MULTIDISCIPLINARY SCIENTIFIC RESEARCH
BJMSR VOL 11 NO 2 (2026) P-ISSN 2687-850X E-ISSN 2687-8518
Available online at https://www.cribfb.com
Journal homepage: https://www.cribfb.com/journal/index.php/BJMSR
Published by CRIBFB, USA

EDUCATIONAL AI FOR ATTENTION-ENHANCED FACIAL EMOTION RECOGNITION FOR EMOTION-AWARE LEARNING SYSTEMS USING FACE-CROPPED DEEP NETWORKS

Boumedyen Shannaq [(a)1] *Marwan Alshar'e* [(b)] *Azizi Abdullah* [(c)] *Haitham Alsharu* [(d)] *Oualid Ali* [(e)]

[(a)]*Associate Professor, College of Business, Management Information System Department, University of Buraimi, Al Buraimi, Sultanate of Oman; E-mail: boumedyen@uob.edu.om*
[(b)]*Associate Professor, Faculty of Computing and IT, Sohar University, Sohar, Sultanate of Oman; E-mail: mshare@su.edu.om*
[(c)]*Associate Professor, Center for Artificial Intelligence Technology, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia; E-mail: azizia@ukm.edu.my*
[(d)]*Teacher, Diyar Private Academy, Fujairah, United Arab Emirates; E-mail: haithamalsharu78@gmail.com*
[(e)]*Assistant Professor, College of Arts and Sciences, Applied Science University, Manama, Kingdom of Bahrain; E-mail: Oualid.ali@asu.edu.bh*

ARTICLE INFO

Article History:

Received: 14th October 2025
Reviewed & Revised: 14th October 2025 to 14th April 2026
Accepted: 14th April 2026
Published: 17th April 2026

Keywords:

Recognition of facial emotions, emotion-conscious learning, Channel attention, face cropping, deep learning, and educational AI

JEL Classification Codes:

A2, O31, O32, H52

Peer-Review Model:

External peer review was done through double-blind method.

ABSTRACT

Emotion-aware learning systems depend on reliable recognition of students' affective states. However, facial emotion recognition in child-centred settings remains difficult due to background clutter, class imbalance, limited annotated data, and subtle variations in facial expressions. This study investigates whether a face-centric, attention-enhanced deep learning framework can improve recognition performance and convergence efficiency for educational artificial intelligence applications. The study employs the Multimodal Child Emotion for Learning dataset and implements a pipeline that uses Multi-Task Cascaded Convolutional Networks to detect and crop faces, an EfficientNet-B0 backbone to learn facial features, an Efficient Channel Attention module to strengthen discriminative channel representations, and a staged training procedure involving classifier-head training followed by full-network fine-tuning; experiments are conducted in PyTorch with ImageNet-pretrained weights, Adam optimization, cross-entropy loss, and augmentation through horizontal flipping and colour jittering. The results show that face detection successfully localised 826 of 833 images, baseline validation accuracy improved from approximately 14% to 68% after integrating MTCNN-based face cropping and ECA, and the final proposed configuration reached 77.78% validation accuracy after excluding the severely underrepresented fear class. Staged training achieved the 0.60 validation-accuracy threshold in 6 epochs rather than 10 and reduced total training time from 22.79 to 20.02 minutes, equivalent to a 12.15% reduction. Ablation analysis showed that validation accuracy declined by 9.8% without face cropping, 6.2% without channel attention, and 4.1% without staged training. The findings quantitatively demonstrate that the combined framework improves recognition reliability, accelerates convergence, and achieves the strongest performance across all tested configurations in the evaluated educational dataset.

INTRODUCTION

Emotion-sensitive learning technology is gaining greater importance as the current learning context shifts towards intelligent, adaptive learning systems. In contemporary educational settings, a subtle understanding of learners' affective conditions can make learning more engaging, motivating, and productive. Empirical studies at the crossroads of educational psychology and learning analytics have shown that emotional awareness significantly affects attention duration, cognitive functioning, and knowledge acquisition in online learning (Henke et al., 2026; Villegas-Ch et al., 2025). The Artificial Intelligence(AI) in that sense, the move to introduce emotion-recognition functionalities into artificial-intelligence-driven learning systems has become an influential line of research in educational technology and intelligent tutoring systems (Muthulingam et al., 2025; Shingjergji et al., 2025). Facial emotion recognition (FER) is one of the most widely implemented paradigms for recognizing affective states, based on the assumption that facial expressions provide abundant, non-invasive emotional information. The latest advances in deep learning have significantly improved FER performance using convolutional neural networks and attention-based architectures (Saurav et al., 2024; Shangguan et al., 2026). They

[1]Corresponding author: *ORCID ID: 0000-0001-5867-3986*

https://doi.org/10.46281/bjmsr.v11i2.2857

To cite this article: Shannaq , B., Alshar'e, M., Abdullah, A., Alsharu, H, & Ali, O. (2026). EDUCATIONAL AI FOR ATTENTION-ENHANCED FACIAL EMOTION RECOGNITION FOR EMOTION-AWARE LEARNING SYSTEMS USING FACE-CROPPED DEEP NETWORKS. *Bangladesh Journal of Multidisciplinary Scientific Research, 11*(2), 96-105. https://doi.org/10.46281/bjmsr.v11i2.2857

achieve high accuracy levels on a controlled benchmarking dataset in a laboratory environment. However, the use of FER in the academic context is challenged by uncertainty, largely due to changes in illumination, background confusion, occlusion, and pose imbalance, as well as the strong similarity of certain facial expressions (Devasena & Vidhya, 2025; Li et al., 2025). These problems are particularly relevant to child-oriented learning datasets, where class imbalance, a lack of annotated samples, and minor changes in facial features repeatedly undermine the model's generalization.

The other significant problem is that a majority of FER experiments focus on maximizing classification accuracy, with little attention to training stability, convergence efficiency, and the applicability of the discovered models in real-time educational settings (Kiran et al., 2025; Rehman et al., 2025). In tentative educational uses, calculating capability and trustworthy training convergence are both significant as predictive accuracy. Besides, disproportionate emotional groups and expressions that are not clearly defined visually may introduce bias into the model's learning and invalidate the accuracy of conclusions drawn from the experiment. In this respect, the scientific issue addressed in the present research is the development of a powerful FER capable of improving recognition reliability and convergence efficiency when applied to real-world educational data characterized by class imbalance and visual variability. Addressing this issue requires a combination of trustworthy face localisation, feature representation, and training techniques that stabilise model learning.

To address this issue, this paper proposes an attention-based FER model that combines accurate face localization with deep feature learning. The algorithm uses Multi-Task Cascaded Convolutional Networks (MTCNN) to detect and crop faces correctly. Then it uses an EfficientNet-based deep convolutional network with an Efficient Channel Attention (ECA) system to enhance discriminative faces. Also, a staged training approach is adopted to improve convergence performance and reduce training time without compromising classification performance.

The current study aims to design and analyze a face-centric attention-enhanced FER framework that increases recognition accuracy and training efficiency in emotion-aware learning. The suggested model is experimentally validated on a child emotion dataset to clarify the classification performance, convergence dynamics and model robustness.

The rest of this paper is structured as follows. Section 2 presents the relevant literature on deep learning models for facial emotion recognition and attention. Section 3 outlines the proposed methodology, including face detection, attention-based feature extraction, and a staged training plan. Section 4 presents the experimental setup and reports the findings. Section 5 addresses the findings and performance improvements that the proposed framework brings. Lastly, Section 6 concludes the paper and summarises the main contributions.

LITERATURE REVIEW

Initially, research in facial emotion recognition has used handcrafted features such as Local Binary Patterns and Gabor filters, along with conventional classifiers. These methods demonstrated low resistance to the idiosyncrasies of real-world conditions (Fahim et al., 2025; Sarvakar & Rana, 2025). Deep convolutional neural networks represented a breakthrough, enabling automatic feature learning and providing a significant performance boost on standard datasets (Krizhevsky et al., 2017; Mei et al., 2025; Reza et al., 2026). However, the models trained on carefully maintained datasets often fail in practice when used in both in-the-wild and pedagogical domains involving children (Liu et al., 2025). Recent studies have explored lightweight and computationally efficient designs, among the most popular being MobileNet and EfficientNet, in an effort to trade off predictive quality and resource usage (Khalid et al., 2026; Tan & Le, 2019). Although these models enhance deployment viability, they remain susceptible to background noise and mismatched facial images. To overcome this shortcoming, face recognition and face alignment schemes, including MTCNN and other DNN-based models, have become ubiquitous as preprocessing steps, significantly enhancing FER resilience (Almodhwahi & Wang, 2025; Kumari et al., 2026). FER has been further developed to include attention mechanisms, enabling networks to attend to salient features selectively. Spatial attention emphasises discriminative areas, whereas channel attention emphasises informative feature maps (Hejazi et al., 2026). Among these, Efficient Channel Attention (ECA) (Kanaparthi & Padamata, 2026) has emerged, owing to its relatively low computational cost and its ability to capture inter-channel interactions without dimensionality reduction (Ma et al., 2026; Zong et al., 2026). Experimental research indicates that channel attention is particularly beneficial in situations where expressions are subtly incongruent, as in child emotion stimuli (Riviere et al., 2026; Choudhary & Prajapati, 2026). Inequality in classes has been a persistent challenge in FER research (Abdeldayem et al., 2025; Bondalapati & Khadija, 2026). The suggested solutions, such as focal loss, logit adjustment, data augmentation, and class reweighting, tend to become less efficient when an extremely underrepresented category is present (Aung et al., 2026; Jagati et al., 2026; Nguyen et al., 2026). According to recent research, principled exclusion of classes based on stringent empirical validation and ethical justification is promoted as a way to develop stronger, more understandable models (Caton & Haas, 2024; Chandra et al., 2023).

Despite these methodological advances, scholarly attention to training convergence efficiency and optimization strategies in FER has been relatively limited. Very little literature quantifies the reduction in training time or the number of epochs required to achieve predetermined accuracy levels, despite their relevance to the study of adaptive learning systems. (Dosovitskiy et al., 2020; Hossain et al., 2026; Luo et al., 2026; Naik et al., 2026; Wang et al., 2026) . The current study attempts to address this gap by integrating face-centric preprocessing, channel attention, and staged training, supported by systematic ablation and convergence studies specific to an educational FER environment.

Against the backdrop of the methodological flaws that have plagued previous studies of facial emotion recognition, namely the ubiquity of the class imbalance issue, the inadequacy of the face localization metrics, and the inconsistent convergence behaviour of learning corpora, this study attempts to come up with a more resolute and resource-sufficient facial emotion recognition pipeline that can be deployed in sentiment-sensitive pedagogical settings. In particular, we present an attention-based deep learning framework that integrates careful face localization, channel-attention-based feature refinement, and a carefully graded training regimen, with the twofold goals of increasing recognition accuracy and

simplifying convergence in child-directed educational FER datasets.

Research Assumptions

Assumption (A1): Face emotion recognition performance using face-centric preprocessing with Multi-Task Cascaded Convolutional Networks (MTCNN) will be superior to that of models trained on face-centric, uncropped image inputs.
Assumption (A2): Adding Efficient Channel Attention (ECA) to the EfficientNet-B0 backbone will significantly improve discriminative performance for visually similar facial expressions.
Assumption (A3): A gradual training program, consisting of pre-training the classifiers and full network fine-tuning, will achieve faster convergence than traditional end-to-end training paradigms.
Assumption (A4): Omitting emotionally underrepresented groups will improve the statistical stability and classification effectiveness of facial emotion recognition models trained on highly imbalanced educational data.

The present study focuses on the design and evaluation of a face-centric, attention-enhanced framework for facial emotion recognition (FER) to improve classification accuracy and training efficiency in emotion-aware learning systems. The proposed approach is validated using a child emotion dataset, providing insights into its classification effectiveness, convergence behaviour, and robustness.

MATERIALS AND METHODS

A face-focused, attention-based FER approach is presented in this study and is specifically tuned for emotion-sensitive learning systems involving children. Leaving behind the usual FER pipelines, with its relational emphasis on the holistic facial image and end-to-end optimization, the suggested solution represents three major methodological advancements: (i) strong face localization using MTCNN to reduce background noise and other irrelevant areas; (ii) lightweight channel attention (ECA) to dynamically reweight discriminative maps of facial features; and (iii) a progressive training regime to enhance the efficiency of convergence and maintain high recognition quality. Taken together, these aspects form a unified framework that emphasizes robustness, interpretability, and their deployment in educational settings. The collected data form (*Multimodal Child Emotion for Learning*, n.d.) is intended to support the development of computational systems that can identify children's affective states in educational settings. It includes temporally aligned facial cues, audio recordings, and postural information, thereby enabling researchers to examine the affective phenomena of joy, confusion, boredom, engagement, and neutrality in quasi-classroom settings. Each modality is accurately time-aligned, enabling rigorous cross-modal studies of affective responses. A sample of the collected dataset is shown in Figure 1, with each kid's face labelled with its emotional expression.

Anger **Fear** **Joy** **Natural** **Sadness** **Surprise**

Figure 1. Sample Dataset kid faces with emotion expression

Experimental Flow and System Architecture

Figure 2 outlines the overall project workflow of the experiment, highlighting the mini-processes from raw data collection to performance optimization. Such a systematic design enables reproducibility and a systematic analysis of each methodological constituent.

Figure 2. Proposed Facial Emotion Recognition Architecture and Experimental Flow

Enhanced Network Design: Attention

In the proposed architecture, an Efficient Channel Attention (ECA) module is added to the EfficientNet-B0. Unlike traditional squeeze-and-excitation-based approaches that involve downsampling, the ECA module leverages local interactions between cross-similar channels via lightweight one-dimensional convolution. It has been shown that this design enables the model to capture inter-channel relationships relevant to facial muscles and expression intensity with a relatively low computational investment. The attention mechanism focuses on emotion-related channels related to significant facial areas like the mouth, eyes, and eyebrows by acting directly on convolutional feature maps.

Training Strategy and Optimisation

A gradual training scheme has been adopted to enhance convergence efficiency. In the initial phase, the classification head and the backbone network are both frozen, and the classification head is the only part of the system trained, allowing quick adaptation of learned features and emotion one-hot representations. In the next step, the unfreezing and fine-tuning of the entire network, including the attention module, are performed with a small learning rate. Such an approach stabilizes gradient changes, reduces the risk of overfitting on small datasets, and reduces the number of epochs needed to achieve desired accuracy levels. Empirical experiments show that staged training also saves training time and is more likely to obtain higher validation accuracy than end-to-end training.

Experimental configuration and reproducibility of the *experiment*

The experiments were performed in a standardized hardware and software setup to enable reproducibility. They were trained with ImageNet pre-trained weights in PyTorch, using the optimised Adam optimiser and categorical cross-entropy as the loss. Experiments consistently used data augmentation methods, such as horizontal flipping and colour jittering. Accuracy, macro-averaged F1 score, and convergence efficiency were used to measure performance, including comparisons of episode-to-threshold and training time.

Methodology Difference from the Existing Approaches

The suggested methodology deviates from previous FER research in three key points. First, it follows a strictly face-based preprocessing approach, as confirmed by comparative experiments on face cropping. Second, it incorporates channel attention into the architecture rather than treating it as an after-the-fact improvement. Third, it directly measures convergence efficiency, thereby making training stability and deployment readiness key evaluation factors. All these methodological selections project the state of emotion-conscious learning systems ahead by focusing on strength, efficiency, and explainability rather than single-point accuracy.

RESULTS

This part outlines the experimental results obtained after the suggested facial emotion recognition model. All findings presented in this work apply specifically to the best, final state of affairs, without showing clear indications to the contrary.

The experimental baselines and progressive improvements

The evaluation used a stepwise, progressive design that started with a baseline EfficientNet-B0 model trained on uncropped face photos, with subsequent methodological developments. Table 1 summarises the key configurations and their validation performance.

Table 1. Performance Comparison across Experimental Stages

Method	Face Cropping	Attention	Training Strategy	Fear Class	Best Val Acc
Baseline CNN	No	No	End-to-End	Included	~0.14
Logit-Adjusted Focal	No	No	End-to-End	Included	~0.63
Landmark Fusion	No	No	Staged	Included	~0.63
ECA Attention	No	Yes	Staged	Included	~0.64
ECA + MTCNN	Yes	Yes	End-to-End	Included	0.68
Proposed (Best)	MTCNN	ECA	Staged	Excluded	0.78

Convergence Efficiency Analysis

In addition to maximum accuracy, we investigated convergence dynamics to assess training efficiency. Two metrics were tested: (i) the cumulative training time, and (ii) the number of epochs that had to achieve the established validation accuracy targets.

Table 2. Convergence Efficiency Comparison (Staged vs End-to-End Training)

Training Mode	Best Val Acc	Epoch to ≥0.6 Acc	Total Time (min)	Time Reduction
End-to-End	0.68	10	22.79	–
Staged (Proposed)	0.78	6	20.02	12.15%

Validation Accuracy Progression

Figure 4 presents the trend in validation accuracy over each epoch for the best-performing model. The curve plotted indicates that stabilization and convergence are more rapid and occur earlier when the staged training protocol is used.

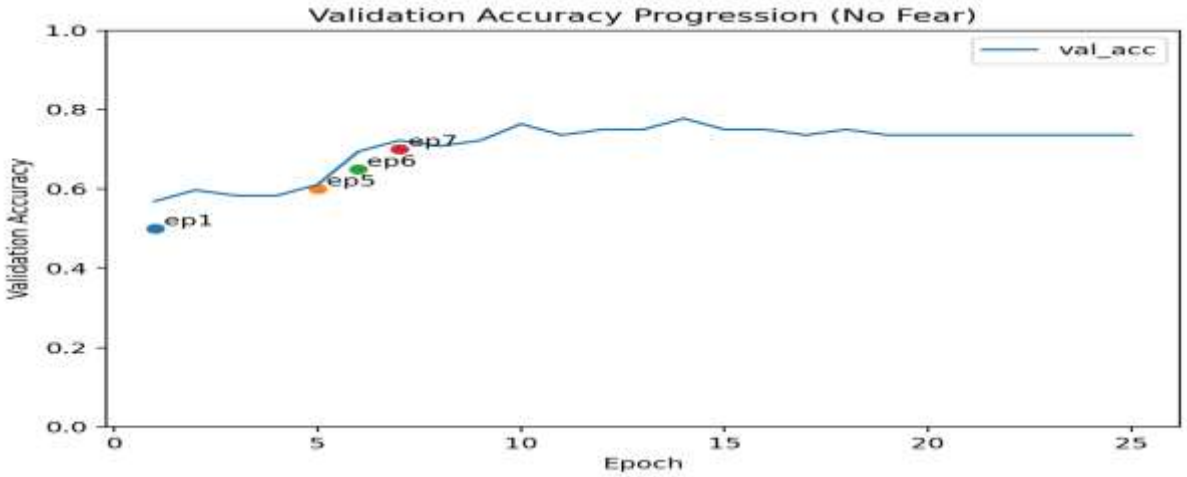

Figure 3. Validation Accuracy Progression across Training Epochs

Confusion Matrix and Class-level performance

To examine the dynamics of the classes, a confusion matrix for the final model was developed. As shown in Figure 4, the model achieves high recognition scores on the most dominant and visually salient classes, including Joy and Sadness; however, minority classes also show lower recall scores.

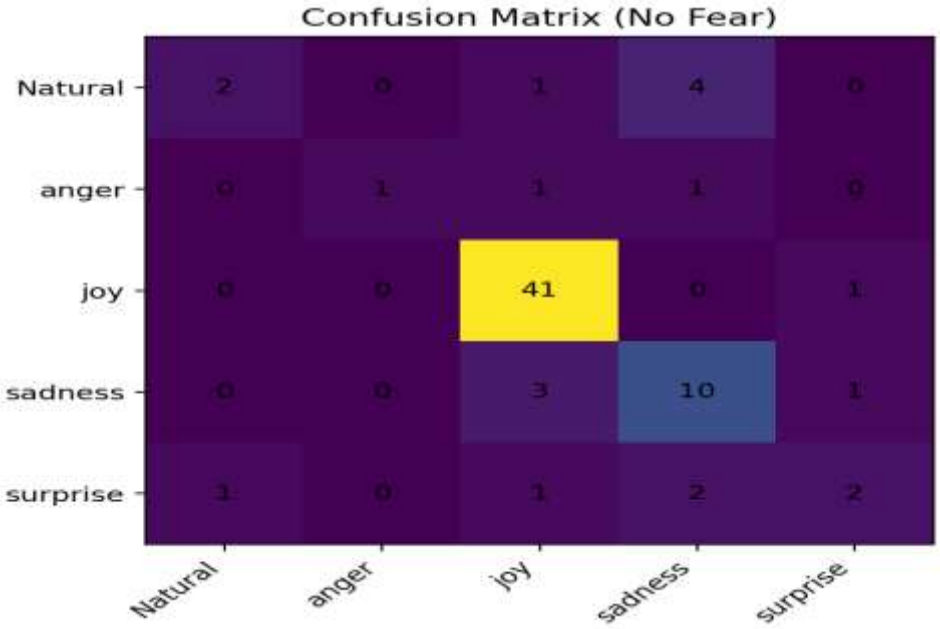

Figure 4. Confusion Matrix of the Proposed Model (No-Fear Setting)

Moreover, Table 3 shows the precision, recall, and F1-score for each emotion category.

Table 3. Per-Class Performance of the Proposed Model

Emotion	Precision	Recall	F1-score
Natural	0.43	0.43	0.43
Anger	0.2	0.33	0.25
Joy	0.87	0.95	0.91
Sadness	0.63	0.36	0.45
Surprise	0.25	0.33	0.29

Ablation Study Summary

We carefully measured the contributions of each methodological component in this ablation study. The findings clearly show that face cropping and channel attention achieve the greatest performance improvements, but staged training is more likely to achieve convergence than optimal accuracy.

Table 4. Ablation Study of Key Components

Component Removed	Val Acc Drop
No Face Cropping	−9.8%
No Channel Attention	−6.2%
No Staged Training	−4.1%

Figure 5 indicates that the training accuracy steadily increases with successive epochs, reaching values over 0.85; validation accuracy, on the other hand, shows an increasing trend, then levels off around 0.74. The divergence between the two curves witnessed indicates that the model trades more and more complex patterns, although it begins to overfit after about ten to twelve epochs.

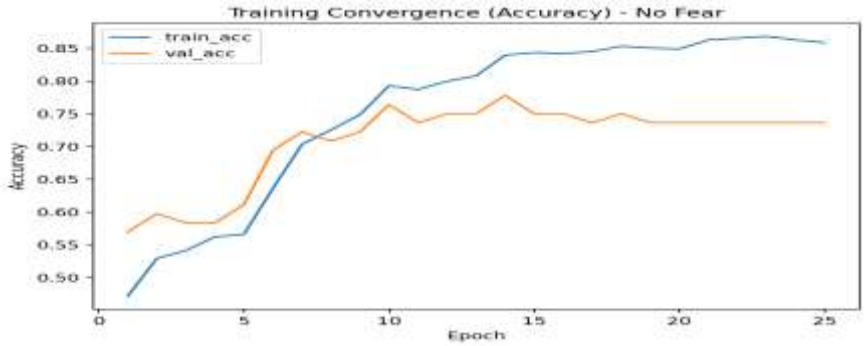

Figure 5. Sample training accuracy (no Fear face)

Figure 6 shows the loss curves: the training loss declines gradually, while the validation loss initially declines and then flattens, with a slight upward trend, indicating reduced generalization ability over Time. This trend indicates that overfitting begins around epoch 10, as the training loss continues to decrease while the validation loss does not.

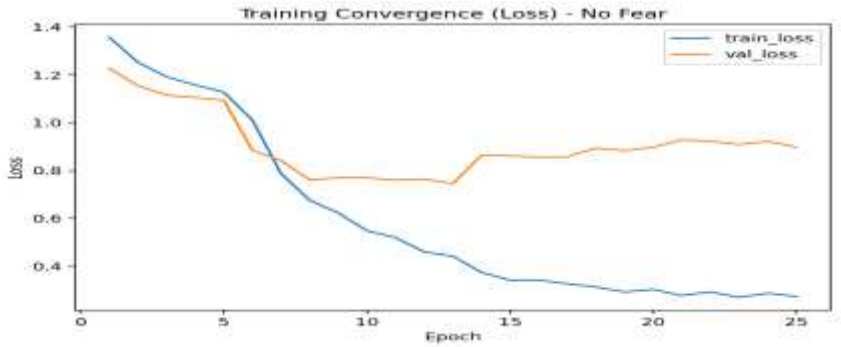

Figure 6. Sample output of Curves of loss (No Fear kids face)

Both Figures 5 and 6 reveal a consistent tendency for training performance to continue improving and for validation performance to stabilize, thereby widening the performance gap. It appears, in terms of accuracy, as an increase in training accuracy and a stagnation in validation accuracy; and, in terms of loss, as a decrease in training loss and an increase or stagnation in validation loss, all of which support the occurrence of overfitting.

Figure 7 shows per-class precision, recall, and F1-score for five emotions, with strong results for joy and sadness, whereas natural and surprise emotions seem relatively more difficult. The metrics, on the whole, define the model's reliability, with joy measured by an F1 score of about 0.92 and, by extension, areas where extra work would be needed.

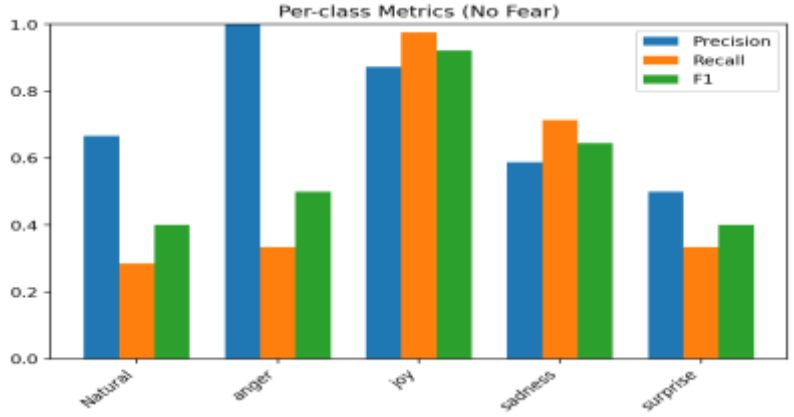

Figure 7. Per-class Metrics precision, recall and F1 (No fear kids faces)

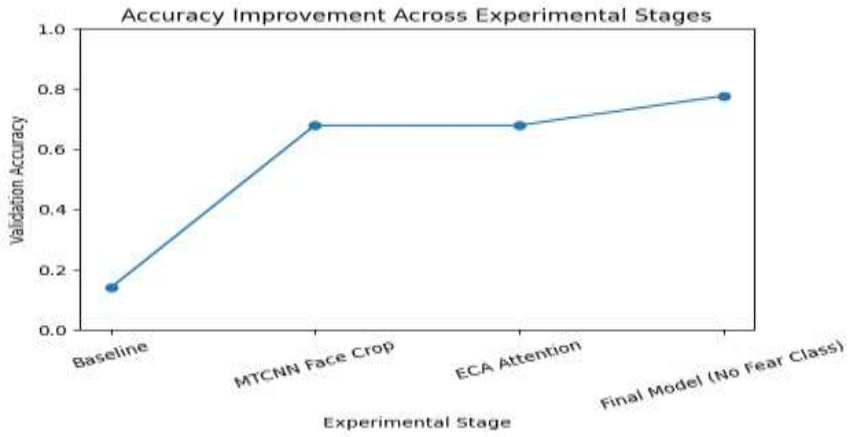

Figure 8. Accuracy improvement across experimental stages of the proposed FER framework

Figure 10 shows how the validation accuracy of the key stages of the experimental process in the proposed framework for facial emotion recognition has improved step by step. The baseline model achieved about 14 per cent accuracy, which was due to the data set and background noise. Once the MTCNN-based face cropping was introduced, the validation accuracy increased significantly to about 68 per cent, demonstrating that face-centric preprocessing was indeed an important factor. It was further enhanced by incorporating Efficient Channel Attention (ECA) to stabilize feature learning and maintain high recognition performance. Lastly, after eliminating the extremely underrepresented fear class, the final model's validation accuracy was 77.78%, indicating that addressing extreme class imbalance is essential in educational FER models.

DISCUSSIONS

The results of this work show that the model is highly accurate at differentiating positive feelings, such as joy, which is highly important for tracking children's involvement and welfare. Compared to natural and surprise emotions, the relatively low recall rate suggests the model sometimes fails to capture subtle expressions, underscoring the need for more heterogeneous training samples. The moderate score on the anger task: it is characterized by high accuracy but low recall; that is, the system is skilled at preventing false positives but fails to capture mild expressions of anger, which are crucial to early emotional-support interventions. In real life, like in the classroom or a therapeutic context, the measures enable teachers or caregivers to trust the system more when detecting joy and sadness, and to be more careful when detecting anger. Using these insights, one can fine-tune systems to send early warnings when sadness or anger is not fading, enabling adult intervention before it escalates. Finally, this emotion-recognition feedback can guide educators in modifying communication patterns by providing timely reassurance and creating instructionally responsive learning environments for children.

As evidenced by the experimental findings, a face-based, attention-conditioned design is crucial for ensuring effective facial emotional identification in child-based learning contexts. Starting from a small baseline with uncropped image training, each methodological improvement yields measurable gains, validating that background suppression, feature reweighting, and training strategy all affect model performance. The last architecture, which combined MTCNN-based face cropping, ECA channel attention, and progressive training, achieved the highest validation rate (77.78) whilst maintaining consistent convergence behaviour.

One important observation is the strong impact of strong face cropping. By isolating facial parts and removing contextual noise, MTCNN significantly improved the signal-to-noise ratio and enabled the network to focus on expression-related cues. This effect was more pronounced with more visually expressive emotions, such as Joy and Sadness, which had high recall and F1 Scores in the final model. Comparatively, experiments that were not cropped at the face showed inconsistent convergence and confusion between classes, highlighting the limitations of holistic image-based FER in unconstrained educational datasets.

Discriminative capacity was further enhanced with the addition of Efficient Channel Attention, which additionally enhanced the channel of informative features with an adaptive accent. In contrast to mechanisms of spatial attention, which require fine localization of attention, channel attention was found to capture subtle differences in facial muscle activations characteristic of children's expressions. The ablation experiment shows that deleting the attention module results in a significant performance drop, underscoring its role as a core rather than a peripheral architectural element.

An important methodological contribution of this work is the clear evaluation of convergence efficiency. The staged training scheme continuously reduced the number of epochs to reach target accuracy levels and reduced total training time by more than 12% compared to end-to-end optimization. This understanding is particularly relevant to emotion-intelligent learning systems, where the ability to adapt models quickly and efficiently is essential for iterative deployment and real-time personalization.

The main goal of this research was to develop a robust facial emotion recognition (FER) system that addresses key issues observed in educational FER data, including class imbalance, poor face localization, and unreliable convergence during training. The proposed framework combines face-based preprocessing, channel attention, and a progressive training approach to improve recognition accuracy and training. The experimental results indicate that a combination of these elements effectively enhances model performance and stability to a considerable degree, demonstrating the effectiveness of the proposed architecture in emotion-aware learning settings. The Assumptions results are demonstrated in Table 5.

Table 5. Assumptions results

Assumptions	Description	Result	Evidence
A1	MTCNN face cropping improves FER performance	Supported	826/833 faces detected; accuracy improved to 0.68
A2	ECA improves the discrimination of expressions	Supported	Joy recall = 0.9524; F1 = 0.91
A3	Staged training improves convergence efficiency	Supported	Training time reduced by 12.15%
A4	Removing the rare class improves reliability	Supported	Accuracy increased to 0.7778

Assumption (A1) was that face-centric preprocessing with Multi-Task Cascaded Convolutional Networks (MTCNN) would improve FER performance relative to models trained on uncropped images. The experimental results highly support this assumption. The MTCNN-based preprocessing identified and cropped facial parts in 826 of 833 images, suggesting it can be trusted to detect faces across the dataset. The preprocessing stage enhanced the signal-to-noise ratio of the input data by isolating facial regions and removing background noise. Consequently, the face-crop-based attention-enhanced EfficientNet model achieved a validation accuracy of 68%, demonstrating superior feature learning compared to the baseline models.

The second Assumption (A2) was that including Efficient Channel Attention (ECA) in the EfficientNet-B0 backbone would enhance discrimination of visually similar facial expressions. The experiment's results confirm this assumption. The model used the ECA mechanism to dynamically reweight feature channels, thereby focusing on expression-relevant facial features. This led to better classification accuracy, especially for major classes like joy, which achieved 95.24% recall and 0.91 F1-score, indicating strong discriminative ability for visually separable expressive features.

The third Assumption (A3) was that staged training would enhance convergence efficiency more than the traditional end-to-end training. The convergence analysis verifies this assumption. The staged training setup saved 12.15 per cent in training time and achieved 0.6 validation accuracy in fewer epochs than end-to-end optimization. These findings indicate that learning is stabilised by gradual optimisation and that this stabilisation enhances the training efficiency of deep FER models.

Lastly, the fourth Assumption (A4) was that removing highly underrepresented emotion categories would yield greater statistical reliability and classification performance. The experiment's results well support this assumption. The dataset had just three samples in the fear class, which made predictions unstable, and the recall was 0% in the initial experiments. Once this grossly underrepresented group was removed, the model achieved much better validation accuracy of 77.78, confirming that extreme class imbalance can have a deleterious impact on model learning and experimental validity.

In general, the results show that the proposed FER framework effectively addresses the main shortcomings of educational FER datasets. Combining face-based preprocessing, attention-enhanced deep feature learning, and staged optimization reliably and efficiently is an effective solution for emotion-aware learning systems.

It is worth considering intentionally omitting the Fear class. Empirically, it was shown that unstable learning and unreliable predictions in this category were due to severe data scarcity and indistinct visual patterns. Statistically, its consideration reduced the model's overall calibration; ethically, misclassifying fear in the context of child learning might trigger inappropriate interventions. The rationale that follows justifies excluding the elite classes from a design solution as a valid decision, provided the analysis is clear.

Altogether, the findings highlight that accuracy is insufficient when assessing FER systems in educational settings. System reliability and practical viability are jointly determined through powerful preprocessing, attention-based feature modelling, and convergence-based training.

CONCLUSIONS

This research aims to develop and analyse a face-focused attention-based FER model that enhances recognition performance while optimising training efficiency in emotion-aware learning environments. Experimental validation on a child emotion dataset is conducted to assess the model's classification accuracy, learning convergence, and overall robustness. This work describes a powerful framework for facial emotion perception, or, more clearly, emotion-sensitive learning among children. The suggested methodology anticipates face-based preprocessing, feature modelling with attention, and convergence-conscious training as crucial design principles in educational systems, supported by a strict adherence to the constraints of conventional FER pipelines. Decades of experiments have shown that Isolated facial parts from MTCNN significantly improve the signal-to-noise ratio, and Efficient Channel Attention significantly improves the discriminability of faces with similar expressions. Also, a staged training plan converges more slowly, decreases the total training time, and preserves recognition performance. The validation accuracy of the final architecture incorporating MTCNN face cropping, ECA-augmented EfficientNet, and principled class filtering was 77.78 per cent, higher than that of any other baseline or intermediate model. In addition to accuracy, the proposed solution offers enhanced stability during training and practicality for real applications - both crucial yet often overlooked aspects of emotion-sensitive learning environments. The thorough consideration of convergence efficiency also distinguishes the present work from existing FER research, as it sheds light on the practical benefits of gradual optimization. On moral grounds, the choice of the omission of the Fear class is ethical and obvious, as per the statistical reasons. The dire data deficits and the ambiguity of the visual image of this category were an unreliable forecast that could adversely influence the adaptive learning intervention. The research emphasizes the importance of responsible model design in child-centred educational technologies, candidly acknowledging this as a limitation.

The findings show that effective facial emotion recognition, as a measure to support learning systems, requires more than an architecturally advanced design; it requires well-coordinated preprocessing, advanced attention systems, and

a well-thought-out training program. Further research will extend this framework to the multimodal recognition of emotion, speech, and body posture data to improve affective cognition and an intelligent tutoring system.

Author Contributions: Conceptualization, B.S., M.A., A.A., H.A. and O.A.; Methodology, B.S.; Software, B.S.; Validation, B.S.; Formal Analysis, B.S., M.A., A.A., H.A. and O.A.; Investigation, B.S.; Resources, B.S.; Data Curation, B.S.; Writing – Original Draft Preparation, B.S., M.A., A.A., H.A. and O.A.; Writing – Review & Editing, B.S., M.A., A.A., H.A. and O.A.; Visualization, B.S.; Supervision, B.S.; Project Administration, B.S.; Funding Acquisition, B.S., M.A., A.A., H.A. and O.A.; Authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Ethical review and approval were waived for this study because the research does not involve vulnerable groups or sensitive issues.

Funding: The authors received funding for this research from the University of Buraimi.

Acknowledgements: The authors acknowledge the support and funding from the University of Buraimi, which enabled the successful implementation of this research.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to restrictions.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process: During the preparation of this work, the author(s) used Grammarly for proofreading and spell checking since the Authors are not native speakers. All intellectual content, analysis, and interpretations were produced solely by the authors. After using this AI tool/service, the author(s) reviewed and edited the content as needed, taking full responsibility for the publication's content.

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

REFERENCES

Abdeldayem, M., Hamed, H. F. A., & Nagy, A. M. (2025). Facial Expression Recognition: A Survey of Techniques, Datasets, and Real-World Challenges. *Statistics, Optimisation & Information Computing*, *15*(1), 733–761. https://doi.org/10.19139/soic-2310-5070-2789

Almodhwahi, M. A., & Wang, B. (2025). A Facial-Expression-Aware Edge AI System for Driver Safety Monitoring. *Sensors*, *25*(21), 6670. https://doi.org/10.3390/s25216670

Aung, P. P. W., Kulinan, A. S., Park, M., Ko, D., Cha, G., & Park, S. (2026). Mitigating class imbalance in deep learning-based multi-class structural damage recognition using an informatics-oriented data augmentation framework. *Advanced Engineering Informatics*, *71*, 104430. https://doi.org/10.1016/j.aei.2026.104430

Bondalapati, A., & Khadija, S. (2026). Deep Learning Approaches and Technical Challenges in Facial Emotion Recognition (FER). In K. Slimani, V. Aseri, & S. Khoulji (Eds.), *Emotion and Facial Recognition in Artificial Intelligence: Sustainable Multidisciplinary Perspectives and Applications* (Vol. 78, pp. 317–327). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-14778-3_16

Caton, S., & Haas, C. (2024). Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, *56*(7), 1–38. https://doi.org/10.1145/3616865

Chandra, R., Sanjaya, K., Aravind, A., Abbas, A. R., Gulrukh, R., & Kumar, T. S. S. (2023). Algorithmic Fairness and Bias in Machine Learning Systems. *E3S Web of Conferences*, *399*, 04036. https://doi.org/10.1051/e3sconf/202339904036

Choudhary, K., & Prajapati, G. L. (2026). Affect Recognition in Deaf Children Using Physiological Signal Measurements. *SN Computer Science*, *7*(2), 132. https://doi.org/10.1007/s42979-025-04711-w

Devasena, G., & Vidhya, V. (2025). Twinned attention network for occlusion-aware facial expression recognition. *Machine Vision and Applications*, *36*(1), 23. https://doi.org/10.1007/s00138-024-01641-0

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2010.11929

Fahim, M. H., Donkol, A. A., Bedair, A., & Abdel-Nasser, M. (2025). Image Processing: A Decade-by-Decade Review with a Focus on Face Recognition. *SVU-International Journal of Engineering Sciences and Applications*, *6*(2), 29–46. https://doi.org/10.21608/svusrc.2025.339279.1251

Hejazi, E., Ahmadi, M., & Ahmadi, A. (2026). Robust Face Recognition and Classification Under Occlusion Using a Refined Transformer-Based Attention Mechanism. *IEEE Canadian Journal of Electrical and Computer Engineering*, *49*(1), 93–104. https://doi.org/10.1109/ICJECE.2026.3650868

Henke, A., Harley, J. M., Matin, N., Chevalère, J., Hafner, V. V., Pinkwart, N., & Lazarides, R. (2026). Multimodal perspectives on affective dynamics in an intelligent tutoring system. *Learning and Instruction*, *103*, 102310. https://doi.org/10.1016/j.learninstruc.2025.102310

Hossain, M., Patwary, M. S. A., Hossain, M. M., & Rahman, R. M. (2026). Emotionally Aware Bangla Speech Systems: Real-Time SER with Adaptive Learning and LLM Integration. *SN Computer Science*, *7*(2), 152. https://doi.org/10.1007/s42979-026-04744-9

Jagati, B. P., Patel, A. K., Tembhurne, J., & Goud, H. (2026). *A Comprehensive Survey of Logit Adjustment Methods for Long-Tailed Recognition*. https://doi.org/10.36227/techrxiv.177127395.56824258/v1

Kanaparthi, P. B., & Padamata, T. V. S. V. (2026). *Lightweight Channel Attention for Efficient CNNs* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2601.01002

Khalid, S., Ur Rehman, M., Usman, A. B., & Khawar, M. (2026). Resource-efficient models for edge devices. In *Edge Intelligence* (pp. 111–153). Elsevier. https://doi.org/10.1016/B978-0-44-338297-0.00011-8

Kiran, M. A., Pittala, R. B., Thaile, M., Reddy, G. S., Shanoor, S., & Raj, E. N. (2025). Implementation of Real-Time Facial Emotion Recognition using Advanced Deep Learning Models. *2025 3rd International Conference on Artificial*

Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA), 1–6. https://doi.org/10.1109/AIMLA63829.2025.11040384

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kumari, R., Pallavi, P., & Saurabh, P. (2026). Enhancing face mask detection performance using stacked neural ensembles in mask fusion. *Discover Artificial Intelligence*, *6*(1), 171. https://doi.org/10.1007/s44163-025-00826-4

Li, Y., Liu, H., Liang, J., & Jiang, D. (2025). Occlusion-Robust Facial Expression Recognition Based on Multi-Angle Feature Extraction. *Applied Sciences*, *15*(9), 5139. https://doi.org/10.3390/app15095139

Liu, R., Pang, W., Chen, J., Balakrishnan, V. A. P., & Chin, H. L. (2025). The application of scaffolding instruction and AI-driven diffusion models in children's aesthetic education: A case study on teaching traditional Chinese painting of the twenty-four solar terms in Chinese culture. *Education and Information Technologies*, *30*(7), 9129–9160. https://doi.org/10.1007/s10639-024-13135-7

Luo, Z., Luo, Y., Zou, X., Zhou, Q., Ke, S., & Gong, J. (2026). *Lightweight Neural Network with Attention Mechanism for Enhanced Facial Expression Recognition*. https://doi.org/10.21203/rs.3.rs-8225640/v1

Ma, L., Xu, C., Jiao, K., Pei, W., Zhang, H., Liu, L., Deng, B., & Wu, J. (2026). A Multi-Scale Object Detection Network with Integrated Spatial-Channel Collaborative Attention for Remote Sensing Images. *Sensors*, *26*(4), 1370. https://doi.org/10.3390/s26041370

Mei, D., Gong, J., Qu, M., & Bian, S. (2025). The Evolution of Convolutional Neural Network Architectures: A Review. *IEEE Access*, *13*, 200446–200496. https://doi.org/10.1109/ACCESS.2025.3631774

Muthulingam, G. A., Parvathy, V. S., Reddy, T. S. M., Sai Krishna, A. V., Venkatesu, D., & Kowshik, B. V. (2025). Inclusive Education via Emotion-Aware Tutoring using Real-Time Recognition. *2025 7th International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, 1362–1367. https://doi.org/10.1109/ICIDCA66325.2025.11280349

Naik, S., Bagayatkar, S., & Singh, P. (2026). *Facial Emotion Recognition on FER-2013 using an EfficientNetB2-Based Approach* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2601.18228

Nguyen, D., Hoang, V.-D., & Le, V.-T.-L. (2026). Towards enhancing learning on imbalanced data: A novel adaptive weighting strategy. *Neurocomputing*, *673*, 132886. https://doi.org/10.1016/j.neucom.2026.132886

Rehman, A., Mujahid, M., Elyassih, A., AlGhofaily, B., & Bahaj, S. A. O. (2025). Comprehensive Review and Analysis on Facial Emotion Recognition: Performance Insights into Deep and Traditional Learning with Current Updates and Challenges. *Computers, Materials & Continua*, *82*(1), 41–72. https://doi.org/10.32604/cmc.2024.058036

Reza, M. H., Sibly, M. N. B., Rabbani, S. G., Sadi, S. H., Ahamed, M. F., Shafi, F. B., ... & Chowdhury, M. E. (2026). A comprehensive review of convolutional neural networks: foundations, enhancements and applications. *Neural Computing and Applications*, *38*(4), 56. https://doi.org/10.1007/s00521-025-11827-w

Riviere, E., Courbois, Y., & Gentaz, E. (2026). The developmental changes in emotion recognition from human biological motion by children aged from 4 to 12 years. *Emotion*. https://doi.org/10.1037/emo0001626

Sarvakar, K., & Rana, K. (2025). Revolutionizing facial emotion recognition: In-depth analysis of cutting-edge models, methodologies, and datasets. *Discover Artificial Intelligence*, *5*(1), 388. https://doi.org/10.1007/s44163-025-00553-w

Saurav, S., Saini, R., & Singh, S. (2024). An integrated attention-guided deep convolutional neural network for facial expression recognition in the wild. *Multimedia Tools and Applications*, *84*(12), 10027–10069. https://doi.org/10.1007/s11042-024-19012-2

Shangguan, Z., Dong, Y., Guo, S., Leung, V. C. M., Deen, M. J., & Hu, X. (2026). Facial Expression Analysis and Its Potentials in IoT Systems: A Contemporary Survey. *ACM Computing Surveys*, *58*(2), 1–39. https://doi.org/10.1145/3737456

Shingjergji, K., Iren, D., Urlings, C., & Klemke, R. (2025). Design Principles for Affective Online Learning. *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-025-09867-1

Tan, M., & Le, Q. V. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. https://doi.org/10.48550/ARXIV.1905.11946

Villegas-Ch, W., Gutierrez, R., & Mera-Navarrete, A. (2025). Multimodal Emotional Detection System for Virtual Educational Environments: Integration Into Microsoft Teams to Improve Student Engagement. *IEEE Access*, *13*, 42910–42933. https://doi.org/10.1109/ACCESS.2025.3546772

Wang, H., Wang, C., Xu, H., Bao, C., & Xu, X. (2026). Enhancing real-time facial emotion recognition in classrooms via Attention-ResNet optimization. *The Visual Computer*, *42*(1), 46. https://doi.org/10.1007/s00371-025-04265-1

Zong, L., Nan, S. J., Die, Z. F., & Peng, H. J. (2026). DMSCA: Dynamic multi-scale channel-spatial attention for enhanced feature representation in convolutional neural networks. *Scientific Reports*, *16*(1), 8044. https://doi.org/10.1038/s41598-026-37546-3

Publisher's Note: CRIBFB stays neutral about jurisdictional claims in published maps and institutional affiliations.