Coordinate-Attention Enhanced EfficientNet for Emotion Recognition in Virtual Learning Environments

Chinthakunta Manjunath[1], *Lavanya*. S P[2*], *Merly* Thomas[3], *Manohar*. M[4], and *Meena*. J[5]

[1]Department of Computer Science Engineering (AI&ML), Vidyavardhaka College of Engineering, Mysuru, India
[2]Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Nitte (Deemed to be University), Bengaluru, India
[3]Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Mumbai, India
[4]Department of Computer Science and Engineering, CHRIST (Deemed to be University), Bengaluru, India
[5]Department of Humanities, Dayananda Sagar College of Engineering, Bengaluru, India

Abstract. In recent times, emotion recognition is a continuous evolving process used to understand an individual's emotional state especially in virtual learning environments. While existing models made improvements, a consistent challenge is to accurately recognize subtle and complex facial emotions that change rapidly, which is common with students in dynamic online sessions. To address, the proposed method CA-EffiecientNet introduces an enhanced EfficientNet architecture integrated with a Coordinate Attention (CA) block, which is specifically designed to capture both fine-grained spatial details and long-range dependencies in facial emotions. The proposed CA-EffiecientNet model improves the models ability to identify subtle signs related to a student's emotional state such as frustration or involvement. The virtual leaning system integrates this proposed CA-EffiecientNet model to continuously monitor learners activity and dynamically adapt educational content. The performance of the CA-EffiecientNet is evaluated on the KDEF and RAF-DB datasets, achieving superior results with the accuracy of 97.6 and 98.57 compared to existing ResNet with Convolutional Block Attention Module (CBAM). The results also demonstrate the efficiency of the proposed CA-EffiecientNet model supported by its integration of MBConv layers, which reduce computational load while preserving model robustness.

1 Introduction

Emotion recognition is a continuously changing method expected to interpret an individual's emotional state, which tends to shift depending on the activity and the surrounding context [1]. Emotion recognition is a wide range of practical applications, including enhancing human–computer interaction, supporting intelligent tutoring platforms, monitoring surveillance systems, assessing patients psychological well-being, and assisting in lie detection [2]. Recent developments in technology and emotional understanding leads to new possibilities for Automatic Emotion Recognition (AER) systems, which are widely adopted in applications such as security surveillance, driver fatigue detection, health monitoring, and interactive simulation design [3]. People across the world express emotions such as anger, joy, or fear through facial cues, but the way these expressions appear differs depending on cultural traditions, spoken language, and religious values [4]. Understanding students emotional states in online learning environments is essential for delivering adapted instruction and addressing the emotional disconnect regularly observed in digital educational settings [5]. AER contributes significantly to improving student learning in educational environments, and its growing presence underscores its relevance to daily interactions [6]. Naturally, language, voice, and facial expressions are the primary methods through which humans convey their emotions. In addition, how people combine these methods to express their emotions is complicated [7]. Recognizing and responding to users' emotions in real time leads to the transformation of virtual experiences that are not only more immersive and personalized but also more aligned with the differences in human behavior [8]. Pandemic-driven changes in education have highlighted the growing significance of socially shared regulation, which is essential for navigating collaborative learning in virtual settings [9]. The progress behind online education is expected to persist, driven by evolving policies that encourage the expansion of virtual classrooms and widespread adoption of online and hybrid learning models by educational institutions [10].

The contribution of the proposed model is as follows:

- The proposed model directly integrates emotion recognition into a functional virtual learning environment to provide a dynamic and personalized educational experience. This allows the system to analyze a student's emotional state, such as frustration or involvement, and

* Corresponding author: lavanya.sp@nmit.ac.in

automatically adjust the learning content to provide a more effective and supportive learning path.

- Introduction of the Coordinate Attention (CA) block is more effective in analyzing facial emotions, allowing recognition of both specific, long-range facial patterns and precise facial signs, which is essential for accurately identifying related emotional states.
- The proposed models built on an EfficientNet50 framework to ensure both high performance and computational efficiency. Using Mobile Inverted Bottleneck Convolution (MBConv), the architecture reduces processing demands and maintains superior accuracy in emotion classification.

The following Section 2 discusses related works that explain the contribution of different authors to emotion recognition. Section 3 presents the problem statement of the paper which is solved in Section 4, and Sections 5 and 6 discuss the results and conclusions.

2 Literature Survey

Shaun George Rajesh et al. [11] employed a multimodal framework to enhance virtual assistants by integrating Facial Emotion Recognition (FER) using ViT-Tiny, Textual Emotion Recognition (TER) using MiniLM, Emotion Fusion Mechanism to combine FER and TER results, and Emotion-aware Response Generation using fine-tuned DialoGPT. The employed model was used for real-time use in resource-constrained environments, such as smart assistants and edge devices, with the aim of improving emotional intelligence and user engagement. However, the employed fusion framework follows fixed rules but does not learn or adapt. Thus, it was unable to handle complex or mixed emotions well and did not improve with experience or personalize to different users.

Michael Shomoye and Richard Zhao [12] developed Deep Learning (DL) model based on the ResNet50 architecture for emotion recognition in Virtual Reality (VR) classrooms The developed model collected a specialized dataset representing VR classroom conditions applied for transfer learning by freezing early layers and fine-tuning deeper ones, and incorporated dropout and L2 regularization to prevent overfitting. Through comparative experiments with VGG19 and MobileNet, their custom ResNet50 consistently performed better than the other models in both full-face and occluded-face scenarios, enabling real-time emotion detection to support educators in adapting their teaching strategies. However, the developed model depends on static facial regions, lacks dynamic temporal analysis, and does not incorporate multimodal indications such as voice or body posture, leading to reduced performance under uncontrolled lighting or camera angles.

Mohammed Aly [13] employed a DL-based ResNet50-CBAM-TCN model as an emotion recognition model specifically designed to monitor student engagement in online classrooms. The employed model integrates the well-known feature extraction capabilities of ResNet50 with Convolution Block Attention Mechanism (CBAM) and temporal model TCNs (Temporal Convolutional Networks). They captured facial images periodically during virtual learning sessions, extracted relevant emotional features, and tracked dynamic changes in expressions such as happiness, sadness, confusion, and boredom over time. However, the employed model depends on facial cues and ignores multimodal signals, such as voice or posture. It struggled with subtle emotional variations and cultural differences.

Abdullatif Ghallab et al. [14] used the ResNet50-CBAM model for facial emotion recognition specifically designed for online learning environments. The integrated ResNet50 deep convolutional neural network for feature extraction for its residual down-sampling modules to preserve spatial details and incorporated the Convolutional Block Attention Module (CBAM) to emphasize relevant facial regions and suppress noise. The model periodically captured students' facial images through a webcam, processed them to detect emotional states such as happiness, sadness, anger, and surprise, and used these insights to dynamically adjust teaching strategies, such as modifying content difficulty or modifying the learning method. However, the model used depends on facial emotions, ignoring multimodal signals, such as voice or posture, and struggles with culturally complex emotional expressions.

Chahak Gautam and Seeja K.R [15] introduced a Histogram of Oriented Gradients (HOG)-Convolution Neural Network (CNN) and Scale-Invariant Feature Transform (SIFT)-CNN model facial emotion recognition. The introduced model includes emotion-specific features, first extracted using HOG and SIFT, which capture edge alignments and key facial regions, such as jawline curvature, eye openness, and mouth shape. The extracted features were then fed into a sequential convolutional neural network composed of three convolutional layers, followed by fully connected dense layers with ReLU activations. The final output layer employs a softmax classifier to categorize facial expressions into seven emotion classes. However, the introduced model lacks temporal analysis, limiting its applicability to dynamic or video-based emotion recognition.

3 Problem Statement

Existing models struggle with the difficulty of accurately recognizing subtle and complex facial emotions, such as the rapid emotions of students in virtual learning, because a student's emotional state can change depending on the activity and situation.

4 Proposed Method

The Proposed method introduces the EfficientNet50 architecture, which focuses on the integration of Coordinate Attention (CA) and optimization of its convolutional blocks to improve facial emotion recognition performance. These modifications aim to strengthen the model's ability to capture both spatial and channel-wise dependencies, thereby filtering the feature representation. Building on this improved recognition framework, an online learning status monitoring system is developed to track and interpret learner engagement in real-time. The final section presents the evaluation strategy used to assess the effectiveness of the proposed system.

4.1 System Architecture

The proposed online learning system integrates emotion analysis with facial expression recognition to improve student learning. It captures students' facial images using a camera during learning sessions, analyzes their expressions using DL techniques, and classifies them into predefined emotional categories. Based on these emotional signs, the system evaluates each learner's current state, such as focus, frustration, or interest, and delivers personalized feedback through the platform. This dynamic interaction allows the system to adapt instructional strategies in real time, supporting a more responsive and emotionally aware virtual classroom experience and system workflow, as shown in Figure 1.

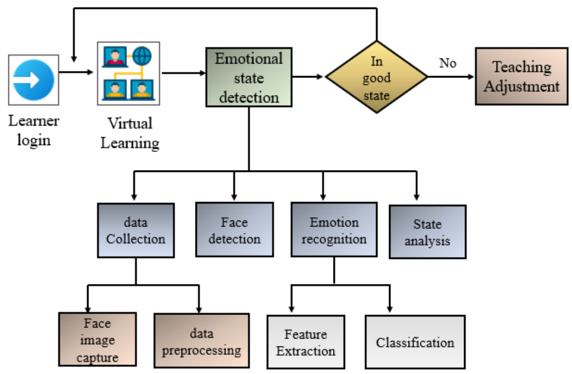

Fig. 1. Work flow of virtual learning system.

Entire Online Learning System Workflow is as follows:

- Account System: Learners begin by registering and logging into the platform to create a personalized profile for their educational journey.
- Learning State: Once logged in, students explore course materials either suggested by the system or discovered through their own search. This phase focuses on active participation in the content.
- Monitoring Emotional Responses: As students interact with the learning material, the system periodically captures facial images using a device camera. These snapshots are analyzed to interpret learners' emotional states, such as interest, frustration, or confusion. If the system detects a positive emotional response, the student continues with the current learning path.
- Adapting the Learning Experience: If signs of disengagement or negative emotions are identified, the system adjusts its learning strategy. This includes simplifying complex topics, slow pace, or introducing related stories and real-world examples to re-engage learners emotionally. These adjustments aimed to create a more supportive and motivating environment.

4.1.1 Evaluating Student Involvement in Online Learning

Students' involvement in online learning evaluated based on focus, emotional responses, engagement level, posture, and course preferences. These indicators help determine attention, difficulty faced, interest, and learning style and are described as follows:

- Focus: In online sessions, the system checks for the students' face every 10 s; if it is missing for over half that time, it is assumed that they are either distracted or experience technical problems.
- Difficulty: If a student repeatedly shows signs of anger or distress during online learning 30 times in sequence, the system identifies the related content as difficult and responds by offering a clearer explanation or connecting the student with the teacher for immediate help.
- Engagement: During online sessions, students' emotions are monitored, and their involvement is measured by the frequency of happy or surprised expressions.
- Involvement Rate: The system monitors how often students keep their heads down during online learning. If this behavior exceeds 50% of the time, it suggests low participation or possible disinterest, which triggers a prompt from the system.
- Learning Preference: The system compares students emotional reactions across different subjects to identify those that activate happiness or surprise. Based on this, similar courses or instructors are recommended to match their positive learning experiences.

4.2 Data Collection

The proposed Co-Attention Enhanced EfficientNet used two datasets, KDEF and RAF-DB, to train and evaluate emotion recognition models that contain facial images with various emotional expressions, and are essential for developing and testing the ability of the models to accurately classify emotions.

4.2.1 KDEF [16]

The KDEF dataset contains a total of 490 JPEG images, showing facial expressions from 70 individuals (35 males and 35 females). Each person exhibited seven unique emotional states, and all images are standardized to a resolution of 72 by 72 pixels.

4.2.2 RAF-DB [17]

The RAF-DB dataset is a large-scale collection of facial images sourced from the Internet and designed to reflect real-world conditions. It includes 29,672 images, each carefully labeled with seven basic emotions: happiness, sadness, surprise, fear, anger, disgust, and neutrality, as well as 11 more nuanced compound emotional states. A total of 12,271 images is allocated for training, with the following distribution: happiness (4,772), sadness (1,982), surprise (1,290), fear (281), anger (705), disgust (717), and neutral expressions (2,524). For testing and evaluation of facial emotional recognition systems, an additional 3,068 samples are provided.

4.3 Data preprocessing

The preprocessing step starts by resizing all input images to a consistent resolution, confirming uniformity in the dataset, and compatibility with the model architecture. Facial regions are then detected and cropped using bounding box techniques to isolate the most informative areas. To improve visual clarity, Histogram Equalization is applied, which improves contrast and highlights subtle features. Pixel values are normalized to a standard range, typically between 0 and 1, to stabilize the training and reduce the sensitivity to illumination variations. To increase the dataset range and improve generalization, augmentation techniques, such as rotation, flipping, and scaling, are used. Finally, the preprocessed images are converted into tensors and organized into batches, ready for input into the feature extraction model.

4.4 Network Structure of Emotion Recognition

The Network Structure of Emotion Recognition introduces the key components of the proposed model, which is an enhanced EfficientNet50 architecture for improving facial emotion recognition by integrating a Coordinate Attention (CA) Block that strengthens the models' ability to capture both spatial and channel-wise dependencies.

4.4.1 Coordinate Attention Block

Coordinate Attention operates in two phases: first, it changes how information is fixed, and second, it improves how attention is generated. This method enables precise location data to support both channel- and long-range interactions. While global pooling is commonly used in channel attention, it often struggles to retain spatial detail. To address this issue, the coordinate attention block breaks down global pooling into two one-dimensional encoding steps. It uses average pooling along the height and width (H,1) and (1, W) to capture the features across the horizontal and vertical axes. This results in two direction-sensitive feature maps that retain positional accuracy in one direction while capturing long-range patterns in the other direction. By inserting the CA mechanism into the EfficientNet50 backbone, the model gained a refined ability to preserve the spatial context while enhancing channel-wise feature representation. This integration is particularly beneficial for emotion recognition tasks where subtle facial signs, such as micro-expressions and localized muscle movements, require both precise spatial localization and robust semantic abstraction. In the modified architecture shown in Figure 2, the CA blocks are strategically inserted after the selected MBConv layers within EfficientNet50. These blocks decompose the global context into two orthogonal attention maps: one for encoding horizontal dependencies and the other for encoding vertical dependencies. Unlike traditional channel attention mechanisms, which depend on global average pooling and risk-losing fine-grained positional signs, CA maintains directional sensitivity by applying separate pooling operations along the height and width dimensions. This produces two complementary feature maps that encode long-range dependencies while preserving spatial precision.

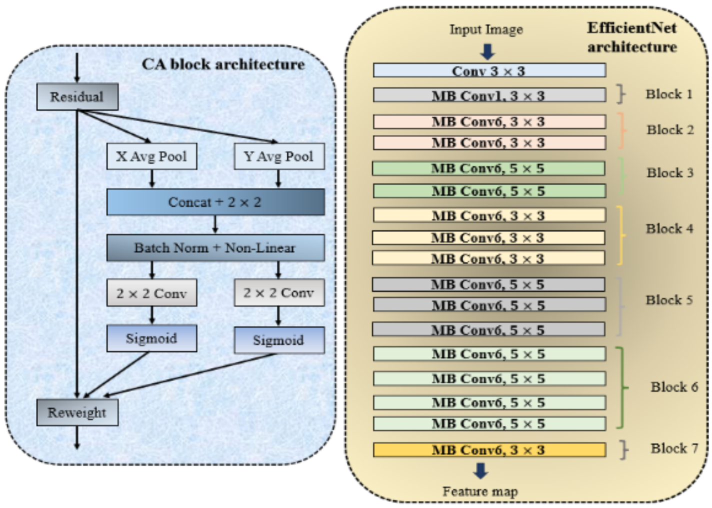

Fig. 2. Architecture representing coordinate attention with efficientnet.

4.4.2 EfficientNet

EfficientNet is a neural network architecture known for its high efficiency and builds on residual networks to deepen the model and improve the feature extraction. Adjusting the number of feature layers and adding extra extraction stages improved this capability. It also balances the input image size to help the network learn richer representations and boost the accuracy. EfficientNet achieves a strong balance between performance and precision through the compound scaling of depth, width, and resolution. The Mobile Inverted Bottleneck Convolution (MBConv) is first introduced in MobileNetV2, plays an essential role in the EfficientNet architecture. This design significantly reduces the computational overhead by approximately a factor of a^2compared with conventional convolutional layers. This efficiency is achieved by leveraging shortcut connections between bottleneck layers that involve fewer parameters than expansion layers and depth-wise separable convolutions. Here, a denotes the kernel size that defines the height and width of the 2D convolutional window. EfficientNet's primary strength is its compound scaling strategy, which systematically balances the depth, width, and resolution to enhance convolutional neural networks (CNNs). This method not only improves feature extraction, but also ensures robust emotion classification across varying input conditions. Additionally, EfficientNet demonstrates strong performance when processing high-resolution images, making it suitable for detailed visual tasks. In the deeper layers of the CNN, the architecture is designed to extract distinctive facial features, which are then classified using a SoftMax layer. The model architecture is customizable, allowing users to define the number of convolutional and fully connected layers and to include components such as batch normalization, dropout, and max-pooling as needed. Given the dynamic nature of facial expressions and features, it is crucial that the attributes selected for recognition remain stable over time. While numerous high-level feature representations have been explored in prior research, this approach focuses on the internal layers of Inception Module C within the InceptionV3 architecture. A multi-feature fusion technique is employed by combining intermediate feature maps from Module C with the final feature vector. These internal layers are systematically evaluated to determine which configuration yielded the most effective performance during fusion.

5 Experimental Results

The proposed Co-attention enhanced EfficientNet model runs on a Windows 10 system using PyTorch as the DL framework. All coding and model development were performed in the PyCharm environment. Training began with a learning rate of 0.0001 and a batch size of 16, using stochastic gradient descent to optimize the parameters. For the performance evaluation, the model is tested on a GPU, CPU, and Jetson Nano. Jetson Nano includes a quad-core ARM Cortex-A57 processor, 512 CUDA cores, and 8 GB of memory, making it suitable for lightweight deployment. The model achieved 32 frames per second on a GPU, 14 s on a CPU, and 19 s on Jetson Nano, showing better efficiency on different platforms. The proposed model supports real-time facial emotion recognition and allows for smooth operation in resource-constrained environments.

5.1 Evaluation metrics

To evaluate the performance of the proposed emotion recognition models in virtual learning environments, classification metrics re used, which helped measure the effectiveness and identify and distinguish emotional states from facial expressions. Accuracy reflects the proportion of correct predictions among all the predictions made. Precision indicates how many of the predicted emotional labels are correct, focusing on the reliability of the positive predictions. Recall shows how many actual emotional instances were successfully detected, emphasizing the model's sensitivity. The F1 score combines precision and recall into a single value, offering a balanced measure that is especially useful when dealing with uneven class distributions and computed using Equations (1), (2), (3), and (4). Together, these metrics provide a clear understanding of the model's strengths and limitations in recognizing emotions, such as dullness, happiness, confusion, and neutrality in faces.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

5.2 Data Splitting

To ensure consistent evaluation and reasonable comparison across models, both the KDEF and RAF-DB datasets are divided into training, testing, and validation sets in an 80:10:10 ratio. This balanced split supports robust learning while preserving sufficient data for an unbiased performance assessment, as shown in Figure 3.

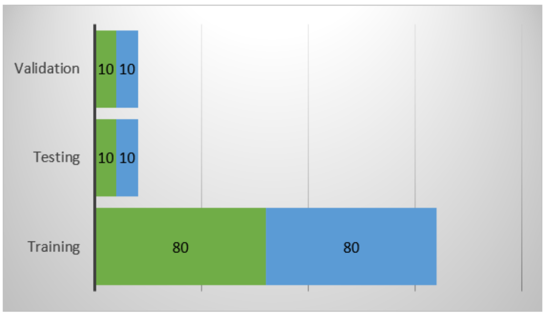

Fig. 3. Graph showing data split for both KDEF and RAF-DB dataset.

5.3 Performance Evaluation

To measure the practical capability of the proposed emotion recognition model in virtual learning environments, a complete performance evaluation is conducted for the datasets used. The proposed Co-Attention Enhanced EfficientNet provides better results than other baseline models. The results not only validate the model's robustness, but also demonstrate its suitability for present deployment in resource-constrained educational settings, as shown in Tables 1 and 2.

Table 1. Evaluation of proposed method comparing with Baseline models by considering KDEF dataset.

Dataset	Model	Accuracy	Precision	Recall	F1 Score
KDEF	ViT-Tiny	92.13	91.54	91.42	92.07
	VGG19	91.05	90.65	90.72	91.45
	MobileNet	94.78	93.85	93.75	94.65
	InceptionV3	95.23	94.01	94.32	95.17
	EfficientNet + CA (proposed)	97.86	96.25	96.32	97.54

Table 2. Evaluation of proposed method comparing with Baseline models by considering RAF-DB dataset.

Dataset	Model	Accuracy	Precision	Recall	F1 Score
RAF-DB	ViT-Tiny	91.25	90.36	90.41	91.05
	VGG19	94.65	93.75	93.62	94.57
	MobileNet	93.45	92.06	92.18	93.28
	InceptionV3	95.84	94.45	94.36	95.75
	EfficientNet + CA (proposed)	98.57	97.46	97.56	98.63

5.4 Complexity and Statistical Analysis

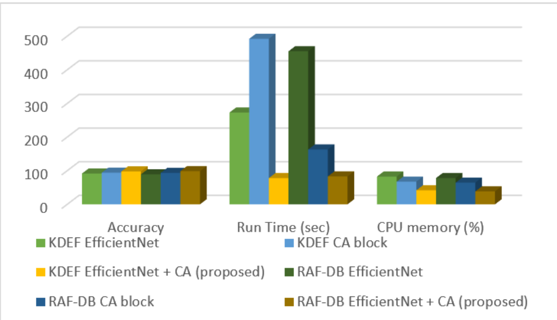

Fig. 4. Representation of accuracy, run time and memory values for both datasets.

Figure 4 illustrates the representation of accuracy, run time and CPU memory of the proposed RAF-DB EfficientNet + CA deployed on a GPU, CPU, and Jetson Nano. This highlights the limits between accuracy, runtime, and memory usage, highlighting the efficiency of the proposed EfficientNet + CA model. To validate the performance difference between the proposed Co-Attention Enhanced EfficientNet model and the baseline models, a paired t-test was conducted on the accuracy scores across multiple emotion classes. The paired t-test yielded a t-value of 3.14 with a p-value of 0.006, indicating a statistically significant improvement at the 0.01 level. This suggests that the proposed model consistently performed better than the baseline in recognizing emotions.

5.5 Ablation Study

To evaluate the contribution of individual components to the proposed method, an ablation study is performed, focusing on architectural and preprocessing variations. Histogram equalization is included to test its role in enhancing contrast and revealing subtle facial features, especially under inconsistent lighting common in virtual classrooms. EfficientNet serves as the backbone for feature extraction, capturing essential facial patterns while keeping the model light weight. To enhance sensitivity to spatial and semantic details, Coordinate Attention (CA) blocks are integrated. MobileNet layers have been explored for their compact architecture, offering a trade-off between speed and performance. Together, these components provide better results that not only identify emotions accurately, but also adapt to the limits of virtual classrooms, enabling dynamic feedback and adapted learning experiences, as presented in Tables 3 and 4.

Table 3. An ablation study to evaluate each component of proposed by considering KDEF dataset.

Dataset	Components	Accuracy	Precision	Recall	F1 Score
KDEF	EfficientNet	91.45	90.87	90.14	91.37
	MobileNet	93.57	92.01	92.00	93.18
	MobileNet + EfficientNet	95.64	94.17	94.27	95.48
	Histogram Equalization (preprocessing)	90.78	89.64	89.45	90.12
	EfficientNet + CA (proposed)	97.86	96.25	96.32	97.54

Table 4. An ablation study to evaluate each component of proposed by considering RAF-DB dataset.

Dataset	Components	Accuracy	Precision	Recall	F1 Score
RAF-DB	EfficientNet	92.54	91.35	91.24	92.41
	MobileNet	95.17	94.24	94.17	95.12
	MobileNet + EfficientNet	96.54	95.89	95.71	96.34
	Histogram Equalization (preprocessing)	89.45	88.76	88.85	89.24
	EfficientNet + CA (proposed)	98.57	97.46	97.56	98.63

5.6 Comparison Study

Table 5 shows comparison range of existing models each chosen for its unique strengths in facial emotion recognition with proposed model. The results consistently show that the proposed method achieves higher accuracy, precision, recall, and F1 score across both KDEF and RAF-DB datasets in emotion recognition for virtual learning environments.

Table 5. Comparison of proposed method with existing methods for both used dataset.

Dataset	Model	Accuracy
KDEF	ResNet50+CBAM+TCN [13]	97.08
	ResNet50+CBAM [14]	96.27
	EfficientNet + CA (proposed)	97.86
RAF-DB	ResNet50+CBAM+TCN [13]	91.86
	ResNet50+CBAM [14]	86.72
	EfficientNet + CA (proposed)	98.57

5.7 Discussion

The proposed Co-Attention Enhanced EfficientNet model shows better performance in emotion recognition for virtual learning by achieving accuracy, precision, recall, and F1 scores than existing models across both the KDEF and RAF-DB datasets. The integration of Coordinate Attention helps the model capture fine spatial and semantic features, which improves its ability to detect subtle emotions, such as frustration or engagement. The model runs efficiently on a GPU, CPU, and Jetson Nano, making it suitable for use in classrooms. Compared to baseline models, such as ResNet50-CBAM and MobileNet, the proposed method performs better in terms of both recognition and resource usage. These results confirm that the model is consistent, fast, and practical for adapting online learning environments to student emotional states.

6 Conclusion

The proposed Co-Attention Enhanced EfficientNet is presented to address the limitations of existing emotion recognition models within virtual learning environments that fail to capture subtle and complex facial emotions of students. An enhanced EfficientNet architecture, integrated with a Coordinate Attention (CA) block, shows better performance for emotion classification. The CA mechanism is particularly effective in preserving crucial positional information through one-dimensional encoding, which is important for identifying subtle facial signs. By combining EfficientNet's scaling strategy and Mobile Inverted Bottleneck Convolution (MBConv) layers, a robust balance between high accuracy and computational efficiency was achieved. The capability of the model for effective operation in various computing environments, including resource-constrained devices, underscores its practical applicability. The results provide superior performance for developing a more responsive and adopted virtual learning system that adapts instructional strategies to student emotional states, thus improving the educational experience. Future work includes expanding the model's capabilities to incorporate multimodal data, such as voice tone and body language, for a more comprehensive and holistic understanding of students' emotional states. Integrating additional data streams can further improve system robustness and the ability to conclude the emotional context in diverse learning scenarios.

References

1. R.A. Jaswal, S. Dhingra, Empirical analysis of multiple modalities for emotion recognition using convolutional neural network. Meas. Sens. **26**, 100716 (2023). https://doi.org/10.1016/j.measen.2023.100716
2. H.M. Shahzad, S.M. Bhatti, A. Jaffar, S. Akram, M. Alhajlah, A. Mahmood, Hybrid facial emotion recognition using CNN-based features. Appl. Sci. **13**, 5572 (2023). https://doi.org/10.3390/app13095572
3. M. Hussain, H.A. AboAlSamh, I. Ullah, Emotion recognition system based on two-level ensemble of deep-convolutional neural network models. IEEE Access **11**, 16875-16895 (2023). https://doi.org/10.1109/ACCESS.2023.3245830
4. M.C. Gursesli, S. Lombardi, M. Duradoni, L. Bocchi, A. Guazzini, A. Lanata, Facial emotion recognition (FER) through custom lightweight CNN model: performance evaluation in public datasets. IEEE Access **12**, 45543-45559 (2024). https://doi.org/10.1109/ACCESS.2024.3380847
5. Y. Huang, D. Bo, Emotion classification and achievement of students in distance learning based on the knowledge state model. Sustainability **15**, 2367 (2023). https://doi.org/10.3390/su15032367
6. Z. Huang, Y. Ma, R. Wang, W. Li, Y. Dai, A model for EEG-based emotion recognition: CNN-BI-LSTM with attention mechanism. Electronics **12**, 3188 (2023). https://doi.org/10.3390/electronics12143188
7. H.D. Le, G.S. Lee, S.H. Kim, S. Kim, H.J. Yang, Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. IEEE Access **11**, 14742-14751 (2023). https://doi.org/10.1109/ACCESS.2023.3244390
8. E.E. Arslan, M.F. Akşahin, M. Yilmaz, H.E. Ilgın, Towards emotionally intelligent virtual environments: classifying emotions through a

biosignal-based approach. Appl. Sci. **14**, 8769 (2024). https://doi.org/10.3390/app14198769

9. D. Ngo, A. Nguyen, B. Dang, H. Ngo, Facial expression recognition for examining emotional regulation in synchronous online collaborative learning. Int. J. Artif. Intell. Educ. **34**, 650-669 (2024). https://doi.org/10.1007/s40593-023-00378-7
10. J. Qi, H. Tang, Z. Zhu, Exploring an affective and responsive virtual environment to improve remote learning. Virtual Worlds **2**, 53-74 (2023). https://doi.org/10.3390/virtualworlds2010004
11. S.G. Rajesh, S.V. Madangarli, G.S. Pisharady, R. Subrahmanyam, Enhancement of virtual assistants through multimodal AI for emotion recognition. IEEE Access. **13**, 102159-102179 (2025). https://doi.org/10.1109/ACCESS.2025.3577664
12. M. Shomoye, R. Zhao, Automated emotion recognition of students in virtual reality classrooms. Comput. Educ.: X Reality **5**, 100082 (2024). https://doi.org/10.1016/j.cexr.2024.100082
13. M. Aly, Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model. Multimed. Tools Appl. **84**, 12575-12614 (2025). https://doi.org/10.1007/s11042-024-19392-5
14. M. Aly, A. Ghallab, I.S. Fathi, Enhancing facial expression recognition system in online learning context using efficient deep learning model. IEEE Access **11**, 121419-121433 (2023). https://doi.org/10.1109/ACCESS.2023.3325407
15. C. Gautam, K.R. Seeja, Facial emotion recognition using handcrafted features and CNN. Procedia Comput. Sci. **218**, 1295-1303 (2023). https://doi.org/10.1016/j.procs.2023.01.108
16. KDEF dataset link: KDEF Database
17. RAF-DB dataset link: RAF-DB DATASET