Computers and Electrical Engineering 111 (2023) 108963

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Diverse distant-students deep emotion recognition and visualization

Ala'a Harb [a], Abdalla Gad [a], Maha Yaghi [a], Marah Alhalabi [a], Huma Zia [a], Jawad Yousaf [a], Adel Khelifi [b], Kilani Ghoudi [c], Mohammed Ghazal [a,*]

[a] *Electrical, Computer, and Biomedical Engineering Department, College of Engineering, Abu Dhabi University, Abu Dhabi, 59911, United Arab Emirates*

[b] *Department of Computer Science and IT, College of Engineering, Abu Dhabi University, Abu Dhabi, 59911, United Arab Emirates*

[c] *Department of Statistics, College of Business and Economics, United Arab Emirates University, Al Ain, 15551, United Arab Emirates*

ARTICLE INFO

Keywords:
Neural Architecture Search (NAS)
Deep learning
Facial emotion recognition
Face alignment
Computer vision
Data augmentation

ABSTRACT

Distance learning through online platforms has recently emerged as a significant addition to the educational system, providing support and enhancing traditional delivery modes in schools. Instructors need visual feedback to monitor students' engagement. This research aims to investigate the emotional state of culturally diverse students in the country throughout the lecture and provide teachers with an interactive dashboard to monitor students' emotional states during a particular lecture. We collected a specialized dataset that includes Arab students posing for various emotions specifically related to lecture scenarios. Face detection and alignment algorithm is then applied. Finally, we utilize Neural Architecture Search (NAS) to optimize emotion classification architecture. The overall accuracy of the model is 86.29% on our collected dataset and 98.84% on the CK+ dataset. The live testing of the system shows the emotions clearly with real-time feedback. The AI-powered dashboard gives instructors insights into student engagement during distance learning, which induces data-driven decisions to optimize teaching strategies.

1. Introduction

Recently, the Arab region had to move to online education delivery to offer students the flexibility to pursue their education anywhere. Due to cultural concerns, Most students feel uncomfortable opening their cameras during lectures, while instructors require feedback on students' apprehension. Emotions can be expressed by nonverbal communication channels, such as facial expressions and physiological reactions, which help detect the learner's motivational state in e-learning environments [1]. Therefore, the computer must understand and interpret these expressions effectively to understand students' emotions and thoughts about the learning materials and provide teachers with visual feedback [2]. However, identifying the student's engagement based on capturing the emotional reactions in real-time during learning tasks is challenging and demands an awareness of the student's emotions. The high demand for intelligent online platforms sheds light on the importance of engaging an effective method of artificial intelligence and machine learning to enhance the distance learning experience for students and teachers. An online platform is where the students interact with various learning activities, such as reading, writing, listening to video tutorials, and online exams. During the student interaction with such activities, they show various emotional states, such as happiness, sadness, neutrality, surprise, anger, and fear. One of the communicative and expressive parts of a human being is the face, as it reflects different expressions to reveal

* Corresponding author.
E-mail addresses: 1046027@alumni.adu.ac.ae (A. Harb), abdalla.gad@adu.ac.ae (A. Gad), maha.yaghi@adu.ac.ae (M. Yaghi), marah.alhalabi@adu.ac.ae (M. Alhalabi), huma.zia@adu.ac.ae (H. Zia), jawad.yousaf@adu.ac.ae (J. Yousaf), adel.khelifi@adu.ac.ae (A. Khelifi), kghoudi@uaeu.ac.ae (K. Ghoudi), mohammed.ghazal@adu.ac.ae (M. Ghazal).

https://doi.org/10.1016/j.compeleceng.2023.108963
Received 8 August 2023; Received in revised form 18 September 2023; Accepted 24 September 2023
Available online 5 October 2023

human emotions by transmitting various emotions. Since human emotions can be expressed through facial expressions, students' engagement statuses can be obtained from recognizing such expressions.

This research aims to investigate the emotional state of diverse students in the Arab region while attending online lectures by experimenting with several deep learning algorithms for feature extraction and classification and providing teachers with an interactive dashboard to monitor students' emotional states on a real-time basis during a particular lecture. The system evaluates the students' interaction with the teaching method and content by offering teachers real-time analysis of students' emotional states during a specific online lecture. It classifies them into Bored, Distracted, Confused, Enjoying, Interested, Neutral, and Worried.

In our study, it is important to note that the data utilized for training and testing are the original images that did not contain any blackout or occlusion over the eyes. However, to present the findings in academic publications and respect the privacy of the individuals in the dataset, we have applied a blackout or occlusion over the eyes in the images shared in the paper. This precautionary measure was taken to protect the identities and privacy of the students while still conveying the essential information and results of our study.

The following is how the rest of this article is organized: Section 2 presents the literary works published by several researchers. The materials and methods of the proposed work are described and illustrated in Section 3. The findings acquired with the proposed work are presented and discussed in Section 4, whereas Section 5 concludes this paper, and the following section provides the references used in this research.

2. Related work

In this section, different emotion recognition methods in educational environments are explored. Then, the role of deep learning for emotion classification and the different datasets used by the literature for evaluation are discussed. Finally, multiple applications introduced by research for students' emotional state monitoring are covered.

2.1. Students' emotion recognition methods

The speech-based emotion classification methods proposed in [3] utilize deep learning and multimodal analysis to detect and recognize emotions in an educational environment. Furthermore, speech emotion recognition supports intelligent teaching applications by providing real-time insights into teachers' emotional states, facilitating the adjustment of teaching methods [4]. The authors in [5] classify students' emotions through spoken speech during online exams. In [6], the speech signal is converted to a 2D spectrum to improve the emotion recognition accuracy. The results show that the speech, pitch, and content provide insight to improve the learning experience. The recent advancement in Natural Language Processing (NLP) shows that it can be used for emotional classification [7]. The authors in [8] perform data mining of online forums filled by students to classify their emotional state. The authors in [9] explore sentiment analysis through students' online comments.

Li et al. [10] utilized facial feature extraction in conjunction with a novel Convolutional Neural Network (CNN) architecture. In contrast, the network introduced in [11] featured a Region-Aware Sub-Net (RASnet), an Expression Recognition Sub-Net (ERSnet), and Multiple Attention (MA) blocks. Another approach presented in [12] incorporated a distinctive CNN architecture rather than using standard pre-trained networks. The current data used for evaluation is limited in variation, which poses a challenge for network generalization and leads to favoring customized architectures as a solution. Also, the paper [13] implements a system to recognize students' emotions using their own standard CNN, which consists of 5 convolutional layers, two pooling layers, and two fully connected layers. The system was integrated into an online educational game to detect four students' emotions while playing through the computer's web camera. The pre-processing applied is all about face detection using Haar Cascades in OpenCV, where faces are cropped, converted to grayscale, and resized. Then, they inject the cropped faces into their CNN network for feature extraction, such as edges, corners, and shapes, through the convolution layer and emotion classification through the fully connected layers. Anger, disgust, fear, happiness, sadness, surprise, and neutral are the emotions detected by the system, with an accuracy rate of 97.18% on combined Cohn Kanade AU-Coded Facial Expression (CK+) [14] and Karolinska Directed Emotional Face (KDEF) [15] data sets on which the model was trained and tested. In all the previous papers, the authors calculate and analyze the confusion matrix for evaluating the system performance in classifying or predicting each emotion. The authors in [16] proposed a facial emotion recognition system based on a Histogram of Oriented Gradients (HOG) features extraction, which is fed into Fast Learning Network (FLN). The overall performance of the method reached 95.04% in accuracy and 72.73% in precision when tested on the Yale Face dataset.

2.2. The role of deep learning in emotion recognition

The appearance of the deep learning concept in machine learning has led to various advancements in computer vision and emotion recognition [17]. Most existing work applies different machine learning algorithms to perform classification based on datasets labeled with Ekman's emotions model, which are six or seven categories: disgust, anger, fear, sadness, happiness, contempt, and surprise [15,18]. Most researchers evaluate their emotions recognition deep learning algorithms by utilizing the existing video-based data sets, such as Japanese Female Facial Expression (Jaffe) [15], KDEF [15], Indonesian Mixed Emotion Dataset (IMED) [18], CK+ [14], Facial Expression Recognition 2013 data set (FER2013) [17], and Indonesian mixed emotion database (IMED) [18]. Some datasets contain images with various profile and frontal views, such as KDEF [15]. It can be noticed from the existing research works that some of these datasets are collected in a controlled environment, such as JAFFE and CK+. In the existing datasets, we can find

Table 1
Comparison between the commonly used datasets and our dataset.

Dataset	Description
Cohn–Kanade AU-Coded Facial Expression (CK+)	• 6 Basic emotions + Neutral • 100 Subject (Females & Males), 593 Video Sequences • Multi western Ethnicity, Age group: 18–50 • Front face view
Indonesian Mixed Emotion Dataset (IMED)	• 7 Basic emotions + Neutral • 15 Subject (Females & Males), 570 Video Sequences • Indonesian Ethnicity, Age Group: 17–32 • Front face view
Japanese Female Facial Expression (Jaffe)	• 6 Basic emotions + Neutral • 10 Subject (Females only), 213 Images • Japanese Ethnicity, Age Group: 17–32 • Front face view
Facial Expression Recognition 2013 (FER 2013)	• 5 Basic emotions + Neutral • 123 subject (Females & Males), 30,000 Images • Multi eastern, Age Group: 2–50 • Different Facial view and facial occlusion
Karolinska Directed Emotional Face (KDEF)	• 5 Basic emotions + Neutral • 170 amateur actors (Males & Females), 4900 Images • Multi eastern, Age group: 20–30 • Frontal view in a controlled environment
Our Diverse Student Emotional Face Dataset	• 6 education-specific emotions + Neutral • 103 Students (Females & Males), 2100 images • Culturally diverse (Considering Arabs), Age group: 17–25 • Front face view was taken in an uncontrolled environment

some of them that are biased toward a specific gender, such as JAFFE, and some of them are biased toward a specific country region, such as JAFEE and IMED, which yields nationality and gender biases, and might not be effective in recognizing emotions across students with different country and cultures. In some papers, two datasets are merged to increase the dataset size. Table 1 describes the existing facial emotion datasets in the literature and briefly compares them with our dataset.

2.3. The application of emotion recognition in educational environment

Several applications of deep learning-based emotion recognition in educational settings are found in the literature. Initial studies focus on recognizing emotions based on micro facial expressions, which shows a system implemented to detect micro-expressions of college students to recognize their anger, disgust, fear, happiness, sadness, and surprise emotions. The system shows that happiness was the easiest detected emotion, with an accuracy rate of 92.5% [19]. In contrast, fear and disgust emotions are relatively difficult to detect since students sometimes express fear as sadness and disgust as anger or sadness. However, the authors of [20] proposed a system that analyzes and recognizes students' emotions from their faces by implementing their own standard Convolutional Neural Architecture, which consists of 4 convolutional layers, four pooling layers, and two fully connected layers. They applied data augmentation techniques to transform the training set images by rotation, shifting, and horizontal flipping. Data was divided into 80% for the training set and 20% left for the training set, and they used the Stochastic Gradient Descent with Momentum (SGDM) optimizer to train the model on the training data through Keras. Fear, disgust, sadness, happiness, anger, and neutral are the emotions detected by that system, with an accuracy rate of 70% on the FER-2013 database. The system shows an effective detection of happy and surprising emotions. However, it is not good enough for detecting fear emotion, as it confuses fear with sadness. Furthermore, the authors emphasize the feasibility of emotional recognition in education as it helps teachers to modify their teaching content accordingly. In [21], the authors explore the automatic recognition of engagement from students' facial expressions, highlighting the reliability of human observers in judging engagement and the successful automation of the process using machine learning. The authors in [22] introduce MOEMO (Motion and Emotion) dashboard. This learning analytics system employs automated emotion detection through facial analysis to measure students' effective engagement and concentration in real-time. These diverse methods demonstrate the significance of emotion recognition in understanding and improving student engagement during online learning environments.

3. Materials and methods

We apply computer vision techniques to overcome challenges in emotion recognition, such as face pose, background, illumination, and occlusion [23]. In addition, Designing a deep neural network requires expert knowledge and unlimited evaluations based on a trial-and-error strategy that consumes time and resources. Hence, further advances in emotion recognition introduce a new concept in the deep learning world called Neural Architecture Search (NAS), which can automatically design network architecture and configure network models [24]. The proposed work illustrated in Fig. 1 shows the different stages we have applied for our proposed work.

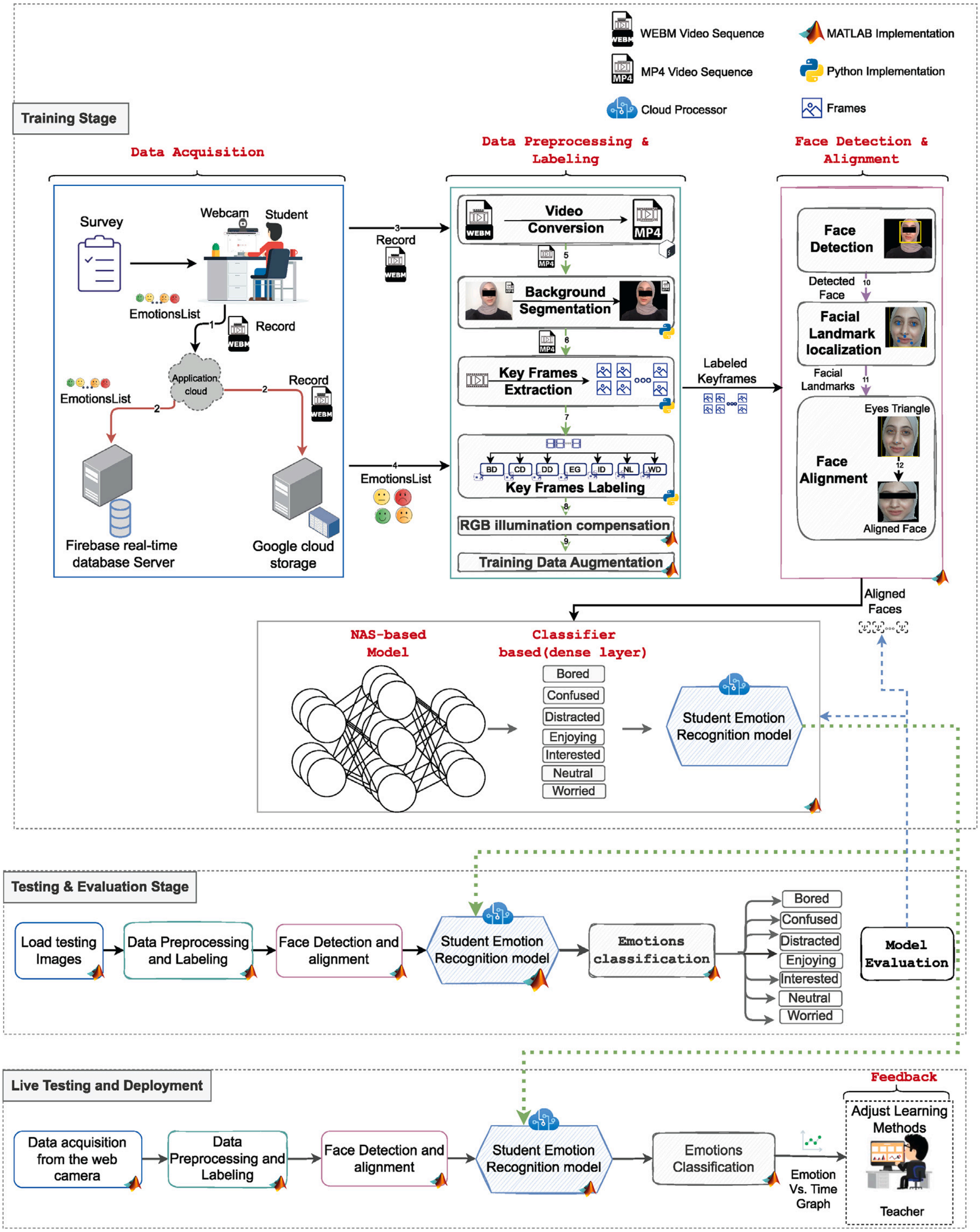


Fig. 1. System overview – Student data collected are passed into pre-processing and labeling during the training stage, followed by face detection and alignment. The aligned faces are used for training and testing the NAS model. Finally, the trained model is accessed from the cloud for emotional classification of the frames acquired from students' webcams during the lecture. The emotions vs. time graph is presented to the instructor for action.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Sad	○	○	○	○	○
Understanding	○	○	○	○	○
Happy	○	○	○	○	○
Angry	○	○	○	○	○
Fear	○	○	○	○	○
Lost	○	○	○	○	○
Worried	○	○	○	○	○
Surprised	○	○	○	○	○
Neutral	○	○	○	○	○
Frustrated	○	○	○	○	○
Bored	○	○	○	○	○
Engaged	○	○	○	○	○
Disappointed	○	○	○	○	○
Enjoying	○	○	○	○	○
Disgust	○	○	○	○	○
Excited	○	○	○	○	○
Confused	○	○	○	○	○
Distracted	○	○	○	○	○
Interested	○	○	○	○	○
Other (please specify)					

Fig. 2. Survey – Emotions list.

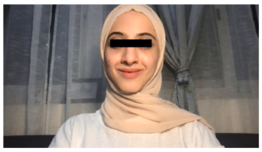


Fig. 3. Normal RGB keyframe.

3.1. Data acquisition

As discussed earlier, according to Ekman, the most current related work about student emotion recognition classifies emotions into six basic emotions. Such emotions might be basic to measuring student interaction with educational methods and materials, such as disgust and fear. Therefore, a survey was distributed among teachers and students from different colleges to determine the appropriate emotions set to assess students' engagement with the teaching methods and materials presented by teachers in an online class. A total of 253 responses were collected from 123 teachers and 130 students. The list of possible emotions to select from is shown in Fig. 2. A large dataset of 2100 images has been collected of university students using a web tool that records seven emotions, including bored, distracted, confused, enjoying, interested, neutral, and worried. The images were captured while students were watching a lesson and selecting their emotions from the list provided on the web tool.

3.2. Data prepossessing stage

Students' videos are captured with a different background where other objects might appear. As a result, extracting the face portion from the video frames becomes a challenge. Some background details behind the face might be considered part of the extracted features by the first conventional layers of the pre-trained network. The method used for background segmentation in this work is the background segmentation customized Python solution based on the landscape MobileNetV3 model followed by a joint bilateral filter. Figs. 3 and 4 demonstrate a sample of one student frame before and after the background segmentation. A black background is selected because the first layers of the pre-trained networks neglect zeros.

Relevant and necessary information retrieval is a crucial task in video analysis and processing for different reasons; one of them is that it is challenging to process a big video reasonably while preserving its semantic features. Hence, extracting keyframes is a crucial initial step in computer vision tasks [25]. Fig. 5 shows a sample of one extracted keyframe from the video sequence of one scene. Keyframes are extracted from the video of a particular student by computing absolute intraframe differences between each consecutive frame in the segmented video. Then, the average is calculated and smoothed to remove noise and, in turn, avoid repeated extraction of frames of similar scenes. The frames in which the average intraframe difference is local maxima are then considered the keyframes and stored under their corresponding label for input to the face detection algorithm.

The illumination compensation is applied to the segmented image to focus more on the student's skin pixels. The gray world algorithm estimates the illumination by computing the inverse mean of each image channel. Then, calculate the smallest average

Fig. 4. Segmented RGB frame.

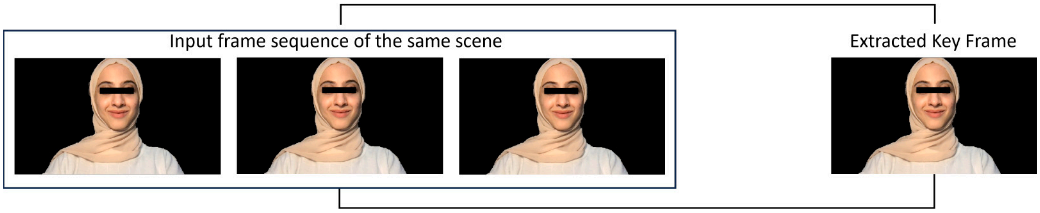


Fig. 5. Extracted keyframe from the video sequence of one scene.

Fig. 6. Normal RGB keyframe.

Fig. 7. Compensated RGB frame.

Fig. 8. Data augmentation.

value and estimate the scaling factor in each RGB channel based on that value. Figs. 6 and 7 shows a sample of one of the compensated RGB keyframes.

We apply reflection, scaling, and translation techniques for image augmentation, as shown in Fig. 8. A horizontal reflection enables a random image reflection in the left/right direction. Vertical and horizontal scaling is applied to crop the edges of the student's face, which may cover the full image or not appear clearly in the image. A vertical and horizontal translation is applied to detect and recognize students' faces in any part of the image. Data augmentation is applied only to the training data, so we generate further data as we have a small dataset. Also, we want to measure the effect of data augmentation on emotion recognition accuracy.

Since we are aiming to recognize students' emotions from their faces, the face area is cropped from each student frame in the training and testing dataset according to the bounding box values calculated by processing all frames through a Multi-task Cascaded Convolutional Networks (MTCNN) [26] as shown in Fig. 1 under Face detection and Alignment box. Afterward, all cropped faces are aligned and resized to match the size of the pre-trained network [244 244 3]. Fig. 9 represents the geometric-based face alignment algorithm implemented. MTCNN is also used to localize the facial landmarks in the detected face and produce a landmark object containing five facial center points (x,y) for the left and right eyes, nose, and mouth corners. The face alignment algorithm implemented in this work is based on eye alignment, where the left and right eye points (x,y) are extracted from the facial landmarks

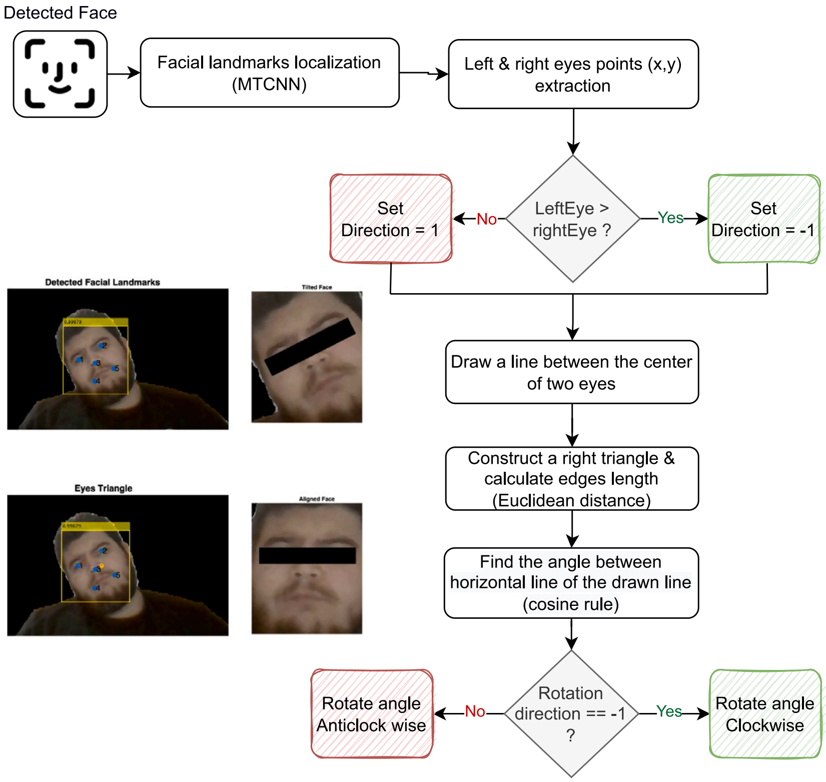


Fig. 9. Face alignment algorithm using MTCNN and geometry.

object, where a line is drawn to connect the left and right points (x,y). Then, a right triangle is constructed according to the line drawn, where all three edges' lengths are calculated by finding the Euclidean distance. The Euclidean distance value of all three edges is used to find the tilt angle. The rotation direction is either done clockwise or anti-clockwise, and it is determined based on which eye is larger; if the y-axis value of the left eye is larger than the right eye, then the rotation will be in the clockwise direction (90-angle). Otherwise, it will be anti-clockwise direction (angle).

3.3. Neural Architecture Search (NAS)

Unlike the existing deep transfer network, recent NAS benchmarks only focus on architecture optimization, despite the effect of training hyperparameters on the obtained model performances [27]. The primary NAS consists of the following steps: the search space to discover the Neural Network (NN) architecture available in the space that suits our problem; it can be either a simple NN or deep NN; the search strategy to select the architecture from a predefined search space, and the performance evaluation to training and evaluate the selected architecture on our dataset. Then, it determines which architecture achieves high predictive performance on unseen/testing data. A discrete search space is a search space where each architecture is a unique combination of blocks. This means that the search space is finite and can be enumerated. Discrete search spaces are typically used in NAS algorithms that rely on reinforcement learning or evolutionary algorithms [24]. The discrete search space concept we have applied included the Sequential, Multi-Layer Perceptron (MLP), 2D convolution layer, 2D max pooling layer, and dropout layer blocks. The process goes over 100 trials before it finds the most optimum architecture.

Training the ResNet-50 network on our collected dataset for emotion classification did not perform well. Although extensive hyperparameter tuning was applied, the dataset has proven to be of higher complexity than the benchmark datasets. For this reason, we followed the path of designing and testing our novel architecture with the help of NAS. Unlike the existing deep transfer network, recent NAS benchmarks only focus on architecture optimization, despite the effect of training hyperparameters on the obtained model performances [27].

Recently, NAS has become common when tackling the emotion classification issue. The EmoNAS network search algorithm introduced in [28] employs 5 cells with 7 nodes and a limited search space of 7 operations to create lightweight architectures, achieving significant performance improvements over other NAS methods across multiple FER datasets. The authors in [29] proposed a network search algorithm for microexpression recognition named AutoMER, which aims to find the most efficient 3-D CNN architecture. The introduced method is based on parallelogram design-based search space and employs a spatiotemporal feature module called "3-D singleton convolution" with various candidate operators and 3-D dilated convolution operators. The algorithm

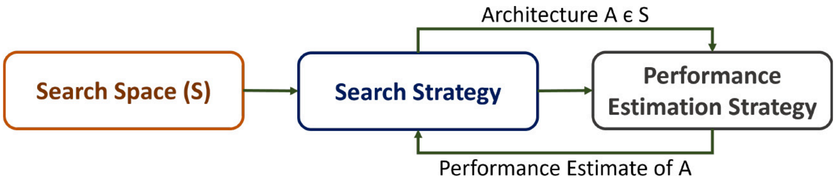


Fig. 10. General overview of the NAS algorithm.

outperformed existing approaches when evaluated on five microexpression datasets. In [30], the introduced network, ENAS-FERNet, is designed through evolutionary NAS for facial expression recognition. The network showed promising results regarding time consumption and performance with noisy data.

The NAS approach tackles the design of Artificial Neural Networks (ANNs) through three key components [24], as shown in Fig. 10: the search space, search strategy, and performance estimation strategy.

The search space defines all the architectures to be explored by the process. The building blocks we define for our approach include Sequential, Multi-Layer Perceptron (MLP), 2D convolution layers, 2D max-pooling layers, and dropout layers. The search strategy guides the exploration of the multiple neural architectures defined in the search space. We select the Bayesian optimization for our application, utilizing accuracy as the primary performance metric. The process goes for 100 trials before selecting the optimum architecture. The dataset used for training and testing in the NAS algorithm contains the student faces labeled according to their corresponding emotion. It is extracted after applying three of the image pre-processing techniques, which are (1) RGB illumination compensation, (2) background segmentation, and (3) keyframes extracted. Fig. 11 shows the NAS-based network architecture implemented with one input layer that accepts an RGB image with a size of 244x244x3, 54 deep layers, and one fully connected layer which classifies emotions into one of the seven categories.

3.4. Performance evaluation matrices

The performance of each model experimented on, whether it is based on multi-class or binary-class classification, is evaluated by looking at the value of the entropy loss computed by the softmax layer or by computing the mean square error. Moreover, each one of the experiments conducted is evaluated based on constructing a confusion matrix for each experiment by aggregating a set of evaluation metrics, such as recall, precision, and accuracy, which helps in indicating how well the model performs for each class as well as how well the model performs overall. However, for the live testing, the performance evaluation matrices are different. First, each participant is asked to express a set of ground truth emotions where the predicted emotion displayed on the top of the bounding box on the participant's face is captured. During the entire live testing period, a video was recorded to notice the duration the participant is asked to express a certain emotion and consider that as ground truth to compare with the predicted emotion shown. The second and the third blocks of Fig. 1 illustrated the process taken toward the testing conducted in the selected testing dataset and the real-time live testing. The softmax equation in (1) defines the prediction score for each class where P is the probability of class i, e^{z_i} is the exponential of each element in vector z, k is the total number of classes and j is the class index.

$$P(\text{class } i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{1}$$

4. Experiments and results

4.1. Dataset and experimental setup

As discussed earlier, experiments are conducted to classify seven emotions. The data is collected from 103 (53 Females + 50 Males) students selected randomly across the department at our college. Unlike existing datasets, our dataset is universal as it contains data for participants from different nationalities and cultures, as seen in Fig. 12. The data are split into training 70%, validation 10%, and testing 20%. All models are trained on the same training set, tuned hyper-parameters on the same validation set, and compared prediction performance on the same unseen testing data. Regarding live testing, another two students are selected for testing the real-time emotion recognition, considering that none previously took part in any of the mentioned data divisions. Table 2 Shows the training parameters.

4.2. Analyze NAS model

Table 3 presents various architectures' performance metrics on the same testing data from the collected dataset. The architectures include VGG16, VGG19, GoogLeNet, ResNet-18, ResNet-50, and a custom NAS-based model. The metrics recorded are Recall, Precision, and Accuracy. VGG16 achieved a Recall of 30.48%, a Precision of 43.21%, and an Accuracy of 30.49%. VGG19 showed slightly improved performance with a Recall of 35.61%, Precision of 38.86%, and Accuracy of 35.62%. GoogLeNet further improved the results, achieving a Recall of 37.71%, a Precision of 47.99%, and an Accuracy of 37.71%. ResNet-18 had comparable results,

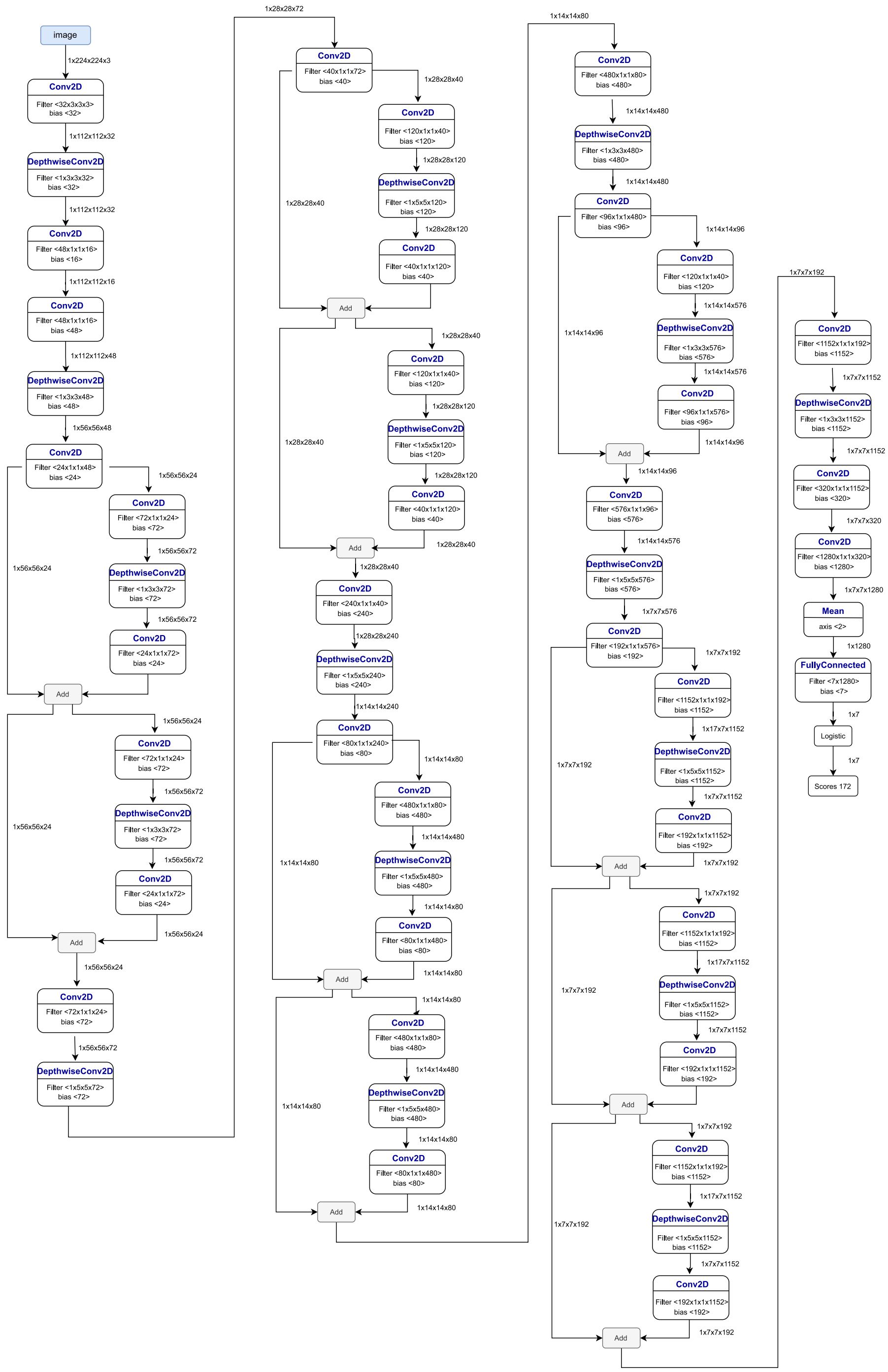


Fig. 11. NAS-based model consists of 54 deep layers.

with a Recall of 36.56%, a Precision of 37.90%, and an Accuracy of 36.56%. ResNet-50 demonstrated significant improvement, achieving a Recall of 66.88%, Precision of 64.66%, and Accuracy of 64.70%. The custom NAS-based model (Ours) exhibited the best overall performance, with an impressive Recall of 82.89%, Precision of 89.37%, and an impressive Accuracy of 86.29%.

Fig. 13 shows the confusion matrix produced after testing the NAS-based model on the testing dataset. The highest prediction accuracy was 98% and 96% for Enjoying and confused emotions, respectively. The lowest prediction accuracy was 73% for worried anxiety, whereas the model misclassified worried mostly as confused and neutral. The model could predict the rest of the emotions in the range of 70%–88%.

Fig. 12. Student samples from our constructed emotions dataset.

Table 2
The parameters for training on the collected dataset.

Training	1470
Validation	210
Testing	420
Learning Rate	0.001
Batch Size	32
Optimizer	Adam Optimizer
Loss Function	Cross Entropy
Epochs	100
Number of Layers	54
Total Learnables	3.5 Million

Table 3
Comparison of the performance of different architectures trained and tested on the collected dataset.

Architecture	Recall	Precision	Accuracy
VGG16	30.48%	43.21%	30.49%
VGG19	35.61%	38.86%	35.62%
GoogLeNet	37.71%	47.99%	37.71%
ResNet-18	36.56%	37.90%	36.56%
ResNet-50	66.88%	64.66%	64.70%
NAS-based (Ours)	**82.89%**	**89.37%**	**86.29%**

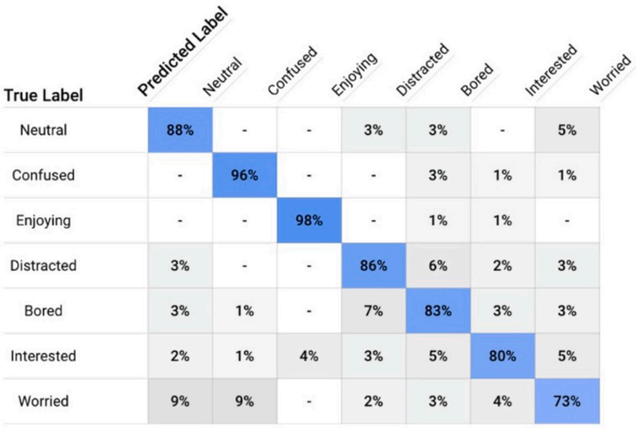
Table 4
Accuracy Comparison of Emotion Recognition Methods on the CK+ Dataset.

Reference	Method	Accuracy
Li et al. [10]	Face Extraction + CNN	97.38
Gan et al. [11]	RASnet + ERSnet + MA blocks	96.28
eXnet [12]	CNN	95.81
ProposedApproach	Face Extraction & Alignment +NAS based CNN	**98.84**

We comprehensively evaluated various state-of-the-art methods for emotion recognition on the CK+ dataset [14], as shown in Table 4. Our proposed approach outperformed the other techniques, combining facial feature extraction, facial alignment, and a CNN architecture based on NAS with an accuracy of 98.84%.

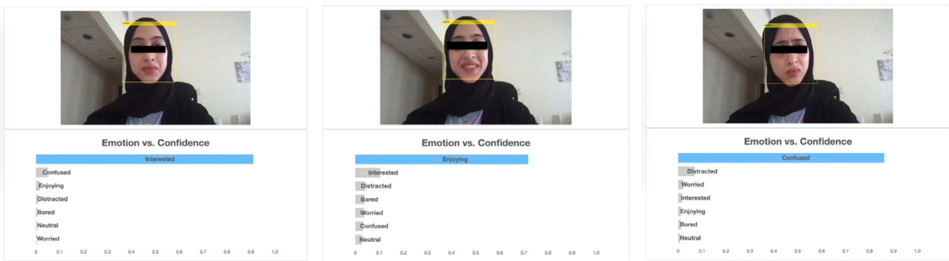
4.3. Live testing

The images captured of students are processed for face extraction and alignment. The processed image is then passed to the trained CNN found in the cloud utilizing the NVIDIA Tesla T4 GPU for inference. The accuracy of our classification model was

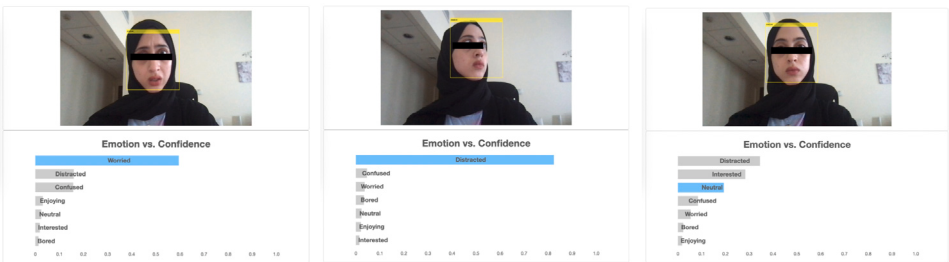


True Label \ Predicted Label	Neutral	Confused	Enjoying	Distracted	Bored	Interested	Worried
Neutral	88%	-	-	3%	3%	-	5%
Confused	-	96%	-	-	3%	1%	1%
Enjoying	-	-	98%	-	1%	1%	-
Distracted	3%	-	-	86%	6%	2%	3%
Bored	3%	1%	-	7%	83%	3%	3%
Interested	2%	1%	4%	3%	5%	80%	5%
Worried	9%	9%	-	2%	3%	4%	73%

Fig. 13. Confusion matrix of NAS Model.



(a) Actual Emotion: Interested - Prediction Confidence: 0.9 (b) Actual Emotion: Enjoying - Prediction Confidence: 0.7 (c) Actual Emotion: Confused - Prediction Confidence: 0.89



(d) Actual Emotion: Worried - Prediction Confidence: 0.6 (e) Actual Emotion: Distracted - Prediction Confidence: 0.86 (f) Actual Emotion: Neutral - Prediction Confidence: 0.2

Fig. 14. Live testing results of NAS model.

evaluated by simulating it as a platform to predict students' emotions in the live testing scenario as a prototype. The live testing results based on the selected NAS for some emotions are depicted in Fig. 14. This approach facilitates the performance analysis by identifying false predictions. Live testing results prove the automated NAS model's capability of recognizing all emotions in real-time scenarios with a high percentage of confidence. Emotions studied in this research are different from the primary emotions that used to be classified by existing research work, and that opens the door to machine learning and the ability to classify a broad group of emotions. Lastly, the model is deployed in the cloud to implement the facial emotion recognition system and prepare it for real-world use. This allows minimal hardware requirements for data acquisition and image processing while machine learning runs online. In addition, Fig. 15 shows a prototype implemented to provide students' emotion-representative analysis based on the testing results so that teachers know students' understanding and interaction with the learning material during several online lectures.

5. Conclusions

To sum up, this research proposed a diverse student emotion recognition system that can be integrated with an e-learning environment to improve the distance learning experience for diverse students. The proposed system can be used widely in schools

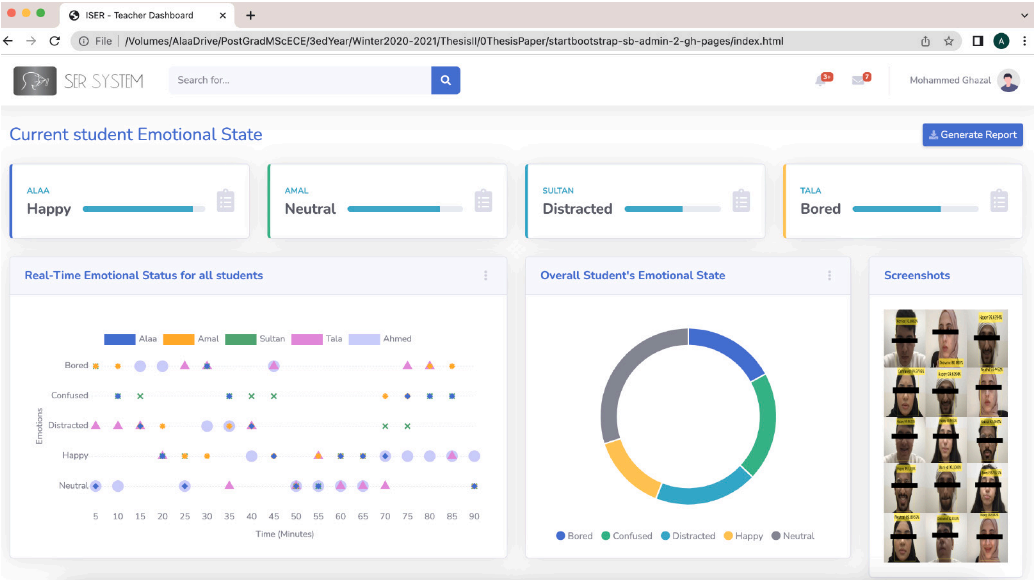


Fig. 15. Interactive dashboard to monitor students' emotional state in real-time.

and universities to offer an intelligent technique for revealing students' engagement with e-learning materials by detecting their emotions based on facial expressions. A customized data collection tool was implemented and distributed among university students to determine suitable emotions in educational settings. The emotions studied in this research differ from those used to be classified by other researchers, which opens the door to the machine learning capability of classifying a broad group of emotions. The main challenges found during the data collection and implementation are that students express their emotions differently. Since the data set tackled students with different cultures, it is even more challenging. The data was trained through NAS to enhance the performance of the ResNet-50 model, and the results show a noticeable improvement in the overall accuracy of 86.29%. Hence, the NAS-based model was selected as the baseline model for the system, and the live testing was based on it. We also trained and tested our proposed model on the CK+ dataset, resulting in an accuracy of 98.84%, outperforming the state-of-the-art methods. The live testing results show minimal confidence in predicting neutral emotion, which might be controlled with a certain threshold. The proposed system can be adopted in an online educational platform linked with the interactive dashboard to provide teachers with real-time analytical feedback about students' understanding and interaction with the learning material during the lecture.

CRediT authorship contribution statement

Ala'a Harb: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Abdalla Gad:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Maha Yaghi:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing — original draft, Writing – review & editing, Visualization. **Marah Alhalabi:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing — original draft, Writing – review & editing, Visualization. **Huma Zia:** Supervision, Investigation, Funding, Writing – review & editing, Validation. **Jawad Yousaf:** Supervision, Investigation, Funding, Writing – review & editing, Validation. **Adel Khelifi:** Supervision, Investigation, Funding, Writing – review & editing, Validation. **Kilani Ghoudi:** Supervision, Investigation, Funding, Writing – review & editing, Validation. **Mohammed Ghazal:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This research work is supported by Abu Dhabi National Oil Company (ADNOC), Emirates NBD and Sharjah Electricity Water & Gas Authority (SEWA), Dubai Electricity and Water Authority R&D Centre as the sponsors of the 3rd Forum for Women in Research (QUWA): Women Empowerment for Global Impact at the University of Sharjah. The authors also acknowledge financial support from Abu Dhabi University's Office of Research and Sponsored Programs. Grant number: 19300792.

References

[1] Tonguç G, Ozaydın Ozkara B. Automatic recognition of student emotions from facial expressions during a lecture. Comput Educ 2020;148:103797. http://dx.doi.org/10.1016/j.compedu.2019.103797.

[2] Dewan M, Murshed M, Lin F. Engagement detection in online learning: a review, 6. 2019, http://dx.doi.org/10.1186/s40561-018-0080-z.

[3] Singh M, Bharti S, Kaur H, Arora V, Saini M, Kaur M, et al. A facial and vocal expression based comprehensive framework for real-time student stress monitoring in an iot-fog-cloud environment. IEEE Access 2022;10:63177–88. http://dx.doi.org/10.1109/ACCESS.2022.3183077.

[4] Zhang S, Li C. Research on feature fusion speech emotion recognition technology for smart teaching. Mob Inf Syst 2022;2022:1–9. http://dx.doi.org/10.1155/2022/7785929.

[5] Wen S. An analysis of emotional responses of students in bilingual classes and adjustment strategies. Int J Emerg Technol Learn (iJET) 2023;18:100–14. http://dx.doi.org/10.3991/ijet.v18i01.37125.

[6] Bekmanova G, Yergesh B, Sharipbay A, Mukanova A. Emotional speech recognition method based on word transcription. Sensors (Basel, Switzerland) 2022;22. http://dx.doi.org/10.3390/s22051937.

[7] Muis KR, Etoubashi N, Denton CA. The catcher in the lie: The role of emotions and epistemic judgments in changing students' misconceptions and attitudes in a post-truth era. Contemp Educ Psychol 2020;62:101898. http://dx.doi.org/10.1016/j.cedpsych.2020.101898, URL https://www.sciencedirect.com/science/article/pii/S0361476X20300631.

[8] Abdelhamid A. Speech emotions recognition for online education. Fusion: Pract Appl 2022;10:78–87. http://dx.doi.org/10.54216/FPA.100104.

[9] Ren Y, Tan X. Research on the method of identifying students' online emotion based on Albert. In: 2021 international conference on intelligent computing, automation and applications. 2021, p. 646–50. http://dx.doi.org/10.1109/ICAA53760.2021.00118.

[10] Li K, Jin Y, Akram M, Han R, Chen J. Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. Vis Comput 2020;36. http://dx.doi.org/10.1007/s00371-019-01627-4.

[11] Gan Y, Chen J, Yang Z, Xu L. Multiple attention network for facial expression recognition. IEEE Access 2020;8:1. http://dx.doi.org/10.1109/ACCESS.2020.2963913.

[12] Riaz M. exnet: An efficient approach for emotion recognition in the wild. Sensors 2020;20(4):1087. http://dx.doi.org/10.3390/s20041087.

[13] El Hammoumi O, Benmarrakchi F, Ouherrou N, El Kafi J, El Hore A. Emotion recognition in e-learning systems. In: 2018 IEEE 6th international conference on multimedia computing and systems. 2018, p. 1–6. http://dx.doi.org/10.1109/ICMCS.2018.8525872.

[14] Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE computer society conference on computer vision and pattern recognition - workshops. 2010, p. 94–101. http://dx.doi.org/10.1109/CVPRW.2010.5543262.

[15] Akhand M, Roy S, Siddique N, Kamal M, Shimamura T. Facial emotion recognition using transfer learning in the deep CNN. Electronics 2021;10:1036. http://dx.doi.org/10.3390/electronics10091036.

[16] Alsemawi MRM, Mutar MH, Ahmed EH, Hanoosh HO, Abbas AH. Emotions recognition from human facial images based on fast learning network. Indonesian J Electr Eng Comput Sci 2023;30(3):1478–87. http://dx.doi.org/10.11591/ijeecs.v30.i3.pp1478-1487, cited by: 0.

[17] Navaz AN, Adel SM, Mathew SS. Facial image pre-processing and emotion classification: A deep learning approach. In: 2019 IEEE/ACS 16th international conference on computer systems and applications. 2019, p. 1–8. http://dx.doi.org/10.1109/AICCSA47632.2019.9035268.

[18] Nafis AF, Adni Navastara D, Yuniarti A. Facial expression recognition on video data with various face poses using deep learning. In: 2020 12th international conference on information technology and electrical engineering. 2020, http://dx.doi.org/10.1109/icitee49829.2020.9271740.

[19] Chen L, Shen X, Yang H, Sui H. Micro expression recognition training in college students 0. 2018, p. 1–3. http://dx.doi.org/10.1109/ACIIAsia.2018.8470376.

[20] Lasri I, Solh AR, Belkacemi ME. Facial emotion recognition of students using convolutional neural network. In: 2019 IEEE third international conference on intelligent computing in data sciences. 2019, p. 1–6. http://dx.doi.org/10.1109/ICDS47004.2019.8942386.

[21] Sharma P, Joshi S, Gautam S, Maharjan S, Khanal SR, Reis MC, et al. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. 2023, arXiv:1909.12913.

[22] Hasnine M, Nguyen Tan H, Tran T, Bui H, Akçapınar G, Ueda H. A real-time learning analytics dashboard for automatic detection of online learners' affective states. Sensors (Basel, Switzerland) 2023;23. http://dx.doi.org/10.3390/s23094243.

[23] Uchôa V, Aires K, Veras R, Paiva A, Britto L. Data augmentation for face recognition with CNN transfer learning. 2020, p. 143–8. http://dx.doi.org/10.1109/IWSSIP48289.2020.9145453,

[24] Ren P, Xiao Y, Chang X, Huang P-Y, Li Z, Chen X, et al. A comprehensive survey of neural architecture search: Challenges and solutions. 2021, arXiv:2006.02903.

[25] Gawande U, Hajari K, Golhar Y. Deep learning approach to key frame detection in human action videos. In: Sadollah A, Sinha TS, editors. Recent trends in computational intelligence. Rijeka: IntechOpen; 2020, http://dx.doi.org/10.5772/intechopen.91188, Ch. 7.

[26] Zhang K, Zhang Z, Li Z, Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 2016;23(10):1499–503. http://dx.doi.org/10.1109/LSP.2016.2603342.

[27] Hirose Y, Yoshinari N, Shirakawa S. Nas-hpo-bench-II: A benchmark dataset on joint optimization of convolutional neural network architecture and training hyperparameters. In: 13th Asian conference on machine learning, in proceedings of machine learning research, vol. 157. 2021, p. 1349–64. http://dx.doi.org/10.48550/arXiv.2110.10165.

[28] Baldeon Calisto M, Lai-Yuen SK. Emonas-net: Efficient multiobjective neural architecture search using surrogate-assisted evolutionary algorithm for 3d medical image segmentation. Artif Intell Med 2021;119:102154. http://dx.doi.org/10.1016/j.artmed.2021.102154, URL https://www.sciencedirect.com/science/article/pii/S0933365721001470.

[29] Verma M, Reddy MSK, Meedimale YR, Mandal M, Vipparthi SK. Automer: Spatiotemporal neural architecture search for microexpression recognition. IEEE Trans Neural Netw Learn Syst 2022;33(11):6116–28. http://dx.doi.org/10.1109/TNNLS.2021.3072290.

[30] Deng S, Lv Z, Galv´an E, Sun Y. Evolutionary neural architecture search for facial expression recognition. IEEE Trans Emerg Top Comput Intel 2023;1–15. http://dx.doi.org/10.1109/TETCI.2023.3289974.