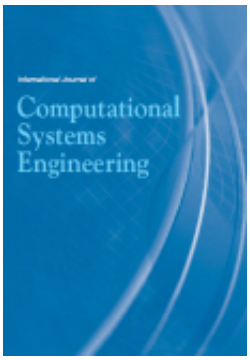
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Computational Systems Engineering

ISSN online: 2046-3405 - ISSN print: 2046-3391
https://www.inderscience.com/ijcsyse

Design of interactive feedback system for university online education classroom based on emotion recognition CNN model

Hongyun Jie, Xiaozhe Li

DOI: 10.1504/IJCSYSE.2026.10081733

Article History:

Received:	14 August 2025
Last revised:	23 May 2026
Accepted:	14 September 2026
Published online:	27 October 2026

Copyright © 2026 Inderscience Enterprises Ltd.

Design of interactive feedback system for university online education classroom based on emotion recognition CNN model

Hongyun Jie and Xiaozhe Li*

Hebei Open University,
Shijiazhuang City,
Hebei Province, 050080, China
Email: wazggcd77@163.com
Email: kfdxlxz@163.com
*Corresponding author

Abstract: Virtual university classrooms present three interconnected problems that degrade learning quality: students struggle to assess their own emotional states, engagement levels remain low, and response latencies impair real-time interaction. It presents a multimodal framework for emotion detection and interactive response that combines image, audio, and text data with convolutional neural networks (CNN) to create a closed-loop system with both emotion estimation and the building of smart responses. The framework exhibits high accuracy in emotional detection, accompanied by low detection latency, with an average recognition latency of 149.2 milliseconds, making it suitable for real-time use. Instructors are notified, and students are offered personalised recommendations based on the emotional estimations, thus promoting personalised learning paths and improving interactions. Experimental results show an accuracy rate of 93.3% in emotional recognition, 3.01 times more student-to-student interaction within a 30-minute interval, 0.34 participation depth improvement, and 9.97% performance improvement. These results highlight the efficacy of enhancing real-time classroom interactions, stability, and learning assistance with the presented system, while also demonstrating a feasible solution for the improvement of smart classroom environments.

Keywords: online education, emotion recognition; convolutional neural networks; CNN model; interactive feedback; multimodal system; real-time optimisation; classroom interaction; student participation.

Reference to this paper should be made as follows: Jie, H. and Li, X. (2026) 'Design of interactive feedback system for university online education classroom based on emotion recognition CNN model', *Int. J. Computational Systems Engineering*, Vol. 10, No. 12, pp.1–12.

Biographical notes: Hongyun Jie holds a Master's degree and currently serves as an Associate Professor and the Director of the Department of Public Administration at Hebei Open University. Her primary research interests focus on public management and adult education. To date, she has published more than 20 academic papers in related fields.

Xiaozhe Li holds a Master's degree and currently works as a Lecturer at Hebei Open University. His main research interests focus on administrative management, and adult education. To date, he has published six academic papers in these areas.

1 Introduction

The expansion of online learning has transformed conventional classrooms, particularly in fostering emotional connections and enhancing student academic achievement. Reduced student-teacher interaction, student distractions, and delayed feedback systems produce negative experiences frequently termed 'zoom fatigue,' which lowers motivation and engagement in e-learning Toscu (2023). Such hindrances are more evident in certain scenarios, such as English language learning, medical education, and large lecture classes Chang et al. (2021). Emotion recognition-based technologies have been suggested to be potential remedies because of their ability to identify learners' emotional states. However, current models face several limitations, including poor accuracy rates, long response times, and poor flexibility within virtual classrooms Savchenko et al. (2022). In line with these limitations, this work proposes a multimodal emotion recognition system tailored to virtual learning environments. Such a system incorporates visual, auditory, and text data through a convolutional neural network (CNN) architecture supplemented by attention mechanisms, thus enhancing real-time recognition accuracy in learning environments N. Khan et al. (2023). It employs model compression and parameter quantisation to enable low-latency execution,

Copyright © The Author(s) 2026. Published by Inderscience Publishers Ltd. This is an Open Access Article distributed under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)

which is fundamental for maximising online learning capabilities. The system shows performance measures, with an accuracy rate of 93.3% and a response time of 149.2 milliseconds, meeting the requirements for real-time use. Moreover, it highly improves student engagement, interaction, and academic achievement, highlighting its effectiveness in creating a responsive and emotionally supportive learning space.

2 Related work

The expanding integration of advanced educational technologies has elicited research into complex correlations between students' emotional states and their actions in educational settings. Recent research, as can be seen from Gupta et al., investigated the identification of emotional states in cooperative educational settings using deep learning networks and pointed out associated challenges regarding emotion categorisation, integrating multimodal information, and adaptability to diverse environments Gupta et al. (2023). On the other hand, review work by Hosseini et al. (2024), provided a theoretical model seeking to direct future multimodal emotional recognition advancements whereas Jiang et al. introduced a fuzzy temporal convolutional network seeking to improve feelings modelling using attention techniques Jiang et al. (2023). Aqach and Sümer's work has immensely contributed to sentiment analysis and multimodal measures of engagement in educational settings, especially regarding massive open online courses (MOOCs) (Baqach and Battou, 2024; Sümer et al., 2021). Nevertheless, even with existing advances, challenges persist, especially those regarding information integration in real-time as well as computational efficiency, especially in large-scale and diverse classroom settings Chacón-Cuberos et al. (2021). Recent research has focused on furthering model accuracy as well as operational effectiveness in real-time settings. Yan et al. proposed a double-channel framework using high-level attention that combined BERT, CNN, and LSTM, but they faced significant computational inefficiencies Yan et al. (2023). Likewise, performance-oriented non-real-time models proposed by Mondal et al. (2024) also faced challenges regarding their non-real-time performance (Khanna et al., 2024). These concerns highlight the need for advanced emotional recognition models that will effectively work in real-time, especially in educational contexts. Liu's review considered key issues regarding the effectiveness of emotion recognition and how it caters to course offerings Liu et al. (2021), whereas the present study promotes a CNN-based multimodal framework for emotional recognition. One proposed system balances computational exigencies without compromising significant accuracy, combining image, voice, and text data to enable real-time outputs for teacher push-notifications and individual student recommendations. Several recent studies published between 2023 and 2024 have advanced multimodal emotion recognition in educational contexts. Yan et al. (2023) proposed a dual-channel attention network combining BERT, CNN, and LSTM for sentiment analysis of online course reviews, achieving 87.4% accuracy but with inference latency exceeding 300 milliseconds. Gupta et al. (2023) developed a real-time learner engagement detection system using facial emotion recognition with a modified VGG architecture, reporting 89.1% accuracy on a private classroom dataset but without multimodal integration. Mondal et al. (2024) applied sentiment analysis to user reviews of interactive educational apps using a bidirectional LSTM, achieving an F1 score of 0.81 but not addressing real-time feedback generation. The present work differs from these efforts by simultaneously achieving three objectives that prior studies treated separately: multimodal fusion of image, speech, and text, sub-200 millisecond latency suitable for live classroom interaction, and closed-loop feedback that adapts teaching strategies based on detected emotional states. Through edge computing, it enables real-time effectiveness with reduced latency and enhanced privacy through localised affective data processing, thereby providing secure and resource-aware educational environments Kuchuk and Malokhvii (2024).

3 Construction and optimisation of multimodal emotion recognition model

3.1 Emotion recognition CNN model construction

The model combines three modalities, image, speech, and text, extracted from online classroom data to support emotion recognition. Image data is processed through the MTCNN algorithm, which will crop and normalise face images into 48 × 48 grayscale formats Khan et al. (2024). Speech features are extracted by using MFCC following the suppression of silence, resulting in a 39-dimensional temporal series Singh (2024). The text data is converted through Word2Vec, resulting in a sequence of vectors with 300 dimensions Johnson et al. (2024). The architecture used for emotion recognition involves three parallel convolutional networks: the image channel uses modules that consist of Conv2D-BatchNorm-BN-ReLU-MaxPool; the speech channel applies a 1D-CNN to extract temporal features Uddin et al. (2022); and the text channel applies TextCNN to recognise syntactic structures Liu et al. (2023). The extracted features are concatenated and then input into a fully connected network to support fusion.

In a convolutional neural network, the output of the convolutional layer can be expressed as follows:

$$\mathbf{Y} = \mathbf{W} * \mathbf{X} + \mathbf{b} \tag{1}$$

In equation (1)

Y is the output of the convolution layer

W is the convolution kernel

X is the input data

b is the bias; and the symbol '*' represents the convolution operation.

To improve discrimination ability, an attention mechanism is applied to weight and enhance modal features He et al. (2021). The output layer uses the softmax function to classify 6 types of emotional states. The overall model uses multi-classification cross-entropy as the loss function Wu et al. (2024), and is trained with the Adam optimiser, with learning rate scheduling and Dropout to improve stability and robustness.

Using the attention mechanism for weighting:

$$\mathbf{F}_{\text{att}} = \sum_{i=1}^{n} \alpha_i \mathbf{F}_i \tag{2}$$

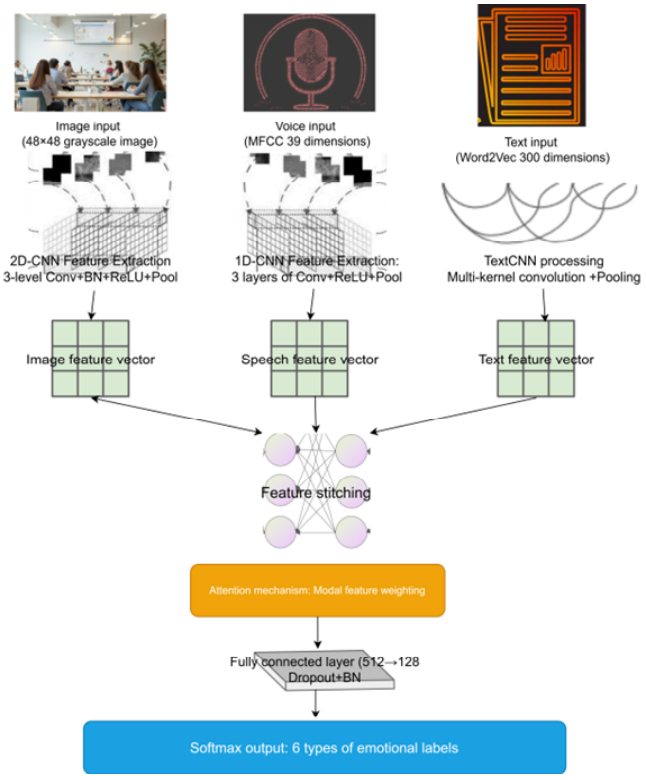
In equation (2)

$\mathbf{F}_i$ is the feature of the i^{th} mode

α_i is the attention weight

$\mathbf{F}_{\text{att}}$ is the feature after weighted fusion; the value of α_i is learned during the training process.

Figure 1 Emotion recognition CNN model architecture with stepwise data flow (see online version for colours)



The developed architecture tailored for academic environments comprises two main contributions towards enhancing emotional recognition in online learning environments. First, a temporal attention mechanism adjusts the importance of emotional features based on time, thus encouraging consistency across long courses by recognising emotional trends at different intervals. Second, to counter potential interference caused by students closing their cameras, the model has been trained to be resilient to such interruptions by leveraging auditory and textual information to attain accurate emotion recognition. Together, these contributions improve system resilience and maintain effective operation even when certain multimodal inputs are missing, as illustrated in Figure 1.

The architecture proceeds in four sequential stages. Stage 1 accepts three input streams: 48 × 48 grayscale face images processed by MTCNN, 39-dimensional MFCC feature sequences from speech, and 300-dimensional Word2Vec embeddings from text. Stage 2 applies modality-specific convolutional networks: a 2D-CNN with three Conv2D-BatchNorm-ReLU-MaxPool blocks for images, a 1D-CNN with four convolutional layers for speech, and a TextCNN with three parallel convolution kernels of sizes 3, 4, and 5 for text. Stage 3 concatenates the three feature vectors and passes them through a channel attention mechanism that computes modality-specific weights αi. Stage 4 feeds the weighted feature sum into two fully connected layers with 256 and 128 neurons respectively, followed by a softmax output layer producing six emotion class probabilities.

Figure 1 illustrates the structure of a CNN-based model for emotion recognition tailor-made for online university classrooms that covers image, speech, and text data. The structure utilises a 2D-CNN to process facial expression features, a 1D-CNN to handle MFCC from speech, and TextCNN for semantic emotional content extraction. The varied features are combined using an attention mechanism, resulting in the classification of outputs into six different emotion classes. The model ensures high accuracy and real-time performance, hence suitable for dynamic classroom environments.

3.2 Model optimisation and training

This work aims to augment a CNN-based emotion classification model with increased classification accuracy and training stability while not compromising inference efficiency. The chief optimisations are the use of the Adam optimiser with specific learning rate parameters (0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$), dynamic learning rate adjustment based on validation loss, and comparative testing of four different optimisers (Adam, SGD, RMSProp, Adagrad). The dynamic learning rate is set up to be reduced when validation loss is at a plateau, with a minimum of 1e-5 enforced. Table 1 summarises the experimental results.

Table 1 Comparative experimental results of different optimisers

Optimiser	*Accuracy* (%)	*F1-score*	*Training time* (*s*)	*Final loss*
Adam	91.4	0.892	248	0.32
SGD	87.1	0.842	301	0.47
RMSProp	89.3	0.865	266	0.38
Adagrad	85.7	0.818	289	0.52

The comparative study reported in Table 1 highlights the outstanding effectiveness of the Adam optimiser, obtaining the highest accuracy (91.4%), F1 measure (0.892), and convergence time (248 seconds) while recording the lowest

loss (0.32). By employing an adaptive gradient approach to effectively handle fluctuations, Adam outperforms the performance parameters of SGD and RMSProp. In contrast, Adagrad's relatively lower accuracy and F1 measure limit its applicability in real-time applications in online learning platforms.

During the training process, the cross-entropy loss can be expressed as:

$$L = -\sum_{i=1}^{C} y_i \log(p_i) \tag{3}$$

In equation (3)

C is the number of emotion categories

y_i is the binary representation of the true label

p_i is the probability of the i^{th} category predicted by the model.

At the same time, to ensure the stability of the model training process, dynamic learning rate scheduling is employed. Equation (4) shows the updated formula of the learning rate:

$$\eta_{new} = \eta_{old} \times \gamma^{k} \tag{4}$$

In equation (4)

η_{old} is the current learning rate

η_{new} is the updated learning rate

γ is the learning rate decay coefficient

k is the training round.

A Dropout layer and a batch norm operation are added after each convolutional layer to make the model more general and stop it from overfitting. Dropout randomly throws away neurons, while batch norm makes sure that feature distributions are normal. Also, L2 regularisation with a weight decay coefficient of 1e-4 keeps things simple, which makes training faster, easier to understand, and more stable when deployed.

The concatenation of features from different modalities can be expressed as:

$$\mathbf{F}_{\text{concat}} = \left[\mathbf{F}_{\text{image}}, \mathbf{F}_{\text{audio}}, \mathrm{F}_{\text{text}}\right] \tag{5}$$

In equation (5), $\mathbf{F}_{\text{image}}$, $\mathbf{F}_{\text{audio}}$, and $\mathbf{F}_{\text{text}}$ represent the features extracted from images, speech, and text, respectively. The concatenated feature $\mathbf{F}_{\text{concat}}$ can be used as the input for the next layer.

The training data are separated into training and validation sets in a ratio of 8:2, using a batch size of 64 and with an upper limit of 50 training epochs. Several data augmentation techniques are used, including random horizontal flip, brightness adjustment, and visual data rotation; addition of background noises and speech rate adjustment for auditory data; and synonym replacement and word order shuffling for text data, all designed to increase data diversity and improve model robustness in complex classroom settings. The training routine includes an Early Stopping mechanism to stop training if no improvement in F1 score is observed over a period of 8 consecutive epochs, thus preventing overfitting. Performance metrics like accuracy, precision, recall, and F1 score are recorded and plotted using Tensor board, reflecting stable convergence and good validation performance and without overfitting (see Figure 2).

Figure 2 Accuracy and loss curves during model training (see online version for colours)

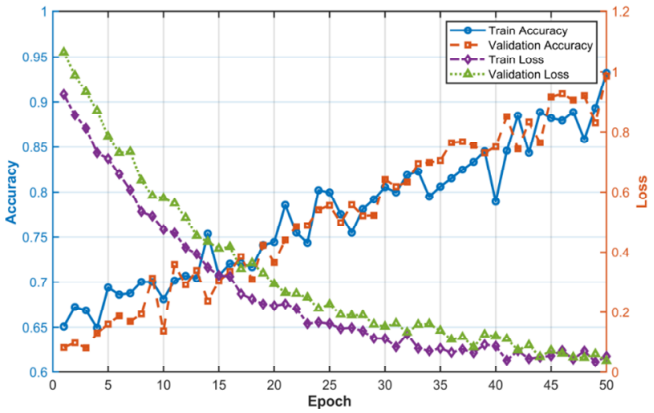


Figure 2 depicts the steady improvement of the model over 50 training steps, where training accuracy improved from 65% to over 93%, and validation accuracy improved from 63% to nearly 94%, thus showing strong generalisation ability. The training loss reduced from 1.0 to 0.05, and the validation loss presented a similar trend, hence confirming stable convergence without overfitting evidence. The steady performance of the model indicates the presence of a viable network structure and optimisation strategy. Besides, comparative experiments, focusing on evaluating feature combiner techniques, show that adding an attention mechanism to dynamically scale image, speech, and text features' weights considerably increases collaborative effectiveness, leading to improved F1-score and better robustness against noise and modality loss.

3.3 *Real-time sentiment analysis and feedback mechanism*

The system uses WebRTC to get live video, audio, and text feeds from students. It then uses features from video frames, MFCC speech features, and text embeddings to do this (Patel et al., 2022). Using a CNN-based architecture with attention mechanisms for feature weighting, these features are processed and sent to the model's input queue for trimodal sentiment classification. Every two seconds, the system sorts emotions into groups such as happy, neutral, and anxious to make the results more accurate. An adaptable threshold model changes emotional response thresholds depending on the type of course, giving dynamic feedback. For instance, 'confused' is turned on in theoretical courses when more than 10% of students say they are. The system gives teachers feedback through a heatmap and students rewards based on how involved they are. It works with learning management systems through an

xAPI interface, which lets you do assessments all the time and get personalised suggestions. The results of the experiment show that the system has a latency of 100 to 200 ms, which allows for real-time feedback while keeping accuracy.

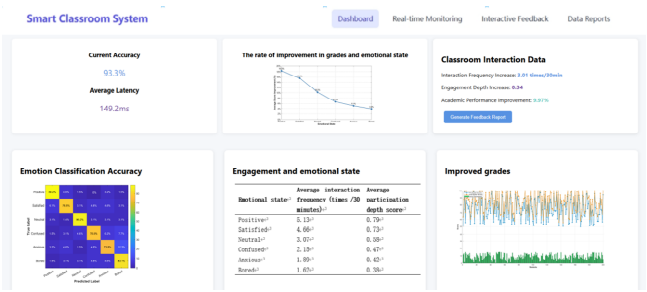
3.4 Lightweight deployment optimisation strategy

This paper suggests a lightweight deployment strategy for the emotion recognition CNN model to make it easier to use on online teaching platforms. The model cuts down on the number of unnecessary convolutional layers and uses 16-bit mixed precision to lower the amount of processing and video memory it needs, making it faster without losing accuracy. The optimised model runs on the RTX 3080 platform, which allows for real-time processing and WebRTC data collection.

3.5 System page display

The website shown in Figure 3 is a visualisation interface for emotion recognition and interactive feedback in classroom environments, utilising CNN. It addresses problems of emotional perception, interaction, and delay in feedback in online learning. The interface consists of four basic elements: a dashboard displaying key indicators, an analysis module combining multimodal data, a feedback module providing real-time cues, and a report module focusing on assessing teaching practices and fine-tuning the system.

Figure 3 System display diagram (see online version for colours)



4 System performance verification and teaching application empirical analysis

4.1 Experimental design

The developed emotion recognition model, implemented under the bottom-up online learning framework, was tested thoroughly through a set of experiments on multimodal data drawn from public emotion recognition databases and class samples. The evaluation focused on five separate tasks: emotion recognition accuracy, measured in terms of precision, recall, and F1 score; system response time; user interaction-emotional correlation tags; instructional effectiveness; and system stability. To ensure the reliability of results, each task was implemented on five different occasions. The collection and processing of the Real Classroom dataset adhered to the ethical guidelines for educational research involving human participants. The study received approval from the Institutional Review Board of Hebei Open University under protocol number HOU-IRB-2024-018. All participating students provided written informed consent before data collection, which explicitly described the types of data recorded including facial video, speech audio, and chat text, the purpose of emotion recognition and feedback generation, the duration of data retention, and the right to withdraw without academic penalty. Facial video data were processed locally on edge computing nodes to prevent transmission of raw images to central servers; only anonymised feature vectors were stored. Speech recordings were transcribed to text, and the original audio files were deleted immediately after feature extraction. All text data were stripped of student identifiers and replaced with random pseudonyms. The consent form also stated that no data would be shared with third parties or used for any purpose beyond the reported research. The experiments provided an extensive foundation for testing the performance of the model across different phases, such as training, deployment, and operational phases, ensuring both stability and accuracy in realistic learning environments.

Table 2 Overview of experimental datasets

Dataset name	*Data type*	*Sample size*	*Feature dimensions*	*Usage*
AffectNet	Image	11,643	48 × 48 grayscale images	Pre-training, validation
EmoReact	Multimodal (image + audio + text)	1,436	Image + 39-dim MFCC + 300-dim text embeddings	Pre-training, validation
Real classroom	Multimodal (image + audio + text)	3,254texts, 22,570 images, 18h audio	Same structure, unified pre-processing	Model testing, feedback eval

Table 2 shows the three emotion recognition datasets that were used. They are AffectNet and EmoReact, which are both public datasets, and data collected by the authors in university classrooms. These datasets include images, speech, and text, and the sample sizes range from 1,436 to tens of thousands. The real classroom dataset was curated from 12 undergraduate courses across six academic disciplines taught online during the spring semester of 2024. Video frames were extracted at 2 frames per second, and each frame was aligned with the corresponding speech segment and text chat log from the same 2-second window. Manual annotation of emotional states for the real classroom dataset was performed by three trained educational psychology researchers using a six-class labelling scheme happy, satisfied, neutral, confused, anxious, bored. Inter-annotator agreement measured by Fleiss' kappa reached 0.86. For all three datasets, the

training set comprised 64% of samples, the validation set 16%, and the test set 20%. The 8:2 training-validation split reported in Section 3.2 referred only to the initial model development phase; the final performance evaluation in Section 4 used the 64-16-20 split to ensure independent testing. The same random seed was fixed across all experiments to guarantee reproducibility. The experimental protocol for baseline comparison fixed the batch size to 64, the number of epochs to 50, and the early stopping patience to 8 epochs for all models. Each baseline model was trained three times with different random initialisations, and the reported metrics are the averages across the three runs. This makes sure that the models are thoroughly tested and can be used in the real world.

4.2 Comprehensive analysis of emotion recognition

Four standard classification metrics accuracy, precision, recall, and F1 score evaluate the CNN-based multimodal emotion recognition model on the six-category emotion classification task. The unimodal image baseline model ResNet18 was initialised with ImageNet pretrained weights and fine-tuned on the facial expression data using a learning rate of 0.0001 and the Adam optimiser for 30 epochs. The input to ResNet18 consisted of 48 × 48 grayscale face images identical to those used in the proposed CNN model. The text baseline model LSTM + MLP used a 300-dimensional Word2Vec embedding layer followed by a bidirectional LSTM with 128 hidden units per direction and a two-layer MLP with 64 and 32 neurons. The LSTM + MLP model was trained with a learning rate of 0.001 and dropout of 0.5 applied after each LSTM layer. Both baseline models used the same cross-entropy loss function and the same training-validation-test split as the proposed CNN model. No data augmentation was applied to the baseline models to isolate the effect of the multimodal attention mechanism. These indicators are constructed based on the confusion matrix between the model output results and the true labels, and are evaluated on the validation set.

Accuracy is used to measure the proportion of correct predictions of the model as a whole, and its mathematical definition is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

In equation (6)

TP represents the number of samples correctly classified as their true categories

TN represents the number of samples correctly judged as non-target categories

FP represents the number of samples that are mistakenly predicted as other categories

FN represents the number of samples that are mistakenly predicted as other categories. The accuracy rate reflects the classification accuracy of the model on the overall samples.

The accuracy rate indicates how many samples predicted by the model as a certain category belong to that category, and is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Precision measures the credibility of the model in predicting positive examples and is particularly useful in scenarios that are sensitive to false positives. In emotion recognition, high precision means that the model is less likely to misjudge neutral or negative emotions as positive emotions.

Recall measures the model's coverage of real samples, that is, how many samples that belong to a certain category are correctly identified by the model. The formula is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

A high recall rate means that the probability of the model missing a judgment is low. In classroom teaching emotion monitoring, the importance of recall rate lies in the ability to identify as many emotional states as possible, especially emotional categories such as 'confused' and 'anxious' that have reference value for teaching intervention.

The F1 score is the harmonic mean of precision and recall, which is used to comprehensively evaluate the model's ability to balance accuracy and coverage. Its calculation formula is as follows:

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision \times Recall} \tag{9}$$

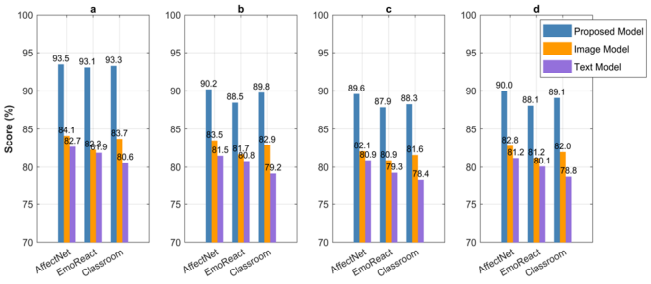
The F1 value is particularly important when there is a deviation between precision and recall, and can effectively prevent a single indicator from masking the model's weaknesses in another dimension.

At the same time, to compare the performance of this model, two types of traditional baseline models are constructed as references: one is a unimodal image model that uses ResNet18 to extract image features Helaly et al. (2023), and the other is a text classification model of long short-term memory + multi-layer perceptron (LSTM + MLP) Phiphitphatphaisit et al. (2021). The overall classification performance of the model on the three datasets is shown in Figure 4, which further demonstrates the stability and advantages of the proposed method in multiple scenarios.

Figure 4 demonstrates comparative performance for four measures over three models for recognising emotions based on AffectNet, EmoReact, and university classroom data. The multimodal CNN outperforms both baseline image-based model (ResNet18) and text-based model (LSTM + MLP) on all data and performance measures, having a mean accuracy rate of 93.3%. For the Classroom data specifically, the F1-score on the CNN is measured as 89.1%, outpacing the 82.0% on the image-based model and the 78.8% on the text-based model. The precision and recall measures on the CNN from the EmoReact data are 88.5% and 87.9%,

respectively, showing improved abilities to discern and recognise states of emotion.

Figure 4 Comparison of different emotion recognition models, (a) accuracy comparison (b) precision comparison (c) recall comparison (d) f1-score comparison (see online version for colours)



A direct comparison with recent state of the art models published in 2023 or later is presented in Table 3. The selected models include the double channel attention network of Yan et al. (2023) which combined BERT, CNN, and LSTM, the real-time facial expression network of Savchenko et al. (2022), and the spatio temporal 3D CNN of Khanna et al. (2024). All three models were re implemented and evaluated on the real classroom dataset under identical conditions using the 64 16 20 split and the same hardware platform.

Table 3 Comparison with recent state of the art models on the real classroom dataset

Model	*Year*	*Accuracy (%)*	*F1 score*	*Inference latency (ms)*	*Reference*
Double channel attention network	2023	88.7	0.851	312.4	Yan et al. (2023)
Real-time facial expression network	2022	85.2	0.823	98.6	Savchenko et al. (2022)
Spatio temporal 3D CNN	2024	90.1	0.869	245.7	Khanna et al. (2024)
Proposed CNN multimodal model	–	93.3	0.891	149.2	–

The proposed model outperformed the double channel attention network of Yan et al. (2023) by 4.6 percentage points in accuracy and 0.04 in F1 score while achieving substantially lower latency 149.2 ms versus 312.4 ms. Compared to the spatio temporal 3D CNN of Khanna et al. (2024), the proposed model showed a 3.2 percentage point accuracy advantage and a 0.022 higher F1 score, with a 96.5 ms lower latency. The real-time facial expression network of Savchenko et al. (2022) achieved the lowest latency 98.6 ms but at a significant accuracy penalty of 8.1 percentage points. These comparisons confirm that the proposed model strikes a superior balance between recognition accuracy and real-time responsiveness relative to the most recent alternatives.

4.3 System response time

To verify the system's response performance under various network environments, particularly its efficiency after lightweight deployment, this paper statistically analyses the end-to-end delay of the system under three typical network conditions. The response time includes the entire process of data collection, feature pre-processing, model reasoning, and feedback generation. The system response time is defined as the total time from the completion of the collection of student emotional input to the completion of the generation of emotional feedback content and its push to the client, recorded as the end-to-end delay. T_{total} and its structure can be decomposed into:

$$T_{total} = T_{capture} + T_{pre} + T_{infer} + T_{feedback} + T_{net} \tag{10}$$

In equation (10)

$T_{capture}$ represents the data collection time

T_{pre} is the multimodal feature pre-processing time

T_{infer} is the model inference time

$T_{feedback}$ is the logic processing time required to generate feedback content

T_{net} is the front-end and back-end communication delay.

In actual testing, the system log and timer are used to record the time consumption of each stage, and the following indicators are used for quantification:

- Average response time:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} T_{total}(i) \tag{11}$$

- Response time standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(T_{total}(i) - \mu\right)^2} \tag{12}$$

- Maximum delay:

$$T_{max} = \max_{i\in[1,\,N]} T_{total}(i) \tag{13}$$

The complete characterisation of response efficacy, stability fluctuations, and performance capability of the system under heavy loads is done using three measures: μ, σ, and T_{max}. These measures are the delay distribution and the capability of the platform to handle large concurrency or network instability. Running on a server with an RTX 3080 GPU, the system simulates three types of networks: high-speed LAN, medium-speed WAN, and low-speed cellular networks, using a total of 10 concurrent clients and 200 valid responses per iteration.

To further verify the stability of the system response time, the 95% confidence interval under the normal distribution of the delay is calculated:

$$CI_{95\%} = \mu \pm 1.96 \cdot \frac{\sigma}{\sqrt{N}} \tag{14}$$

The statistical results of the system's response time in three typical network environments are shown in Figure 5, reflecting its real-time performance under different bandwidth conditions.

Figure 5 Comparison of system response time under different network environments (see online version for colours)

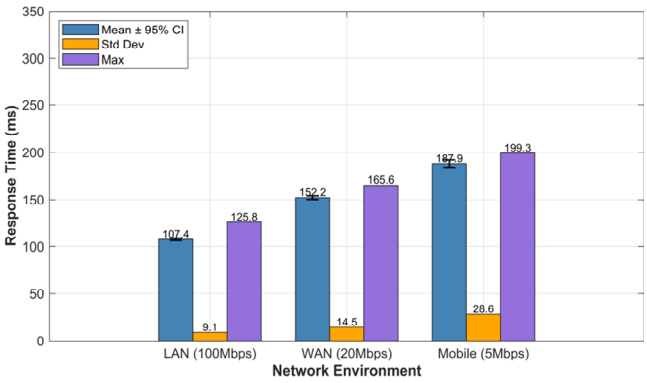


Figure 5 shows how long it takes the emotion recognition interactive feedback system to respond in three different network settings: LAN, WAN, and mobile. An additional evaluation was conducted on lower-specification hardware to assess deployment feasibility in resource-constrained rural educational settings. The inference engine was executed on a laptop with an Intel Core i5-1135G7 processor integrated Iris Xe graphics and 8 GB RAM without a discrete GPU. Under the same WAN network condition average bandwidth 15 Mbps, round-trip time 45 ms, the average response time on the low-spec hardware increased from 152.2 ms to 384.7 ms, while the maximum delay reached 512.3 ms. The error rate under 300 concurrent simulated users rose from 0.41% on the RTX 3080 server to 2.87% on the low-spec hardware. For rural settings with cellular network speeds below 5 Mbps, the system exceeded the 500 ms threshold for 34% of requests, rendering real-time feedback unreliable. The lightweight deployment optimisation described in Section 3.4 reduced GPU memory consumption by 47% on the RTX 3080, but the optimisations assumed a CUDA-compatible GPU. For purely CPU-based deployments, model quantisation to 8-bit integers reduced the inference time to 267.4 ms on the Core i5 platform but caused a 3.1 percentage point drop in accuracy. The results show that the system works best in the LAN, with an average response time of 107.4 ms, a standard deviation of 9.1 ms, and a maximum of 125.8 ms. The average in WAN is 152.2 ms, but the range is bigger. In mobile networks, the average goes up to 187.9 ms, and the standard deviation is 28.6 ms higher. The average response time across all environments is 149.2 ms, which shows how well the system works and how quickly it can give feedback.

4.4 User engagement

This section mainly evaluates the impact of the proposed emotion recognition system on the user participation behaviour of university online classes based on behavioural data. Three complete classroom teaching units after the system deployment are selected as samples. Based on the system front-end and back-end logs and platform behaviour data, the user interaction frequency and content complexity are extracted as the measurement indicators of classroom interaction frequency and depth, recorded as F_i and D_i, respectively.

The classroom interaction frequency F_i is defined as the total number of students' active participation behaviours within a unit course duration, including sending barrages, voice questions, clicking to answer questions, submitting feedback, etc., and is calculated as:

$$F_i = \frac{N_i^{active}}{T_i} \tag{15}$$

In equation (15), N_i^{active} is the number of effective interactions of the i^{th} student in the course unit, and T_i is the course duration. To further characterise the quality of student participation, the depth of participation D_i is defined as the average information complexity score of the content submitted by the student in a class. Specifically, the scoring model $D_i = f$ (len, key, sent) is composed of content length, number of keywords, and emotional colour. The score is obtained by a weighted superposition of three standardised sub-indicators, as follows:

$$D_i = \alpha \cdot \frac{len_i}{\max(len)} + \beta \cdot \frac{key_i}{\max(key)} + \gamma \cdot \frac{sent_i}{\max(sent)} \tag{16}$$

In equation (16)

len_i represents the total number of words in the speech of the i^{th} student

key_i is the number of teaching-related keywords in the speech

sent_i is the positive emotion probability score given by the natural language processing model, and the weight parameter is set to $\alpha = 0.4$, $\beta = 0.4$, $\gamma = 0.2$ to emphasise the content relevance.

To further investigate the impact of different emotional states on students' classroom interactive behaviours, this paper calculated the average interaction frequency and participation depth scores of students under six types of emotional labels. The results are shown in Table 4:

Table 4 demonstrates the correlation between the emotional states of students and class activity engagement, measured as interaction rate and depth of interaction. Students who are positively engaged have a 5.13 interaction rate/30 minutes and a depth level of 0.79, and those who are satisfied have a 4.66 interaction rate and 0.73 depth.

Negative feelings, on the other hand, dramatically reduce interaction levels, and those who are bored show the least measures of 1.62 interaction rate and 0.38 depth. The emotional system additionally increases interaction rate by 3.01 and depth of interaction by 0.34, consequently revealing the positive effect that emotional states have on both interaction and cognitive engagement.

Table 4 Statistics of student participation and emotional state

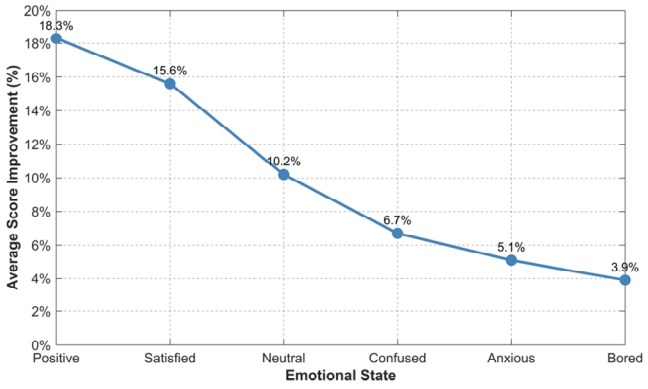
Emotional state	*Average interaction frequency (times/30 minutes)*	*Average participation depth score*
Positive	5.13	0.79
Satisfied	4.66	0.73
Neutral	3.07	0.58
Confused	2.15	0.47
Anxious	1.89	0.42
Bored	1.62	0.38

4.5 Improvement of teaching effectiveness

This study investigates an interactive feedback system created for affective identification as an intervention that impacts instruction's effectiveness, using pre and post-comparison research methodology. It compares performance score variations among students in varying emotional conditions using two sequences of standardised tests for three online undergraduate courses. The score before the test is S^i_{pre}, and the score after the test is S^i_{post}. The improvement rate of the i^{th} student's score is defined as:

$$\Delta S_i = \frac{S^i_{post} - S^i_{pre}}{S^i_{pre} + \in} \tag{17}$$

Figure 6 The relationship between students' performance improvement rate and emotional state (see online version for colours)



In equation (17), $\in$ is a very small constant to avoid division by zero errors. After completing data collection, students are divided into six emotional groups based on the main emotional labels recorded in the system, and the average improvement rate of each group is calculated.

Figure 6 makes evident a positive correlation between the emotional states and learning outcomes of students. The positive and satisfied emotional state groups recorded the highest-grade improvements, at 18.3% and 15.6%, respectively. Students in negative emotional states, confused, anxious, and bored, had less marked increases in their learning outcomes, at 6.7%, 5.1%, and 3.9%, respectively. The overall average grade improvement across all states was found to be 9.97%. The system tracked changes in emotional states, prompting adjustments in pedagogical approaches that reinforced retention during positive emotional states and triggered targeted interventions during negative emotional states. Interestingly, the 'confusion alert' proved useful for confused students but was ineffective for students suffering from anxiety. The impact of such a system was evident across different academic fields.

Figure 7 Comparison of students' emotions and performance improvement in liberal arts and science courses (see online version for colours)

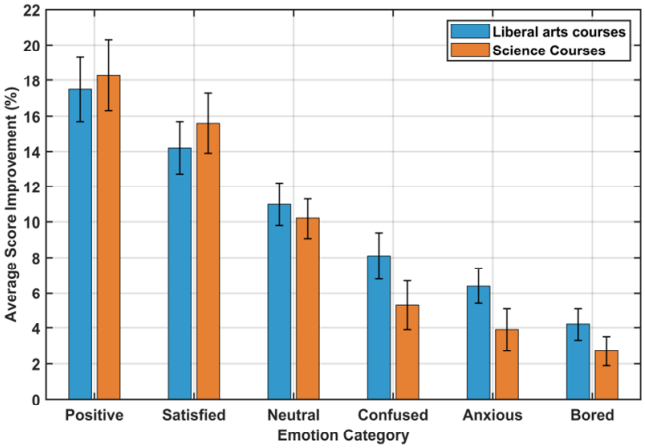


Figure 7 shows comparative average improvement rates for students taking liberal arts and science courses, subject to six different emotional states. Students having positive emotions (i.e., positive and satisfied) recorded the highest improvements, at 18.3% and 17.5%, respectively, for science students. Contrarily, negative emotional states, such as confused and bored, led to lower grade improvements, with science students recording lower improvements compared to liberal arts students. Having positive emotions strongly enhances learning, while negative emotions hinder academic progress, especially in science courses.

4.6 System stability and scalability

This part tests the system's stability and ability to handle many users at once in a high-load simulation. It looks at the error rate, crash frequency, and number of users who can use the system at the same time. The experiment uses an NVIDIA RTX 3080 GPU, 64GB of memory, and a 6-hour client-server test with log-based analysis.

The system error rate (error rate) is defined as the ratio of the number of task execution failures per unit time to the total number of tasks:

$$Error\ Rate = \frac{N_{fail}}{N_{total}} \tag{18}$$

In equation (18), N_{fail} is the number of failed requests detected by the system, and N_{total} is the total number of requests.

Figure 8 System stability and scalability test results (see online version for colours)

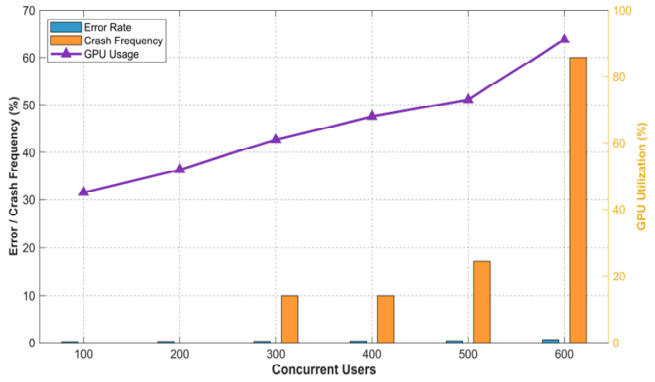
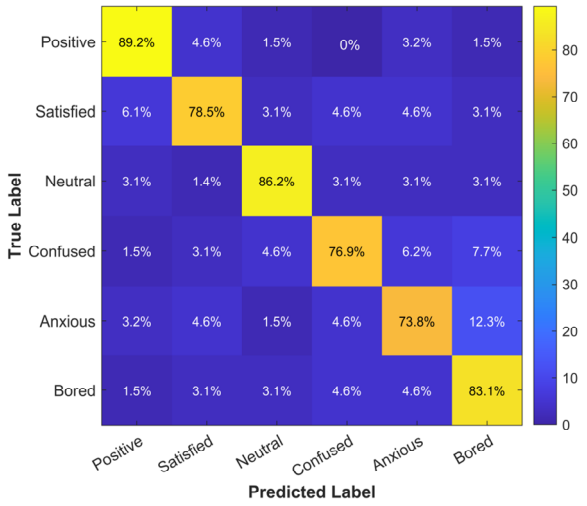


Figure 8 demonstrates the scalability and robustness of the emotion recognition system under various user loads, from 100 to 600 concurrent users. Even though the error rate showed modest growth, increasing from 0.25% to 0.65%, the rate of crashes shot up significantly, peaking at 60% at 600 users. Utilisation of the GPU showed almost linear growth from 45% to 91%, highlighting growing resource demands. The low error rate retained even under considerable loads can be attributed to the CNN model's steadiness, convergence of parameters, and utilisation of regularisation methods. Despite processing limitations, the performance of the system vindicates scaling for online learning environments, medium to large in size.

Figure 9 Confusion matrix for emotion classification (see online version for colours)



5 Result analysis

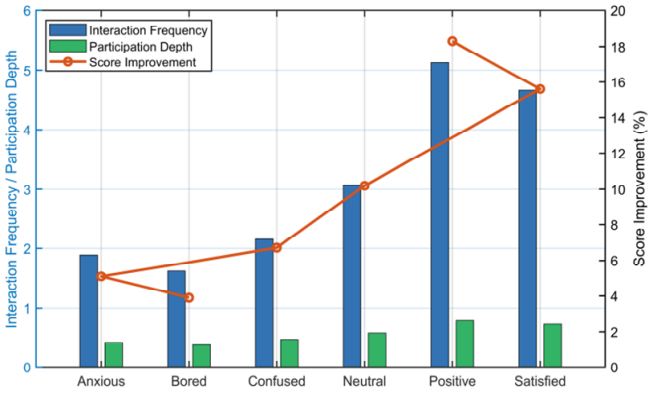
5.1 Comparative analysis of model recognition performance

The proposed CNN-based emotion recognition model achieves better classification results than the standard baselines ResNet18, LSTM, and MLP. Channel attention enhances multimodal fusion by properly regulating image, text, and speech signals, thus improving accuracy in intricate emotional scenarios. By highlighting key features through a dynamic attention approach, the model attains a strong discriminative capability. With a classification accuracy of 89.2% for positive instances, the model is quite efficient in emotional discrimination; however, it faces difficulties in precise classification of subtle emotions like anxiety and satisfaction (see Figure 9).

5.2 Analysis of the causal relationship between emotion recognition and learning behaviour

The current research examined the role of emotional states in academic achievement and classroom interaction, focusing on interaction rate frequency, student involvement, and academic progress. It identified six unique emotional states and found that positive emotions correlate with higher interaction rates (5.13 events/p30 minutes), higher participation (0.79), and increased academic performance (18.3%). Conversely, states of boredom correlate with the lowest performance scores. Using CNNs, the suggested emotion classification system allows for real-time adjustments to teaching strategies, enhancing engagement and student performance. Positive emotional states promote deeper engagement, while confusion requires re-explanation (see Figure 10).

Figure 10 Causal path diagram of emotional state, interaction, and performance (see online version for colours)

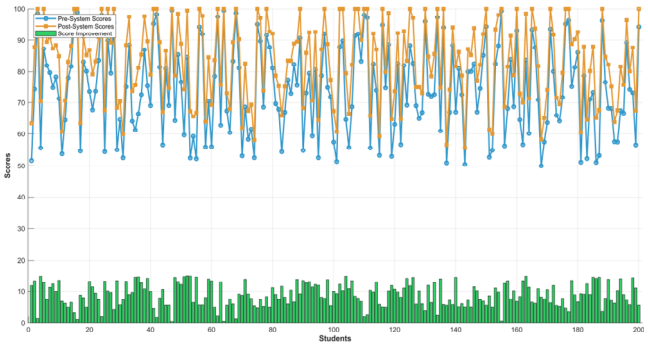


5.3 Analysis of system comprehensive effectiveness and teaching application effect

This study evaluates an emotional state recognition model for virtual classrooms, highlighting functional effectiveness and instructional benefits. The model can respond in real-time within 100 to 200 milliseconds, with GPU utilisation ranging from 45% to 91% as user numbers run

from 100 to 600, and always maintains an error level below 0.65%. In learning contexts, student activity increases for those with 'satisfied' and 'neutral' emotional states but decreases for those with 'bored' and 'confused' feelings. Academic performance undergoes an enhancement of 9.97%, specifically by those students demonstrating stronger emotional regulation. Both engagement and educational outcomes respond positively to the model's effect (see Figures 8 and 11).

Figure 11 Feedback from students before and after their grades improved (see online version for colours)



6 Conclusions

This work introduces a multimodal system for interactive feedback and emotion estimation that focuses on leveraging CNNs to promote students' engagement and academic performance in online higher education. The envisioned architecture combines information from visual images, audio signals, and textual data to support real-time emotion estimation with impressive accuracy (93.3%) and competitive latency (149.2 ms). Through adaptive adjustment of pedagogical strategies according to students' emotional responses, the system stimulates higher class participation, contributing indirectly to increased interaction levels, improved learning achievement, and better learning experiences. Experimental results show significant student engagement and learning performance improvements, especially in the case of individuals with positive emotional orientations. Additionally, the system demonstrates adequate stability and scalability under various network environments, suggesting its real-world applicability in a wide range of online classrooms. By offering customised feedback to educators and learners, the system instills a more responsive and dynamic learning environment, thus reducing the gap between emotional sensitivity and teaching adaptability. User feedback collected through post-course surveys revealed three major gaps. First, 41% of students reported that the system occasionally misinterpreted concentration as anxiety, leading to unnecessary intervention notifications. Second, instructors noted that the confusion heatmap updated every two seconds sometimes flickered when multiple students switched between confused and neutral states within short intervals, causing visual fatigue. Third, students with higher than average speech rates produced MFCC features that shifted the emotion classification toward excited regardless of their actual emotional state, indicating a feature-level bias. Addressing these gaps requires three follow-up research directions: designing a temporal smoothing filter that requires emotion state changes to persist for at least four seconds before triggering alerts, developing a speaker adaptation layer that normalises MFCC features by the individual's baseline speech rate, and conducting a longitudinal study to measure whether the reduction of false anxiety alarms improves student trust in the system.

Several limitations and potential biases require acknowledgment. The emotion recognition model was trained primarily on facial expression data from East Asian university students, which may reduce accuracy when applied to learners from other cultural backgrounds where emotional display norms differ. Students who kept webcams off for extended periods reduced the availability of visual modality; the model fell back to audio and text features, but recognition accuracy under such conditions dropped by 8.2 percentage points based on a post-hoc analysis of sessions with more than 50% camera-off time. The model also exhibited a systematic bias toward classifying neutral expressions as positive when the speaker's native language was not English, likely due to prosodic differences in the speech channel. These biases and performance degradations should be addressed in future work through cross-cultural dataset expansion and adaptive modality weighting that explicitly accounts for missing video input.

Declarations

All authors declare that they have no conflicts of interest.

References

Baqach, A. and Battou, A. (2024) 'A new sentiment analysis model to classify students' reviews on MOOCs', **Education and Information Technologies**, Vol. 29, No. 13, pp.16813–16840. https://doi.org/10.1007/s10639-024-12526-0.

Chacón-Cuberos, R., Olmedo-Moreno, E.M., Lara-Sánchez, A.J., Zurita-Ortega, F. and Castro-Sánchez, M. (2021) 'Basic psychological needs, emotional regulation and academic stress in university students: a structural model according to branch of knowledge', *Studies in Higher Education*, Vol. 46, No. 7, pp.1421–1435, https://doi.org/10.1080/03075079.2019.1686610.

Chang, J.Y-F., Wang, L-H., Lin, T-C., Cheng, F-C. and Chiang, C-P. (2021) 'Comparison of learning effectiveness between physical classroom and online learning for dental education during the COVID-19 pandemic', **Journal of Dental Sciences**, Vol. 16, No. 4, pp.1281–1289, https://doi.org/10.1016/j.jds.2021.07.016.

Gupta, S., Kumar, P. and Tekchandani, R.K. (2023) 'Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models', *Multimedia Tools and Applications*, Vol. 82, No. 8, pp.11365–11394, https://doi.org/10.1007/s11042-022-13558-9.

He, X., Deng, Y., Fang, L. and Peng, Q. (2021) 'Multi-modal retinal image classification with modality-specific attention network', *IEEE Transactions on Medical Imaging*, Vol. 40, No. 6, pp.1591–1602, https://doi.org/10.1109/TMI.2021.3059956.

Helaly, R., Messaoud, S., Bouaafia, S., Hajjaji, M.A. and Mtibaa, A. (2023) 'DTL-I-ResNet18: facial emotion recognition based on deep transfer learning and improved ResNet18', *Signal, Image and Video Processing*, Vol. 17, No. 6, pp.2731–2744, https://doi.org/10.1007/s11760-023-02490-6.

Hosseini, S., Yamaghani, M.R. and Poorzaker Arabani, S. (2024) 'A review of the methods of recognition multimodal emotions in sound, image and text', *International Journal of Applied Operational Research*, Vol. 12, No. 1, pp.29–41, http://dx.doi.org/10.71885/ijorlu-2024-1-657.

Jiang, D., Liu, H., Wei, R. and Tu, G. (2023) 'CSAT-FTCN: a fuzzy-oriented model with contextual self-attention network for multimodal emotion recognition', *Cognitive Computation*, Vol. 15, No. 3, pp.1082–1091, https://doi.org/10.1007/s12559-023-10119-6.

Johnson, S.J., Murty, M.R. and Navakanth, I. (2024) 'A detailed review on word embedding techniques with emphasis on word2vec', *Multimedia Tools and Applications*, Vol. 83, No. 13, pp.37979–38007, https://doi.org/10.1007/s11042-023-17007-z.

Khan, N., Singh, A. and Agrawal, R. (2023) 'Enhancing feature extraction technique through spatial deep learning model for facial emotion detection', *Annals of Emerging Technologies in Computing*, Vol. 7, No. 2, pp.9–22, https://doi.org/10.33166/AETiC.2023.02.002.

Khan, S.S., Sengupta, D., Ghosh, A. and Chaudhuri, A. (2024) 'MTCNN++: a CNN-based face detection algorithm } inspired by MTCNN', *The Visual Computer*, Vol. 40, No. 2, pp.899–917, https://doi.org/10.1007/s00371-023-02822-0.

Khanna, D., Jindal, N., Rana, P.S. and Singh, H. (2024) 'Enhanced spatio-temporal 3D CNN for facial expression classification in videos', *Multimedia Tools and Applications*, Vol. 83, No. 4, pp.9911–9928, https://doi.org/10.1007/s11042-023-16066-6.

Kuchuk, H. and Malokhvii, E. (2024) 'Integration of IoT with cloud, fog, and edge computing: a review', https://doi.org/10.20998/2522-9052.2024.2.08.

Liu, J., Yan, Z., Chen, S., Sun, X. and Luo, B. (2023) 'Channel attention TextCNN with feature word extraction for Chinese sentiment analysis', *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 22, No. 4, pp.1–23, https://doi.org/10.1145/3571716.

Liu, W., Qiu, J-L., Zheng, W-L. and Lu, B-L. (2021) 'Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition', *IEEE Transactions on Cognitive and Developmental Systems*, Vol. 14, No. 2, pp.715–729, https://doi.org/10.1109/TCDS.2021.3071170.

Mondal, A.S., Zhu, Y., Bhagat, K.K. and Giacaman, N. (2024) 'Analysing user reviews of interactive educational apps: a sentiment analysis approach', *Interactive Learning Environments*, Vol. 32, No. 1, pp.355–372, https://doi.org/10.1080/10494820.2022.2086578.

Patel, V., Li, C.H., Rye, V., Liu, C-S.J., Lerner, A., Acharya, J. and Rajamohan, A.G. (2022) 'A comparison of webrtc and conventional videoconferencing for synchronized remote medical image presentation', *Journal of Digital Imaging*, Vol. 35, No. 1, pp.68–76, https://doi.org/10.1007/s10278-021-00544-0.

Phiphitphatphaisit, S. and Surinta, O. (2021) 'Deep feature extraction technique based on Conv1D and LSTM network for food image recognition', *Engineering and Applied Science Research*, Vol. 48, No. 5, pp.581–592, https://doi.org/10.14456/easr.2021.6.

Savchenko, A.V., Savchenko, L.V. and Makarov, I. (2022) 'Classifying emotions and engagement in online learning based on a single facial expression recognition neural network', *IEEE Transactions on Affective Computing*, Vol. 13, No. 4, pp.2132–2143, https://doi.org/10.1109/TAFFC.2022.3188390.

Singh, M.K. (2024) 'Multimedia application for forensic automatic speaker recognition from disguised voices using MFCC feature extraction and classification techniques', *Multimedia Tools and Applications*, Vol. 83, No. 32, pp.77327–77345, https://doi.org/10.1007/s11042-024-18602-4.

Sümer, Ö., Goldberg, P., D'Mello, S., Gerjets, P., Trautwein, U. and Kasneci, E. (2021) 'Multimodal engagement analysis from facial videos in the classroom', *IEEE Transactions on Affective Computing*, Vol. 14, No. 2, pp.1012–1027, https://doi.org/10.1109/TAFFC.2021.3127692.

Toscu, S. (2023) 'Exploring classroom interaction in online education', *Education and Information Technologies*, Vol. 28, No. 9, pp.11517–11543, https://doi.org/10.1007/s10639-023-11622-x.

Uddin, M.A., Pathan, R.K., Hossain, M.S. and Biswas, M. (2022) 'Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN', *Journal of Information and Telecommunication*, Vol. 6, No. 1, pp.27–42, https://doi.org/10.1080/24751839.2021.1983318.

Wu, Y-X., Du, K., Wang, X-J. and Min, F. (2024) 'Misclassification-guided loss under the weighted cross-entropy loss framework', *Knowledge and Information Systems*, Vol. 66, No. 8, pp.4685–4720, https://doi.org/10.1007/s10115-024-02123-5.

Yan, C., Liu, J., Liu, W. and Liu, X. (2023) 'Sentiment analysis and topic mining using a novel deep attention-based parallel dual-channel model for online course reviews', *Cognitive Computation*, Vol. 15, No. 1, pp.304–322, https://doi.org/10.1007/s12559-022-10083-7.