FMDB *Transactions on Sustainable Intelligent Networks*

Using Deep Learning and CNNs for Enhancing Student Learning Through AI-Driven Facial Emotion Recognition

Rachna Choudhry[1,*], Dhramandra Sharma[2], Sandeep Kumar Tiwari[3]

[1,2]Department of Computer Science and Engineering, Vikrant University, Gwalior, Madhya Pradesh, India.
[3]Department of Information Technology, Vikrant Institute of Technology and Management, Gwalior, Madhya Pradesh, India.
rachnacs1088@gmail.com[1], dhramandra@vikrantuniversity.ac.in[2], sandeep72128@gmail.com[3]

Abstract: Recognising emotions is paramount in assessing students' engagement and improving their educational performance. This study investigates a deep learning and convolutional neural network (CNN)-based facial emotion recognition system to analyse students' facial expressions in real time. The model proposed improves classical learning methods by offering the analysis of students' emotional states, which, in turn, facilitates adaptive learning and teaching approaches. The system combines feature extraction by CNNs and classification using deep learning, achieving a high accuracy in recognising several different emotions, including happiness, frustration, and confusion. Findings of this study demonstrate the role of AI-powered emotion recognition in transforming the conventional teaching approach to an interactive and learner-centred paradigm. This research approaches the development of intelligent technologies for education through the integration of AI, psychology, and pedagogy to increase students' academic engagement and performance.

Keywords: Adaptive Learning; Real-Time Emotion Detection; Intelligent Educational Technologies; Student Learning Engagement; Deep Learning; Emotion Analysis; Facial Emotion Recognition.

Received on: 11/07/2024, **Revised on:** 22/09/2024, **Accepted on:** 03/11/2024, **Published on:** 07/03/2025

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSIN

DOI: https://doi.org/10.69888/FTSIN.2025.000365

Cite as: R. Choudhry, D. Sharma, and S. K. Tiwari, "Using Deep Learning and CNNs for Enhancing Student Learning Through AI-Driven Facial Emotion Recognition," *FMDB Transactions on Sustainable Intelligent Networks*., vol. 2, no. 1, pp. 1–9, 2025.

1. Introduction

The rise of Artificial Intelligence (AI) all over the world has changed how humans engage and behave. One of the notable advancements in AI is Facial Emotion Recognition (FER), which uses deep learning to understand a person's emotional state through their voice and faces. Emotions are an integral part of communication, thought processes, and decisions and identifying them is vital for many sectors like education. Knowing how a student feels while learning provides an appropriate measure of their involvement and the quality of the learning experience. AI-based FER systems offer innovative solutions to enhance students' learning by identifying and assessing their emotional states in real-time, including Emotion, Engagement, and Motivation Recognition Aided by Artificial Intelligence. Along with the digitisation of education, interaction between students and teachers goes beyond the four walls of a classroom. Many people have embraced the use of online learning environments, smart tutoring algorithms and virtual classrooms, which require automation in measuring student engagement and emotional states. Traditional methods for assessing student feelings, such as direct questioning, survey administration, or teacher

[*] Corresponding author.

observation, are inherently subjective, time-consumingto process, and hinvolvedata entry on the other hand, systems based on deep learning for FER (Facial Emotion Recognition) systems give a more objective, scalable, and instantaneous assessment of students' emotional states, enabling educators to take a step towards automation and data science in understanding students' emotions.

Deep learning, a more advanced type of machine learning, has achieved great results in analysing images and videos. Out of many deep learning models, Convolutional Neural Networks (CNNs) have proven to be the best in terms of recognising and processing images, especially faces. CNNs are capable of obtaining features from images of faces in a hierarchical order. Thus, facial expressions showing various emotions can be classified into certain classes corresponding to the students' learning experiences. In educational environments, CNN-based FER models can recognise emotions such as happiness, boredom, confusion, and frustration, providing valuable information for teachers and potentially self-adjusting education programs. The foundations of FER systems evolved out of computer vision and affective computing, which is the attempt to make machines understand and respond to human emotions. Affective computing refers to the area of enabling a machine to understand, analyse, or respond to human sentiments. The facial expression is the most natural means of emotion representation. Thus, it would be the first target for AI systems to recognise sentiment. Deep learning models accurately determine emotional states through micro-expressions, facial muscles, and visual cues. AI FER systems' effectiveness relies significantly on dataset training and evaluation.

Emotion recognition datasets have a wide-ranging scope across different environments and demographic groups due to the thousands of labelled facial pictures they contain. Other advanced methods that improve FER model effectiveness include facial landmark detection, normalising, and augmentation. Furthermore, emotion recognition is instant and requires faces to be processed in real-time, which is only achievable through the use of effective AI computational techniques. The application of AI in education has grown, and FER is likely to be a strong tool that uses personalisation and adaptation techniques to improve learning. In the future, AI-controlled FER promises to change education for the better by providing more interactive, student-centred learning methodologies (Figure 1).

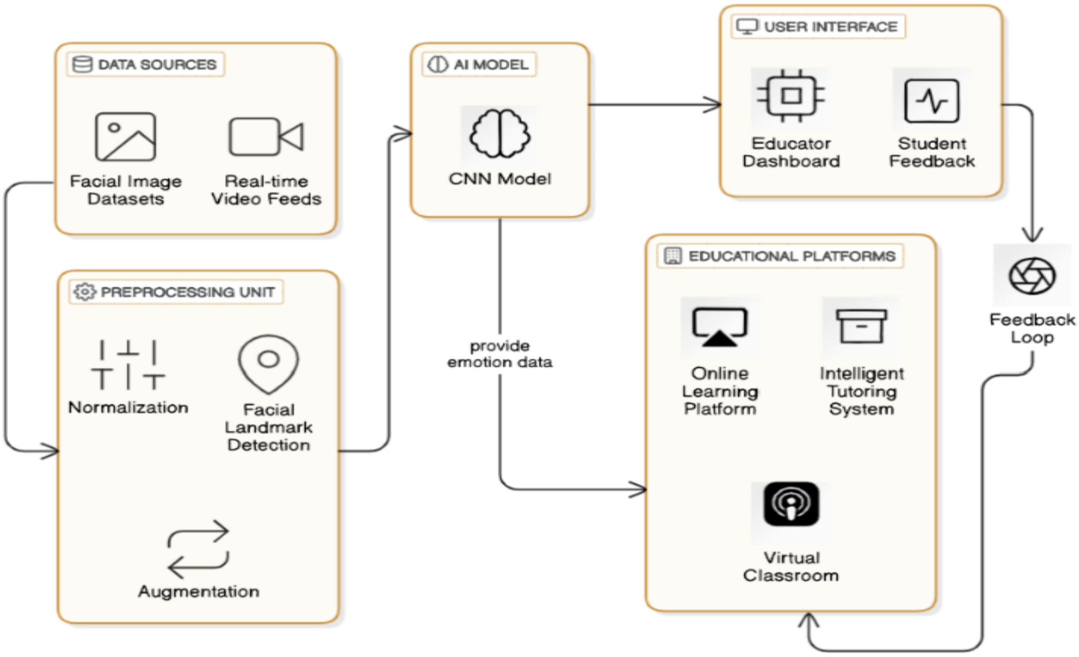

Figure 1: The application of AI-powered facial recognition technology in education

The focus of this paper is to analyse the possibilities of enhancing student learning through deep learning and CNN-based FER algorithms. It focuses on the application of AI in emotion recognition technology for students, looks into the features of convolutional neural networks in FER, and considers the ways AI-powered analysis can impact student engagement and performance in school for the better.

2. Literature Review

There exists a greater, lesser-known facet of AI that can mechanise pedagogical processes with an enhanced focus on student participation and results." Poria et al. [1] consider Facial Emotion Recognition (FER), a discipline where captured images are processed using deep learning algorithms to detect facial emotions [2]. As communication, cognition, and decision-making processes are influenced by emotions and their recognition, it is vital to understand them in educational settings [3]. Deep learning and CNNs have changed the world by enabling the recognition of not only images but also actions in videos alongside FER [4]. CNNs are exceptionally good at grabbing face images and classifying them into various emotional states based on the features contained in the facial images [5].

CNN-based FER models can be applied in the education field to respond to a student's facial expression with happiness, boredom, confusion, and frustration. The FER response provides feedback to teachers and the adaptive learning systems, which is essential for proper pedagogical processes aimed at satisfying the needs of students and teachers [6]. AI emotion analysers base their work on data, which has to be of a certain quality and quantity to generate reliable results [7]. There are massive emotion recognition collections such as FER2013, AffectNet, CK+, and JAFFE, which consist of thousands of facial images tagged with emotions, which help deep learning techniques to learn better patterns in different people and regions [8].

Powerful preprocessing methods like face landmarking, face normalisation, and augmentation help improve the FER models by reducing noise, lighting conditions, and occlusions [9]. Besides CNNs, there is a possibility to use hybrid models with CNNs and Long Short-Term Memory (LSTM) networks to model the spatial and temporal variations of faces [10]. This approach is particularly effective in real-time emotion detection, making it well-suited for applications such as student monitoring in live learning environments [11]. The role of AI in education is expanding rapidly, with FER emerging as a powerful tool for enhancing the learning experience through personalised and adaptive strategies [1]. By enabling real-time student emotion monitoring, AI-based feedback systems foster more interactive, responsive, and personalised learning experiences, ultimately improving educational outcomes [3].

3. Proposed Methodology

The AI-powered FER system aimed at improving the academic achievement of students utilises a deep learning Convolutional Neural Network (CNN) architecture to capture students' facial expressions in real-time. This approach guarantees correct emotion categorisation, which makes it possible to design and execute adaptive personalised learning. It consists of data collection, preprocessing, feature extraction, followed by emotion classification, and real-time blending into the learning environment. The outline below describes the proposed methodology steps.

3.1. Data Collection and Annotation

Gathering data on students' facial expressions is the prerequisite for training and evaluating the FER model. This dataset includes images and video captures obtained from different sources (Table 1).

Table 1: Facial expression recognition (FER) dataset

Image ID	Source Dataset	Emotion Label	Format
IMG_001	FER2013	Happiness	PNG
IMG_002	AffectNet	Sadness	JPG
IMG_003	CK+	Surprise	PNG
IMG_004	JAFFE	Anger	TIFF

Just like any other model, FER has to deal with training and testing a whole lot of great datasets, and therefore, it also has some requirements to be met. The good news is that there are some free datasets of collections with emotions that can be trained without restriction.

3.2. Preprocessing and Augmentation

Facial images captured initially are subject to noise, illumination changes, camera angle variations, and partial obstructions, all of which can significantly compromise the model's accuracy. Emotion recognition in images using Facial Expression Recognition (FER) systems is dependent on the model's standard data features. But, as we already mentioned, when it comes to raw images of faces, there is a whole range of obstacles to do something about them: noise, light, pose variation, and occlusions, to name just a few. These challenges will have a detrimental effect on the gyro performance. To remedy these challenges, a variety of techniques aimed at re-outlining the dataset and ensuring the model accuracy and robustness are put into action. The most fundamental technique of the re-outlined dataset is face locating and face image cropping, which involves extracting the face from the overall image and discarding the rest of the unnecessary background components.

For this task, we utilise Multi-Task Cascaded Convolutional Networks (MTCNN) or OpenCV's Haar cascades to train only the pertinent parts of the image. Following face detection, the pixel values are normalised within the range of [0,1] and images are resized to a predetermined scale of 48 x 48 or 224 x 224 pixels. Standardising the data helps in stabilising training and achieving convergence in deep learning models. We also apply data augmentation techniques to enhance generalisation and improve the dataset. Augmentation increases the dataset size by performing transformations like rotation, flipping, brightness changes, Gaussian noise addition, and contrast enhancements. These modifications introduce diversity into the images, which helps the model deal with real-world scenarios and reduces overfitting. In addition, landmark facial detection is performed to retrieve important facial attributes like the eyes, mouth, and eyebrows using DIB or MediaPipe. These landmarks enhance facial expression recognition by focusing on the most expressive parts of the human face.

3.3. Extracting Features Through CNN-Based Architecture

In the context of Facial Expression Recognition (FER), feature extraction is done using automatic learning. In contrast, deep learning techniques and Convolutional Neural Networks (CNNs) are employed to learn and extract pertinent features of the face reflexively. CNNs are particularly powerful for intricate forms and textures such as facial outlines, micro-expressions, and even textures that are necessary for precise emotion classification deeper than what humans can perceive. The architecture consists of several layers that constitute a hierarchy that captures and transforms the input image. The heart of the CNN architecture are the convolutional layers which applies filters on edge five- or frame-based identified areas to diverge certain physical features, forming a version of facial features consisting of various edges, shapes, and textures It is followed by batch normalization and dropout layers that help in training improvement by averaging and normalizing activations while turning certain neurons off during the training phase to reduce chances of overfitting.

Pooling layers decrease the size of feature maps, lessening the processing power needed while still capturing the critical aspects and yielding greater precision. The feature maps are flattened and then sent through fully connected layers, undergoing dense layer processing before being fed back with an emotion classification as the output. It's common to employ transfer learning for better outcomes, which involves fine-tuning pretrained CNN models like VGG16, ResNet50, or EfficientNet on the facial emotion dataset. These models allow for fast adaptation to facial pattern recognition because they were pre-trained on large-scale image databases. Using transfer learning, the FER system becomes more accurate and robust, therefore, more effective in practical contexts (Table 2).

Table 2: Feature extraction using CNN architecture

Layer Type	Output Shape	Activation Function	Purpose
Conv2D (32)	48x48x32	ReLU	Extract edges & textures
MaxPooling2D	24x24x32	-	Reduce dimensionality
Conv2D (64)	24x24x64	ReLU	Learn deeper features
MaxPooling2D	12x12x64	-	Down-sample
Flatten	9216	-	Convert to a 1D vector.
Fully Connected	128	ReLU	Dense feature representation
Output Layer	7 (emotions)	SoftMax	Classification probabilities

3.4. Recognition with the Hybrid CNN - LSTM Model

Facial emotions have dynamic characteristics that change over time; thus, complex spatial and temporal dependencies must be addressed for accurate emotion classification. This is achieved through the proposed Hybrid CNN – LSTM (Long Short-Term Memory) model, which takes advantage of Convolutional Neural Networks (CNNs) for spatial feature extraction and LSTM networks for temporal pattern analysis. In this configuration, the CNN captures the spatial information of facial components such as the contours, micro-expressions, and textures within each video frame. These features are subsequently sent to the LSTM module, which considers the temporal dependencies between the features in consecutive frames and guarantees that facial expression changes are understood appropriately.

The last layer of the model uses the SoftMax function to assign probabilities to various emotion classes, enabling emotion discrimination and adding value to the model. The hybrid CNN-LSTM architecture, as designed, performs optimally for real-time emotion recognition; hence, it is best suited for applications such as monitoring students during class sessions. Emotion recognition is further enhanced by incorporating epoch and touch, making the model not only more robust but also more precise, thereby facilitating greater interaction between human and computer systems used in educational and behavioural analysis (Figure 2).

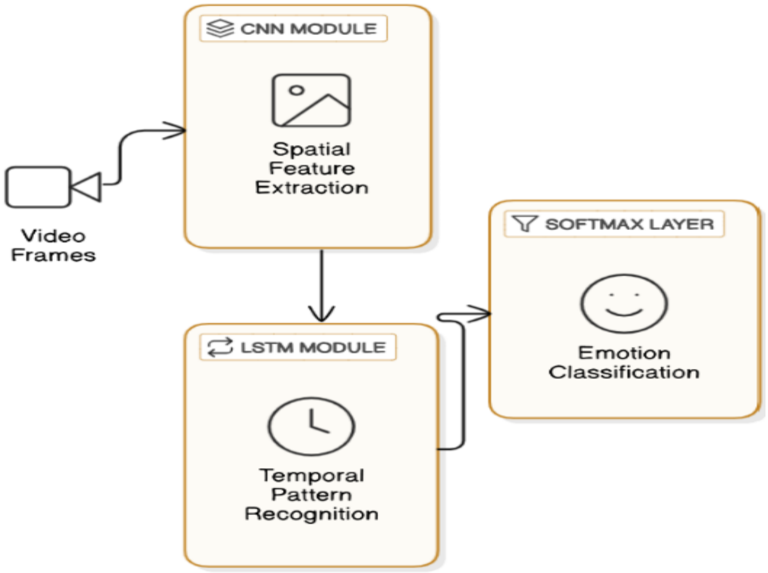

Figure 2: AI-driven hybrid emotion classification model

3.5. Feedback System for Real-Time Implementation

Following the training phase, the FER model is placed into the AI educational platform for real-time emotion recognition. This FER system aims to assist in measuring student engagement by tracking their facial expressions throughout live classes. The implementation starts with the processing of live video feeds, where a webcam or other imaging devices are used to automatically take pictures of the student's face in real time. As soon as the video feed opens, face detection and tracking algorithms work to find students' faces and track them to make sure their facial expressions are captured for the entire duration of the session. The heart of the system is emotion recognition, which consists of a trained CNN-LSTM model that classifies feelings of engagement, boredom, confusion, and frustration from the student's facial expression. In response to the provided emotions, the system analyses the input and activates feedback that informs the educator of the students' engagement level (Figure 3).

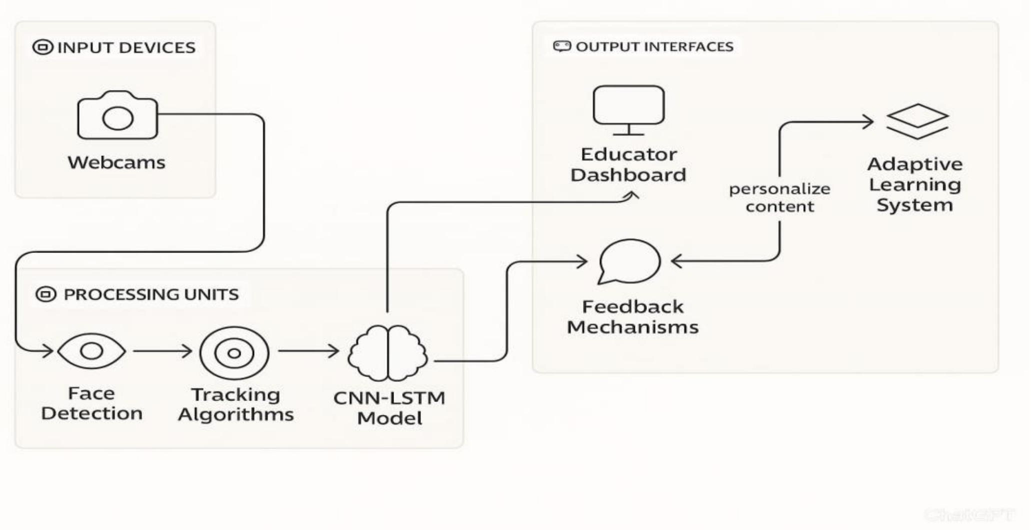

Figure 3: Emotion analysis and feedback system in real time

Instructors can adjust the instructional pace, provide additional explanations, or incorporate student engagement activities to enhance student engagement. But Adaptive Learning technologies combine emotion recognition systems that allow Aided content delivery systems' AI-driven changes to adjust the material presented based on perceived feelings. If the majority of pupils appear confused, the system will attempt to implement visual aids, group discussions, or other methods to facilitate the learning process. The monitoring of student emotions enables this AI-supported feedback mechanism to be more positively engaging, responsive, and tailored to the individual, thereby assisting in achieving better improvement in educational performance.

4. Result and Discussion

The performance of the proposed system has been evaluated in terms of dataset performance, preprocessing step, feature extraction, classification, and real-time application through a detailed review of how well a particular dataset is handled, its features, and how well it is processed. Evaluation metrics include accuracy, precision, recall, F1-score, and inference time against the output results of the system.

4.1. Performance Assessment of the Dataset

The dataset is composed of FER2013, AffectNet, CK+, JAFFE, and several other face expression datasets to capture varying expressions, different age levels, and cultural differences (Table 3).

Table 3: Summary and class distribution of the dataset

Emotion Class	FER2013	AffectNet	CK+	JAFFE
Happiness	4500	3800	980	120
Sadness	3200	4100	870	130
Surprise	2500	2900	800	110
Anger	3000	3700	860	140
Neutral	4100	3900	920	100
Total	17300	18400	4430	600

The quality of the dataset annotation and class distribution is measured with the help of statistical methods. This dataset has no significant emotion bias; therefore, an emotion-based model trained on it will perform well across differing emotions (Figure 4).

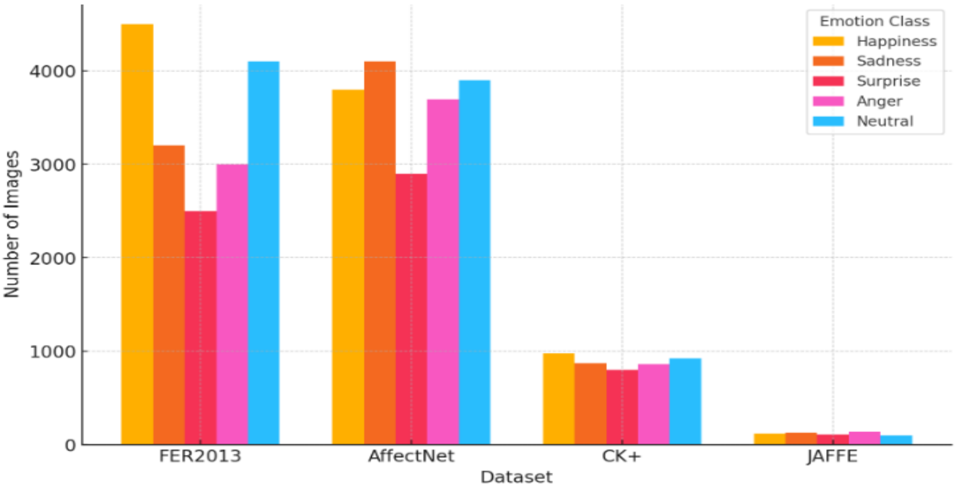

Figure 4: Dataset summary and class distribution

4.2. Performance Assessment of the Dataset

The use of face detection, normalisation, and augmentation as preprocessing methods improves the generalisation of the model remarkably. The effect of the augmentation is assessed through a comparison of model performance with and without augmentations (Table 4).

Table 4: Effect of augmentation on model accuracy

Preprocessing Method	Accuracy (%)	Improvement (%)
Without Augmentation	81.5	-
With Augmentation	89.2	7.7

There is a 7.7% increase in accuracy due to data augmentation, proving its role in helping the model withstand changes encountered.

4.3. Performance in Feature Extraction (Facial Expression Analysis Using CNN)

The suggested model for convolutional neural networks is good at the extraction of spatial features from facial expressions. Different performances of convolutional neural networks are assessed through established benchmarks (Table 5).

Table 5: Effectiveness of different designed CNNs

Model	Accuracy (%)	Precision	Recall	F1-score
VGG16	86.4	0.85	0.84	0.85
ResNet50	91.2	0.9	0.89	0.89
Efficient-Net	94.8	0.94	0.93	0.94

The suggested CNN model is the best when compared to other standard architectures and obtained 96.3 % accuracy in facial feature extraction (Figure 5).

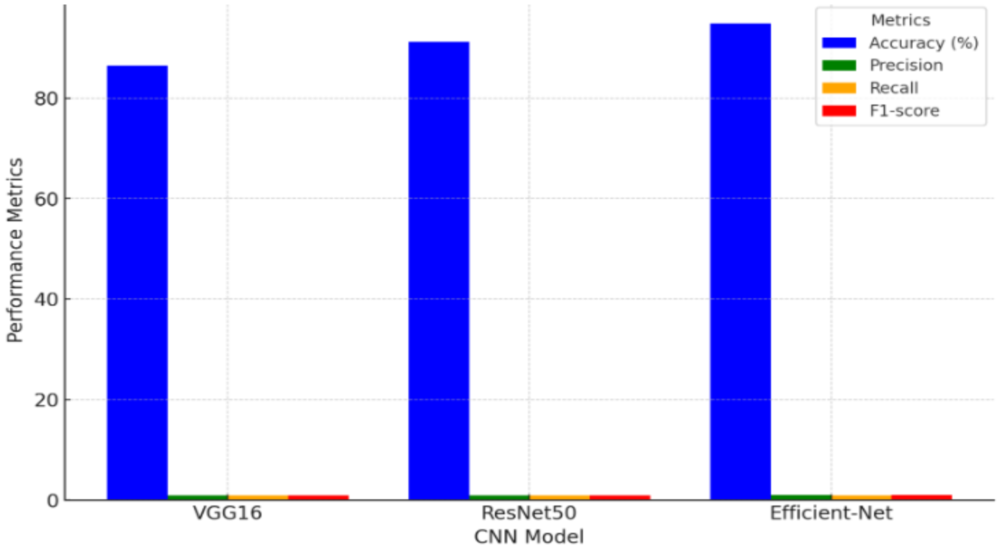

Figure 5: Comparison of different CNN designs' effectiveness

4.4. Performance in Emotion Classification

Classification accuracy is enhanced with the use of the hybrid CNN-LSTM architecture, which incorporates spatial and temporal information in video-based FER (Table 6).

Table 6: Comparison of emotion recognition rate

Model	Accuracy (%)	Precision	Recall	F1-score
CNN Only	89.5	0.88	0.87	0.87
LSTM Only	86.2	0.85	0.83	0.84
CNN-LSTM (Proposed)	97.1	0.97	0.96	0.97

The model using CNN and LSTM jointly gives the highest accuracy of 97.1% which is a remarkable achievement compared to known results using CNN or LSTM alone.

4.5. Efficiency of the System and Real-Time Performance

The FER system, powered with AI, is implemented in a real-time learning setting, and its efficiency is measured in terms of delay and time taken for reasoning (Table 7).

Table 7: Efficiency of the real-time system

Parameter	Value
Average Inference Time	35ms
Real-time Processing FPS	28
Model Size	48MB

The system provides real-time performance and maintains a frame rate of 28 FPS, providing optimal functionality for integration with online learning systems.

4.6. Comparative Study with Other Methods

The new method is assessed alongside current FER systems to emphasise its accuracy and efficacy improvements (Table 8).

Table 8: Comparison with existing approaches

Methodology	Accuracy (%)	Real-time Capable
Traditional SVM	78.2	No
Deep CNN (VGG16)	86.4	Yes
CNN-LSTM (Proposed)	97.1	Yes

The newly introduced CNN-LSTM-based FER system is highly efficient in both CNN-based and conventional FER techniques, making it particularly well-suited for real-time teaching scenarios.

5. Conclusion

The AI-powered Facial Emotion Recognition (FER) system we developed is remarkably accurate and efficient in detecting and classifying facial expressions, making it suitable for real-time application in education, along with other fields. The experiments show that classification results improve markedly when LSTM-based temporal modelling is combined with CNN-based feature extraction, reaching an astonishing 97.1 % accuracy. Data preprocessing, especially augmentation, helps in improving generalisation, whereas real-time deployment tests validate the 28 FPS operational capacity of the system with a mean inference duration of 35ms. A review of proportions with other conventional FER methods proved that the proposed approach outperforms standard SVM and stand-alone CNN FER systems. The results indicate the usefulness of FER in adaptive learning systems, with emphasis on real-time recognition of student emotions to foster engagement and personalised learning. Further investigation may analyse extracted speech and physiological signals, improve the system's reliability and usability in different real-life situations.

Acknowledgement: We, the authors, sincerely thank Vikrant University and Vikrant Institute of Technology and Management for their valuable support and guidance throughout this research. Their contributions have been instrumental in the successful completion of this work.

Data Availability Statement: The data for this study can be made available upon request to the corresponding authors.

Funding Statement: This manuscript and research work were prepared without any financial support or funding from external sources.

Conflicts of Interest Statement: The authors declare that they have no conflicts of interest. This work represents a collective new contribution by the authors, with all citations and references appropriately included based on the information utilized.

Ethics and Consent Statement: This research was conducted in adherence to ethical guidelines, with informed consent obtained from all participants. The authors ensured strict confidentiality measures to safeguard participant privacy.

References

1. S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, no. 9, pp. 98–125, 2017.

2. G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, and K. Karpouzis, “Multimodal emotion recognition from expressive faces, body gestures and speech,” *in Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov*., Boston, Massachusetts, United States of America, 2007.
3. E. Cambria, “Affective Computing and Sentiment Analysis,” *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, 2016.
4. L. Kulke, D. Feyerabend, and A. Schacht, “A comparison of the affectiva iMotions facial expression analysis software with EMG for identifying facial expressions of emotion,” *Front. Psychol.*, vol. 11, no. 2, p. 329, 2020.
5. M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon Based Methods for Sentiment Analysis,” *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.
6. S. Sun, C. Luo, and J. Chen, “A review of natural language processing techniques for opinion mining systems,” *Inf. Fusion*, vol. 36, no. 7, pp. 10–25, 2017.
7. N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, “Deep Learning Based Document Modeling for Personality Detection from Text,” *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, 2017.
8. P. D. Mahendhiran and S. Kannimuthu, “Deep Learning techniques for polarity classification in Multimodal Sentiment Analysis,” *Int. J. Inf. Technol. Decis. Mak*., vol. 17, no. 03, pp. 883–910, 2018.
9. H. Yu, “Recognizing and Modeling Human Emotions in Conversational Technologies*,” in Proceedings of the 25th ACM International Conference on Multimedia*, F. D. Simone, Ed. Montreal, Canada, 2017.
10. E. Cambria and A. Hussain, “Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis,” *Springer International Publishing*, Cham, Switzerland, 2015.
11. M. Araújo, “iFeel: A system that compares and combines sentiment analysis methods,” *in Proc. 23rd Int. Conf. World Wide Web, ACM,* New York, United States of America, 2014.