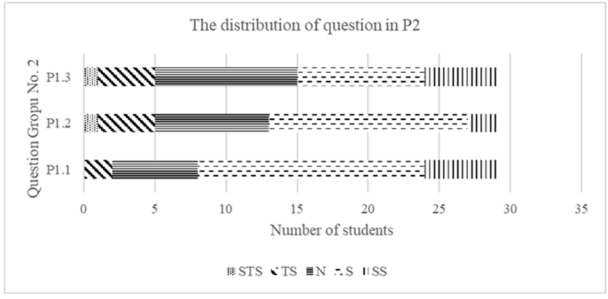

Fig. 7 The number of students who answered Question Group No. 2 (P2) (Challenge)

For Fig. 6 and 7, STS stands for strongly disagree, TS for disagree, N for neutral, S for agree, and SS for strongly

agreeing. Table 2 shows that 243 out of 310 students (7 classes) received a total score of ≥ 7. Nineteen students received a total score of <7 and answered less than seven questions correctly, indicating a lack of attention to the topic.

TABLE II
RESPONDENTS' ANSWERS TO QUESTION GROUPS P1 AND P2

Question	Number of Answers on a scale of 4	Number of Answers on a scale of 2	Percentage of respondents that chose Answer 4
Engagement Category			
1	193	117	62.26
2	138	172	44.52
3	168	142	54.19
4	167	143	53.87
5	174	136	56.13
7	177	133	57.10
Challenge Category			
1	181	129	58.39
2	149	161	48.06
3	199	111	64.19

The response to question P1.4 is the most supportive of the sleepy face category. The relevance of P1.4 and question P1.5 reinforces each other, namely students' motivation to participate in other activities [29].

B. Facial Expression Detection

Face detection was accomplished by FER utilizing the Haar-Cascade classifier, which is used in this work as "haarcascade_frontalface_default.hml". Every image is scaled to 640 by 640 and cropped to match the detected face region. This strategy differs from that used by Hangaragi [8]. Figure 7 depicts the results of a study with 310 students, which revealed that 121 were drowsy and 189 were neutral.

Fig. 8 An example of a face detection result in the detection process

Fig. 8 shows the feature extraction findings, which are used as input by the CNN model. Drowsy and neutral features have distinct characteristics. Drowsy expressions include closed eyes and a closed or open mouth. A neutral expression is one in which the eyes are open and the mouth is closed. Based on the findings, it is clear that there are differences in the position of points on facial landmarks in neutral and drowsy eyes and mouths, particularly in drowsy eye landmarks, where detected points tend to be tighter than in neutral eyes with open eyes; this also applies to differences in open and closed mouths.

C. CNN Model

The entire amount of dataset from face detection and feature extraction is the same: 4923 images. The retrieved features are then classified using the CNN model, and the classification results are returned. This detection result is likewise recorded separately, with drowsy results saved in the "drowsy" folder and neutral results saved in the "neutral" folder. Drowsy detected 1611 faces, while neutral detected 3312.

Extraction of facial features results in a prediction procedure shown in Fig. 9, and the accuracy value is computed using a validation model. Validation of the model using the training model yields training, validation, and test accuracy of 99.7%, 94.9%, and 99.0%, respectively. When comparing the accuracy and loss in training, validation, and testing, as shown in Fig. 10, it is clear that the model performs well with high accuracy values, particularly above 80%, and without overfitting. If we focus on the accuracy and loss of the training data, we can see that the model's performance on the training data continues to improve as the loss decreases, and the model's accuracy increases.

Fig. 9 An example of facial feature extraction results in the prediction process

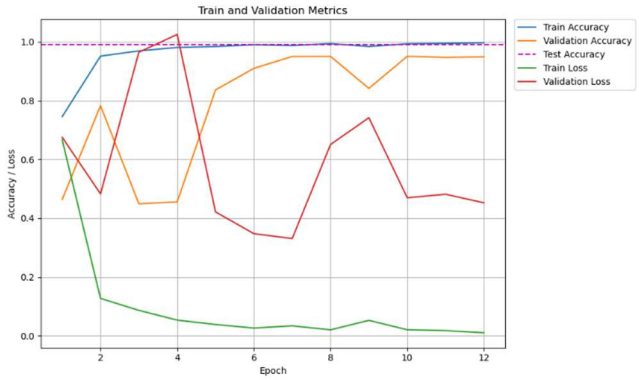

Fig. 10 Accuracy values during training, validation, and testing, as well as loss during validation

TABLE III
THE CLASSIFICATION REPORT FROM THE CNN MODEL

	Precision	*Recall*	*F1-score*	Support
Drowsy	1	0.98	0.99	45
Neutral	0.98	1	0.99	56
Accuracy			0.99	101
Macro avg	0.99	0.99	0.99	101
Weighted avg	0.99	0.99	0.99	101

The confusion matrix (CM) evaluates the performance of the classification CNN model how the model predictions relate to the real labels, and we can calculate numerous assessment metrics such as accuracy, precision, recall, and F1-score. The prediction model is carried out on test data, and a confusion matrix is obtained. Based on the prediction results, it can be seen that the model can predict the data well. Then we also obtain a classification report or summary of the model's performance in classifying test data, namely precision, recall, F1-score, support, accuracy, macro AVG, and weighted AVG, according to Equations 1–5, with the values shown in Table 3. The accuracy results of the CNN

model show good system performance with an accuracy value close to one and low losses approaching a value of zero, as illustrated in Fig. 11.

The Drowsy label achieves 100% precision, 98% recall, and 99% F1-scores, respectively. Neutral labels score 98%, 100%, and 99%, respectively. The macro average of precision, recall, and F1-score is near 0.99, indicating that overall performance across all classes and labels is high or good. The weighted average of precision, recall, and F1-score is near 0.99, indicating that overall performance for all classes and labels is high or good, even when considering class data imbalance. Thus, the test results indicate that the model can accurately classify "drowsy" and "neutral," as shown in Fig. 12.

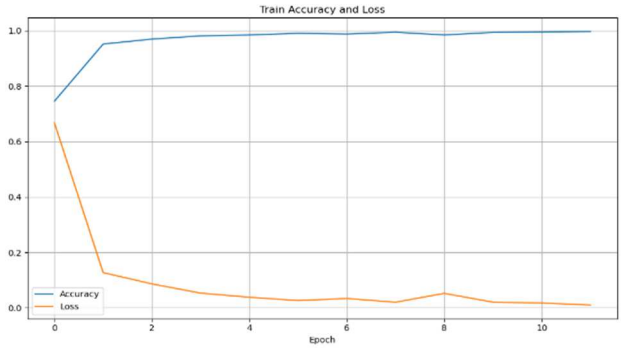

Fig. 11 The results of the confusion matrix (CM)

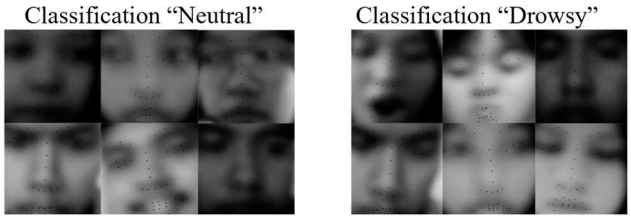

Fig. 12 Results of the classification of neutral and drowsy expressions

D. Alarm System Hardware

The detection of 310 students was divided into eight (eight) monitor screens, each of which photographed a maximum of 40 students.

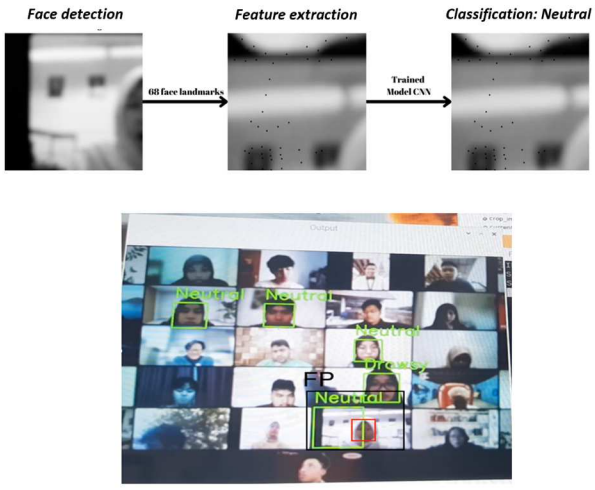

Fig. 13 Prediction results using false-positive (FP) predictions

The detection technique is limited to faces identified in the "drowsy" and "neutral" conditions. On multiple monitor panels, the detection focuses on several students taken on video and indicates that they are asleep or indifferent as shown in Fig. 13.

The facial extraction technique is performed at a number of landmark sites [18],[28] including the brows, eyes, nose, and mouths. The number of data successfully detected and retrieved for facial features was drowsy and neutral. Fig. 14 and 15 indicate the results of face detection and feature extraction, as well as a visualization of the process.

Fig. 14 Face extraction procedure using "neutral" data

Fig. 15 Face extraction procedure using "drowsy" data

It does not exclude the chance of detecting errors throughout the detection process. In Fig. 13, it can be seen that some students have their eyes open yet are labeled as drowsy; this is known as false-positive (FP) since the projected class or classification is incorrect. During the detection process, there are several FP detection results where the frame contains a face, but the prediction box detects an area that is not a face as the detected face, so the detection continues until the model is able to classify. This is because the Intersection Over Union (IoU) value is < threshold, which is the ratio of the intersection between the predicted and ground truth bounding boxes, as shown in Fig.13.

This monitoring and alarm system also includes voice alarms generated by text-to-speech with the Pyttsx3 module. The log contains the alert results obtained during the detection process for validation data. The alarm will sound every 10 minutes to notify you of the total number of pupils in online classes detected by drowsy. The alarm merely tells the lecturer about drowsy's status; it is designed to notify the lecturer how frequently or when students are recognized as being drowsy or not paying attention in class within a 10-minute period. Drowsy conditions indicate that the majority of students are having difficulty focusing their attention. This is corroborated by the signal shown on the left side of Fig. 13.

Based on the recorded log findings, it is clear that the system may deliver changing detection results every 5 seconds, with each detection having a varying FPS. The system requires 5 seconds to process detection between frames. If the log data is distributed in plot form, a plot graph will be generated, as shown in Fig. 16. The highest number of detection results is 21 students in the same frame. However, when compared to the 21 students captured by the camera during the detection process, 7 students' faces were not detected, so the students' expressions could not be classified or were labeled False-Negative (FN).

Table 4 shows the amount of facial expression detection findings in one frame on a single monitor screen, ranging from 1 to 8. On the first monitor, 11 facial expressions (5 drowsy and 6 neutral) were identified from the 20 students

caught on camera, or around 55% of faces had successful facial expressions detected. The system can also provide warnings to lecturers about how often students in the class are detected as being drowsy or inattentive within a 10-minute interval. From the results of the questionnaire data in the Knowledge section, the majority of students were paying attention to the material. This was also obtained from the detection results, which showed that there were indeed several students detected by Drowsy, which showed that the students were not paying attention, but the detection results showed more data in neutral expression.

TABLE IV
THE CLASSIFICATION REPORT FROM THE CNN MODEL

Monitor screen	Number of faces detected			Percentage of faces captured on camera (%)
	Screen	Drowsy	Neutral	
1	20	5	6	55
2	24	7	11	75
3	31	3	14	55
4	30	7	12	63
5	28	8	13	75
6	22	4	8	55
7	23	5	11	70
8	29	4	12	55

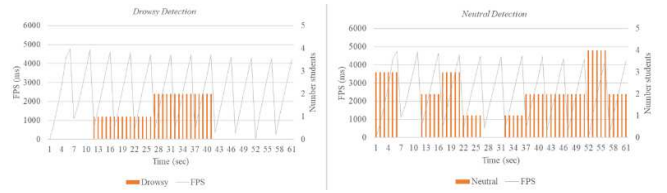

Fig. 16 The number of frames per second (FPS) for drowsy and neutral detection within the first minute

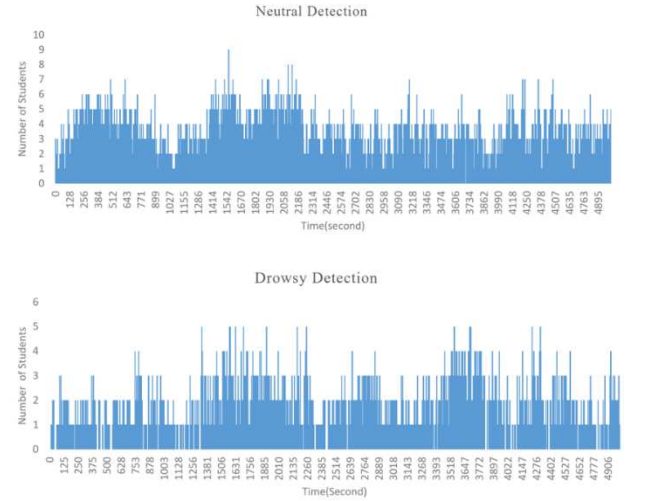

Fig. 17 Changes in the number of detections throughout time

Fig.16 and 17 exhibit graphs that correlate with detection findings and activate the alarm log. According to Fig. 16 the results of facial expression recognition using the CNN model range from 1000 to 4900 frames per second. According to the findings of this study using a CNN model to determine student FER in mobile learning, the CNN model's training results revealed an accuracy of 99.7% in training, 94.9% in validation, and 99.0% in testing. The CNN model's prediction outputs can identify 44 drowsy expressions from 45 data points (97.8%) and 56 neutral expressions from 56 data points (100%). The classification report shows that the Drowsy label has 100% precision, 98% recall, and a 99% F1-score, whereas the Neutral label has 98%, 100%, and 99% respectively. The macro Avg value for precision, recall, and F1-score is close to 0.99, suggesting high to good performance across all classes and labels. The test results indicate that the "drowsy" and "neutral" classifications are good, with a weighted average performance of precision, recall, and F1-score close to 0.99, indicating that overall performance for all classes/labels is high or good when data imbalances are taken into account. The CNN model for recognizing facial expressions utilizing face landmarks for mobile learning combined with Haar detection produced good results with a weighted average performance of precision, recall, and F1-score close to 0.99. When compared to previous studies, which found that facial expressions with CNN had a 97.32% accuracy, the results of this study are nearly identical at 99.72%. The future study could include data preparation to improve performance with augmentation and reduce inference time by employing a lighter CNN.

The test results demonstrate that the system can monitor well and provide alarms every 10 minutes based on drowsy and neutral detection, with a total of 1611 and 3312 faces detected as drowsy and neutral, respectively, and an alarm log obtained every 10 minutes showing 130, 150, 245, 233, 205, 217, 240, and 191 students detected as drowsy. The findings of the questionnaire in the Knowledge portion confirm the validity of the "neutral" facial expression, as they reveal that most students are attentive to the information, as well as some students who are detected as drowsy, indicating that students are not attentive. This finding greatly supports research [31], which focuses on automatic attendance tracking every 10 seconds by identifying the faces of drowning students in mobile learning applications.

IV. CONCLUSION

The research conclusions are that the mobile learning platform, which includes a monitoring and alert system, is intended to improve student-focused attention. Statistical analysis results serve as input data for the concentration-level decision-making system. The concentration level will be utilized to activate the alert and communicate it via chat, allowing students to focus on their studies. The system was created by merging facial expression recognition (FER) and decision making with a convolutional neural network (CNN). The designed solution is then tested on the Raspberry Pi 4 and Pi Camera V2. Face detection via camera is accomplished using the Haar-Cascade classifier, which extracts face features. The results of CNN model face detection utilizing landmark features indicate good system performance values near 0.99. The study mobile learning system can deliver alarm warnings in the form of monitoring warnings to lecturers about how frequently students are detected as tired during a 10-minute interval, which works effectively. Future research can include data pre-processing to improve augmentation performance, employing a lightweight CNN model that can be easily applied to mobile edge devices.

ACKNOWLEDGMENT

We thank the Department of Engineering Physics, which has provided facilities in obtaining data, to Directorate of

Research and Community Services, Institut Teknologi Sepuluh Nopember (DRPM-ITS) for providing research funding in the Indonesian Collaborative Research scheme in 2022.

REFERENCES

[1] M. A. Almaiah, M. M. Alamri, and W. Al-Rahmi, "Applying the UTAUT Model to Explain the Students' Acceptance of Mobile Learning System in Higher Education," *IEEE Access*, vol. 7, pp. 174673–174686, 2019, doi: 10.1109/ACCESS.2019.2957206.

[2] S. Criollo-C, A. Guerrero-Arias, Á. Jaramillo-Alcázar, and S. Luján-Mora, "Mobile learning technologies for education: Benefits and pending issues," *Applied Sciences*, vol. 11, no. 9, 2021, doi:10.3390/app11094111.

[3] M. Uther, "Mobile learning—trends and practices," *Education Sciences*, vol. 9, no. 1, pp. 10–12, 2019, doi: 10.3390/educsci9010033.

[4] S. Qun, "The Development of Mobile Education Resource Database under the Concept of Ubiquitous Learning," *Proceding - 2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA) 2021*, pp. 725–728, 2021, doi:10.1109/icmtma52658.2021.00167.

[5] C. Demazière *et al.*, "Enhancing higher education through hybrid and flipped learning: Experiences from the GRE@T-PIONEeR project," *Nuclear Engineering and Design*, vol. 421, no. February, 2024, doi:10.1016/j.nucengdes.2024.113028.

[6] K. Alhumaid, M. Habes, and S. A. Salloum, "Examining the Factors Influencing the Mobile Learning Usage during COVID-19 Pandemic: An Integrated SEM-ANN Method," *IEEE Access*, vol. 9, pp. 102567–102578, 2021, doi: 10.1109/access.2021.3097753.

[7] M. N. Hasnine, H. T. T. Bui, T. T. T. Tran, H. T. Nguyen, G. Akçapõnar, and H. Ueda, "Students' emotion extraction and visualization for engagement detection in online learning," *Procedia Computer Science*, vol. 192, pp. 3423–3431, 2021, doi:10.1016/j.procs.2021.09.115.

[8] S. Hangaragi, T. Singh, and N. Neelima, "Face Detection and Recognition Using Face Mesh and Deep Neural Network," *Procedia Computer Science*, vol. 218, pp. 741–749, 2022, doi:10.1016/j.procs.2023.01.054.

[9] G. Kaur *et al.*, "Face mask recognition system using CNN model," *Neuroscience Informatics*, vol. 2, no. 3, p. 100035, 2022, doi:10.1016/j.neuri.2021.100035.

[10] D. Bhagat, A. Vakil, R. K. Gupta, and A. Kumar, "Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN)," *Procedia Computer Science*, vol. 235, no. 2023, pp. 2079–2089, 2024, doi: 10.1016/j.procs.2024.04.197.

[11] D. Wang, H. Yu, D. Wang, and G. Li, "Face recognition system based on CNN," *Proceedings - 2020 International Conference on Computer Information and Big Data Applications CIBDA 2020*, pp. 470–473, 2020, doi: 10.1109/cibda50819.2020.00111.

[12] B. R. Ilyas, B. Mohammed, M. Khaled, and K. Miloud, "Enhanced Face Recognition System Based on Deep CNN," *Proceedings - 2019 6th International Conference on Image and Signal Processing and their Applications ISPA*, 2019, doi:10.1109/ispa48434.2019.8966797.

[13] D. Ciraolo, M. Fazio, R. S. Calabrò, M. Villari, and A. Celesti, "Facial expression recognition based on emotional artificial intelligence for tele-rehabilitation," *Biomedical Signal Processing and Control*, vol. 92, no. May 2023, p. 106096, 2024, doi: 10.1016/j.bspc.2024.106096.

[14] C. N. Duong, K. G. Quach, I. Jalata, N. Le, and K. Luu, "MobiFace: A Lightweight Deep Learning Face Recognition on Mobile Devices," *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems BTAS*, 2019, doi:10.1109/BTAS46853.2019.9185981.

[15] R. Ravi, S. V. Yadhukrishna, and R. Prithviraj, "A Face Expression Recognition Using CNN LBP," *Proceedings 4th International Conference on Computing Methodologies and Communication ICCMC 2020*, pp. 684–689, 2020, doi:10.1109/iccmc48092.2020.iccmc-000127.

[16] J. Wang, R. Cao, P. N. Chakravarthula, X. Li, and S. Wang, "A critical period for developing face recognition," *Patterns*, vol. 5, no. 2, p. 100895, 2024, doi: 10.1016/j.patter.2023.100895.

[17] Y. Wang, M. Cao, Z. Fan, and S. Peng, "Learning to Detect 3D Facial Landmarks via Heatmap Regression with Graph Convolutional Network," *Proceedings 36th AAAI Conference on Artificial Intelligence, 2022*, vol. 36, pp. 2595–2603, 2022, doi:10.1609/aaai.v36i3.20161.

[18] D. Cakir and N. Arica, "Cascading CNNs for facial action unit detection," *Engineering Science and Technology, an International Journal*, vol. 47, no. October, p. 101553, 2023, doi:10.1016/j.jestch.2023.101553.

[19] S. Dwijayanti, R. R. Abdillah, H. Hikmarika, Hermawati, Z. Husin, and B. Y. Suprapto, "Facial Expression Recognition and Face Recognition Using a Convolutional Neural Network," *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems ISRITI 2020*, pp. 621–626, 2020, doi:10.1109/isriti51436.2020.9315513.

[20] C. Ashwini and V. Sellam, "An optimal model for identification and classification of corn leaf disease using hybrid 3D-CNN and LSTM," *Biomedical Signal Processing and Control*, vol. 92, no. February, p. 106089, 2024, doi: 10.1016/j.bspc.2024.106089.

[21] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022, doi:10.1016/j.gltp.2022.04.020.

[22] N. Deshpande, F. Nunnari, and E. Avramidis, "Fine-tuning of Convolutional Neural Networks for the Recognition of Facial Expressions in Sign Language Video Samples," *7th Workshop on Sign Language Translation and Avatar Technology SLTAT 2022*, June, pp. 29–38, 2022.

[23] A. Nasayreh *et al.*, "Jordanian banknote data recognition: A CNN-based approach with attention mechanism," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 4, p. 102038, 2024, doi: 10.1016/j.jksuci.2024.102038.

[24] D. Yang *et al.*, "An efficient multi-task learning CNN for driver attention monitoring," *Journal of Systems Architecture*, vol. 148, no. September 2023, p. 103085, 2024, doi: 10.1016/j.sysarc.2024.103085.

[25] G. Rajeshkumar *et al.*, "Smart office automation via faster R-CNN based face recognition and internet of things," *Measurement: Sensors*, vol. 27, November 2022, p. 100719, 2023, doi:10.1016/j.measen.2023.100719.

[26] C. R. Kumar, S. N, M. Priyadharshini, D. G. E, and K. R. M, "Face recognition using CNN and siamese network," *Measurement: Sensors*, vol. 27, March, p. 100800, 2023, doi: 10.1016/j.measen.2023.100800.

[27] J. Wan *et al.*, "Robust and Precise Facial Landmark Detection by Self-Calibrated Pose Attention Network," *IEEE Transactions on Cybernetics*, vol. 53, no. 6, pp. 3546–3560, 2023, doi: 10.1109/TCYB.2021.3131569.

[28] R. Verma, N. Bhardwaj, A. Bhavsar, and K. Krishan, "Towards facial recognition using likelihood ratio approach to facial landmark indices from images," *Forensic Science International: Reports*, vol. 5, no. October 2021, p. 100254, 2022, doi: 10.1016/j.fsir.2021.100254.

[29] R. Salhab and W. Daher, "University Students' Engagement in Mobile Learning," *European Journal of Investigation in Health, Psychology and Education*, vol. 13, no. 1, pp. 202–216, 2023, doi: 10.3390/ejihpe13010016.

[30] P. González-Gaspar *et al.*, "Analixity: An open source, low-cost analysis system for the elevated plus maze test, based on computer vision techniques," *Behavioural Processes*, vol. 193, no. November, 2021, doi: 10.1016/j.beproc.2021.104539.

[31] C. Hughes and A. Akkari, "Education needs a refocus so that all learners reach their full potential," *The Conversation*, March, 2021.

[32] K. A. Adamson and S. Prion, "Reliability: Measuring Internal Consistency Using Cronbach's α," *Clinical Simulation in Nursing*, vol. 9, no. 5, pp. e179–e180, 2013, doi: 10.1016/j.ecns.2012.12.001.