engineering proceedings

MDPI

Proceeding Paper

Detection of Students' Emotions in an Online Learning Environment Using a CNN-LSTM Model †

Bilkisu Muhammad Bashir * and Hadiza Ali Umar

Department of Computer Science, Bayero University, Kano PMB 3011, Nigeria
* Correspondence: bilkisubashirbilhaq@gmail.com
† Presented at the 5th International Electronic Conference on Applied Sciences, 4–6 December 2024; https://sciforum.net/event/ASEC2024.

Abstract

Emotion recognition through facial expressions is crucial in fields like healthcare, entertainment, and education, offering insights into user experiences. In online learning, traditional methods fail to capture students' emotions effectively. This research introduces a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) model to recognize learning emotions (interest, boredom, and confusion) during online lectures. A custom dataset was constructed by mapping action units from FER2013, CK+48, and JAFFE datasets into three learning-related categories. Images were preprocessed (grayscale conversion, resizing, normalization) and divided into training and testing sets. The CNN layers extract spatial facial features, while the LSTM layers capture temporal dependencies across video frames. Evaluation metrics included accuracy, precision, recall, and F1-score. The model achieved 98.0% accuracy, 97% precision, 98% recall, and 98% F1-score, surpassing existing CNN-only methods. This advancement enhances online learning by enabling personalized support and has applications in education, psychology, and human–computer interaction, contributing to affective computing development.

Keywords: facial emotion recognition; convolutional neural network; long short-term memory; learners' emotion and action units

check for updates

Academic Editor: Nunzio Cennamo

Published: 2 December 2025

Citation: Bashir, B.M.; Umar, H.A. Detection of Students' Emotions in an Online Learning Environment Using a CNN-LSTM Model. *Eng. Proc.* **2025**, *87*, 116. https://doi.org/10.3390/engproc2025087116

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Emotions play a crucial role in human communication, as facial expressions and gestures provide nonverbal cues essential for interpersonal relationships. Facial emotion recognition (FER) has emerged as a powerful tool for identifying emotions in real-time, mimicking human coding skills and aiding in various fields such as medicine, computer vision, and image recognition [1]. Emotion recognition primarily relies on facial expressions and speech, facilitating cross-cultural communication and enabling computers to respond appropriately to users' needs [1]. While humans naturally recognize emotions, computers require advanced algorithms to achieve similar accuracy. FER finds applications in entertainment, social media, healthcare, criminal justice, and education, where it helps analyze user reactions and improve interaction [2]. Recognizing emotions dynamically enhances communication by accurately interpreting individual expressions, fostering deeper connections between humans and technology.

In online learning, emotions significantly impact engagement and learning outcomes. Researchers have explored various modalities to enhance electronic learning (e-Learning), leveraging IoT devices such as webcams and fitness bands for real-time interaction. While

online education has grown rapidly, concerns remain regarding effective communication and faculty skepticism. Studies suggest online courses can match traditional learning in outcomes, yet emotional engagement remains a challenge [3]. This study addresses these concerns by proposing a CNN-LSTM-based approach tailored to e-learning contexts. By leveraging FER for dynamic engagement monitoring, the framework aims to bridge the emotional gap in online education and foster a more supportive and adaptive digital learning environment.

2. Related Works

Several studies have explored the application of deep learning in emotion recognition. For example [4], researchers employed CNNs for emotion recognition and reported 92.5% accuracy on a custom dataset. In another work [3], researchers investigated student emotion recognition during online conferences, achieving 88.7% accuracy on FER2013. Similarly [2], other researchers implemented CNN-based models for facial expression recognition on the CK+ dataset and obtained 88.2% accuracy. While studies argue that online courses can match traditional face-to-face courses in learning outcomes, concerns persist regarding effective communication and faculty skepticism [5]. More recently [1], researchers applied an attentional CNN to FER2013, achieving state-of-the-art performance with 99% accuracy. While these studies highlight strong results, most focus on general emotion categories and do not fully adapt to online learning contexts. To address this gap, our work introduces a custom dataset tailored to e-learning and proposes a hybrid model combining CNN and LSTM for sequential emotion recognition. Table 1 shows the comparative review of related work and the proposed methodology.

Table 1. Comparative Table of Related Works and Our Proposal.

Author	Method	Dataset	Accuracy	Notes
[2]	CNN	CK+	88.2%	Facial expressions
[3]	-	FER2013	88.7%	Online e-learning
[4]	CNN	Custom	92.5%	General FER
[5]	Attentional CNN	FER2013	99.0%	General FER
Our Work (2024)	CNN-LSTM	Custom	98.0%	Tailored to e-learning

3. Methodology

The proposed methodology employs a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) deep learning model for facial emotion recognition in online learning environments. This approach is chosen due to its high accuracy, fast computation, and minimal data requirements. The process begins with image acquisition, where a learner's face is detected and preprocessed through resizing, grayscale conversion, and normalization. Data preprocessing plays a crucial role in improving model performance by cleaning noise, handling missing values, and standardizing features. After preprocessing, the dataset is split into training and testing sets to evaluate the model's effectiveness. The CNN is used for spatial feature extraction, identifying key facial patterns, while LSTM captures temporal dependencies, helping the model recognize emotional transitions over time, as shown in Figure 1 below. This combination enhances the model's ability to detect emotions such as boredom, confusion, and interest in an online educational setting.

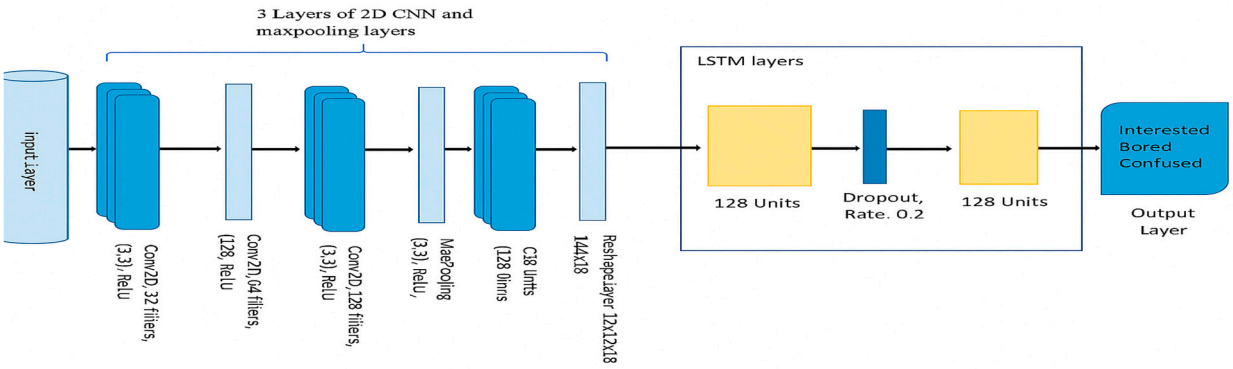


Figure 1. Proposed Methodology.

The CNN-LSTM architecture is particularly beneficial for analyzing sequential image data, such as video frames, where understanding the temporal order is essential. CNN layers extract spatial features by applying convolutional filters, followed by pooling layers to reduce dimensionality while preserving essential details. The output from the last CNN layer is reshaped into a sequential format compatible with LSTM layers, which process the data as a time series to capture changes in facial expressions. Dropout layers are incorporated to prevent overfitting by randomly deactivating a portion of connections during training. The final LSTM layer produces a sequential representation, which is then passed through a dense layer with three output units representing emotion classes: interested, bored, and confused. A SoftMax activation function is applied to generate probabilities for each emotion class, enabling the model to predict emotions based on the highest probability score.

3.1. Dataset

The dataset was constructed by combining samples from FER2013, CK+, and JAFFE. Facial Action Units were mapped into three learning-related categories: interest, boredom, and confusion. Images were resized to 48 × 48 grayscale and normalized. The final dataset contained over 6000 samples, divided into 80% training and 20% testing sets.

3.2. Neural Network Architecture

The proposed CNN-LSTM architecture integrates convolutional layers for spatial feature extraction with recurrent LSTM layers for temporal sequence modeling. Specifically, three convolutional layers were employed with 32, 64, and 128 filters, respectively, each using a kernel size of 3 × 3 and ReLU activation. Each convolutional layer was followed by a max-pooling layer (2 × 2) to progressively reduce dimensionality while retaining key features. The final convolutional output was flattened and reshaped into a sequential format compatible with recurrent processing.

Two stacked LSTM layers, each with 128 hidden units, were then used to capture temporal dependencies across sequential facial frames. Dropout layers (rate = 0.5) were introduced after each LSTM layer to mitigate overfitting by randomly deactivating neurons during training. Finally, a fully connected dense layer with three output units was applied, corresponding to the target emotion categories: interested, bored, and confused. A SoftMax activation function generated normalized probability distributions over the three classes, enabling robust classification.

In total, the model comprised approximately 1.2 million trainable parameters. This balanced configuration of convolutional and recurrent components ensured both efficient feature extraction and accurate temporal modeling, making the architecture particularly effective for emotion recognition in online learning environments.

3.3. Experiment Planning

The model was trained on an NVIDIA GPU using TensorFlow/Keras. Batch size was 64, learning rate 0.001 with Adam optimizer, and training lasted 100 epochs. Early stopping was applied to avoid overfitting.

3.4. Evaluation Metrics

Performance was measured using accuracy, precision, recall, and F1-score. Graphs of training/testing accuracy and loss were analyzed to monitor model learning. Comparisons were made with state-of-the-art FER models as shown in Table 2 below.

Table 2. Performance Matric Comparison.

Algorithm	Dataset	Accuracy	Precision	Recall	F1-Score
CNN model	FER 2013	97%	94%	96%	95%
CNN-LSTM(Proposed Model)	FER 2013	98%	97%	98%	98%

This hybrid CNN-LSTM model is particularly well-suited for emotion classification tasks that require both spatial and temporal analysis. CNN layers efficiently extract essential facial features, while LSTM layers process sequential dependencies, making the model capable of detecting subtle emotional changes. This approach is especially useful in online learning environments where students' emotional engagement can fluctuate over time. The trained model is evaluated using various performance metrics, including accuracy, precision, recall, and F1-score, ensuring its effectiveness in real-time emotion recognition. By providing insights into students' emotional states, this model can help educators adapt their teaching strategies, improve engagement, and enhance the overall learning experience in virtual classrooms.

4. Results and Discussions

This section presents the results of the CNN-LSTM model. Compared to [1], who achieved 99% accuracy using an attentional CNN, our model achieved 98.0%. Although slightly lower, our approach focuses on a custom dataset specific to e-learning emotions, making it more relevant for educational applications.

The trade-off between general accuracy and domain adaptation is highlighted: while [1] optimized for generic FER, our approach is designed to recognize emotions directly tied to learning engagement. This novelty makes the work valuable for personalized education technologies.

Results Comparisons

Figure 2a presents a comparative analysis of precision, recall, and F1-score for the three classified emotions: bored, confused, and interested. Precision (blue) indicates how many of the predicted instances for each emotion were correctly classified, while Recall (orange) measures how well the model identifies actual instances of each emotion. The F1-score (green) balances both precision and recall to provide an overall performance metric. The bars for all three metrics are closely aligned, suggesting that the model maintains a strong and consistent classification ability across all emotion categories. The high values across these metrics indicate that the CNN-LSTM model effectively recognizes learners' emotions with minimal misclassification. Table 3 below shows the comparison of results for different methodologies and the state of the art.

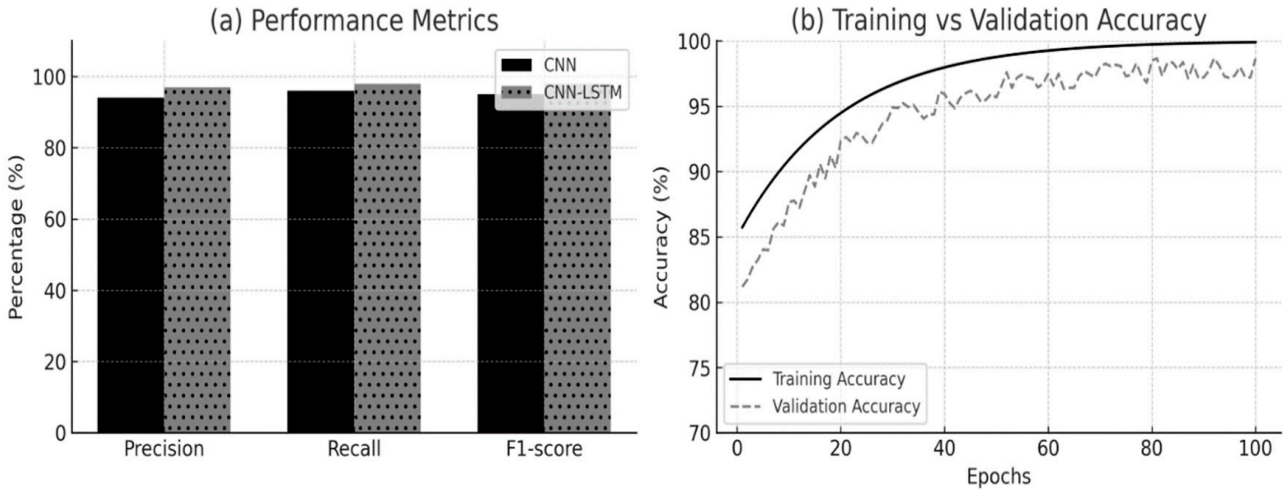


Figure 2. (**a**) Performance evaluation graph for the FER 2013 dataset. (**b**) Training and validation of loss function for the FER 2013 dataset.

Table 3. Result Comparison with State-of-the-Art Models.

S/No	Tittle	Author	Methodology	Number of Emotions Detected	Dataset	Accuracy
1.	Using Convolutional Neural Networks (CNNs) for Emotion Recognition	[4]	Deep Learning Based on CNN	4	Custom Dataset	92.5%
2.	Students' emotional recognition during online conferences during e-learning	[3]		6	FER 2013 dataset	88.7%
3.	Convolution neural networks are used to recognize facial expressions (CNN models)	[2]	CNN MODEL	7	CK+	88.2%
4.	Emotion Recognition through Facial Expression Using CNN	[1]	Attentional CNN	6	FER 2013 dataset	99.0%
5.	Our work	2024	CNN-LSTM	3	Custom dataset	98.0%

Figure 2b is an accuracy plot, which tracks the model's training and testing accuracy over 100 epochs. The training accuracy (blue) shows a steady increase, quickly approaching 100%, while the testing accuracy (orange) follows a similar trend but with noticeable fluctuations. These fluctuations, particularly around epochs 60 and 80, suggest potential overfitting, where the model performs exceptionally well on training data but struggles with new, unseen data. Despite these variations, the overall high accuracy for both training and testing suggests that the model successfully learns to classify emotions, but further optimization such as regularization or early stopping could help stabilize the performance for better generalization.

5. Conclusions

FER (Facial Expression Recognition) was introduced to address the challenge of automatically recognizing and understanding human emotions through facial expressions. Human beings convey a wealth of emotional information through their facial expressions, which play a significant role in interpersonal communication and social interactions. Understanding emotions from facial expressions is crucial in various fields, including psychology, human–computer interaction, affective computing, and artificial intelligence.

Current online teaching and learning systems often rely on binary indicators of comprehension (understood vs. not understood), which fail to capture the complexity of learners' emotional states. This study introduces a CNN-LSTM model capable of recognizing three critical emotions in online learning environments boredom, confusion, and interest. By moving beyond simplistic classifications, the proposed approach provides a more nuanced understanding of learner engagement.

In the Results and Discussion section, we present the outcomes of these approach and discuss the performance of the proposed emotion classification for detecting learners' emotions in the online learning environment. We analyze the classification results, evaluate the effectiveness of the new emotion classification scheme, and compare it to existing approaches. Furthermore, we provide insights into the implications and potential applications of our proposed classification in the context of online learning.

Finally, the success of the proposed deep learning model will be evaluated through rigorous experimentation and validation on the modified dataset of learners' emotions. Performance metrics such as accuracy, precision, recall, F1-score, and other relevant evaluation measures will be used to assess the model's effectiveness in emotion classification. The proposed model can be used to promote personalized support and improve learning outcomes. Future work includes expanding the dataset and exploring other modalities for emotion recognition.

Future Work

Future work includes expanding the dataset and exploring other modalities for emotion recognition, including multimodal emotion recognition (voice + posture), larger datasets, and real-time classroom deployment.

Author Contributions: Conceptualization, B.M.B.; methodology, B.M.B.; validation, B.M.B. and H.A.U.; formal analysis, B.M.B.; resources, B.M.B.; data curation, B.M.B.; writing—original draft preparation, B.M.B.; writing—review and editing, H.A.U.; supervision, H.A.U.; funding acquisition, B.M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://www.kaggle.com/datasets/msambare/fer2013 (accessed on 10 November 2021).

Acknowledgments: I would like to acknowledge the valuable administrative and technical support provided during the course of this research. The assistance from laboratory staff and technical teams in data collection, processing, and analysis has been instrumental in ensuring the successful completion of this study. Their expertise and contributions have significantly enhanced the quality of this work. Additionally, I extend my appreciation to the institutions and organizations that facilitated access to resources, datasets, and computational tools necessary for this research. Their support, whether through equipment, materials, or research facilities, has been essential in carrying out the experiments and achieving the objectives of this study.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Networks
LSTM	Long Short-Term Memory
FER	Facial Expression Recognition
AU	Action Units

References

1. Ahmed, L.J.M.; Haruna, A.; Sani, U.; Mohammed, S.; Mohammed, B. CNN-LSTM deep learning based forecasting model for COVID-19 infection cases in Nigeria, South Africa and Botswana. *Health Technol.* **2022**, *12*, 1259–1276. [CrossRef]
2. Nour, N.; Elhebir, M.; Viriri, S. Face expression recognition using convolution neural network (CNN) models. *Int. J. Grid Cloud Comput. Appl.* **2020**, *11*, 1–11. [CrossRef]
3. Panichkriangkrai, C.; Silapasuphakornwong, P.; Saenphon, T. Emotion recognition of students during e-learning through online conference meeting. *Sci. Eng. Health Stud.* **2021**, *15*, 21020010. [CrossRef]
4. Badrulhisham, N.A.S.; Mangshor, N.N.A. Emotion Recognition Using Convolutional Neural Network (CNN). In Proceedings of the 1st International Conference on Engineering and Technology (ICoEngTech) 2021, Perlis, Malaysia, 15–16 March 2021. [CrossRef]
5. Zaher, S.; Masoud, W. Emotion recognition through facial expression using attentional CNN. *IJCSIS* **2023**, *21*, 7. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.