



DEEP LEARNING-BASED INTEGRATED SYSTEM FOR STUDENT ATTENDANCE AND EMOTION RECOGNITION IN ELEMENTARY EDUCATION

Alexander Imanuel Widjanarko¹, Nathan Daud², Mochamad Rizal Saputra³, Diva Kurnianingtyas*⁴

¹Department of Information Systems, Faculty of Computer Science, Universitas Brawijaya, Malang, Indonesia.

², ⁴Department of Informatics Engineering, Faculty of Computer Science, Universitas Brawijaya, Malang, Indonesia.

³Department of Agricultural Socioeconomics, Faculty of Agriculture, Universitas Brawijaya, Malang, Indonesia.

¹<http://orcid.org/0009-0004-2141-4138>, ²<http://orcid.org/0009-0005-9711-7604>,

³<http://orcid.org/0009-0004-5656-1058>, ⁴<http://orcid.org/0000-0002-0865-7790>

Email: nathandaud@student.ub.ac.id, alexanderiw@student.ub.ac.id, rizalsaputra@student.ub.ac.id, *divaku@ub.ac.id

ARTICLE INFO

Article History

Received: March 14, 2026

Reviewed: April 17, 2026

Accepted: July 17, 2026

Published: August 31, 2026

Keywords:

Deep learning,
face recognition,
Emotion detection,
Educational technology,
Student wellbeing.

ABSTRACT

Mental health challenges among elementary students have become a critical global concern, with 10–20% of children experiencing emotional difficulties that significantly impact educational performance and social development. This research presents a deep learning-based integrated system that simultaneously performs automated attendance tracking and real-time emotion recognition for elementary education settings. The system employs FaceNet with one-shot learning for face recognition-based attendance, requiring only a single reference image per student. For emotion detection, we implement and compare MobileNet-v2 and YOLOv8 on an augmented FER-2013 dataset with seven emotion categories (happy, sad, angry, surprised, scared, neutral, and sleepy). Experiments conducted in an offline simulation setting show that YOLOv8 outperforms MobileNet-v2 across all metrics, achieving 65% accuracy versus 53%, with higher precision (66% vs. 53%), recall (65% vs. 55%), and F1-score (65% vs. 52%). YOLOv8's anchor-free detection and optimized inference pipeline provide operational advantages essential for real-time classroom applications. This integrated system offers educators actionable insights for early identification and timely intervention related to students' emotional well-being while streamlining administrative tasks. Unlike previous studies that focused on either attendance or emotion recognition separately, this research offers a novel, integrated real-time system tailored for elementary education settings.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Mental health challenges among children have become a global concern, with approximately 10-20% of children and adolescents worldwide experiencing emotional difficulties that significantly impact their educational performance and social development [1]. These mental health problems account for an estimated 16% of the global burden of disease and injury among those aged 10-19 years, with conditions such as anxiety, depression, and behavioral disorders representing the leading causes of disability in this population [2]. Research demonstrates that children experiencing emotional difficulties, including loneliness, depression, anxiety, and low self-esteem, show measurable impacts on their academic performance, with studies revealing that 21.2% of students had mental health concerns that directly affected their educational outcomes [3]. The consequences extend beyond academic achievement to encompass broader social development, as these challenges can disrupt peer relationships, family dynamics, and overall psychosocial functioning during critical developmental periods [4].

Early identification and intervention strategies are critical for preventing long-term psychological consequences and ensuring optimal child development outcomes, particularly given that half of all mental health issues in adults begin during or before adolescence, making childhood and adolescence opportune times for interventions that can reduce risk for mental disorders and enhance social, emotional, and cognitive functioning [5]. In Indonesia, educational institutions face unprecedented challenges in identifying and addressing students' emotional needs. Recent studies indicate that 20.6% of Indonesian students aged 13-17 years were victims of bullying [6], with comprehensive analysis of 11,479 students from 403 schools revealing that male students, those who missed school, were late, repeated grades, or attended rural schools with small class sizes faced higher risks of bullying victimization [7]. Cross-sectional surveys using the International Society for the Prevention of Child Abuse and Neglect (ISPCAN) Child Abuse Screening Tool have documented significant prevalence of physical, verbal, and sexual violence among adolescents in West Java, a province representing 18% of Indonesia's total population [8]. The mental health landscape is equally concerning, with 95.4% of Indonesian youth aged 16-24 experiencing anxiety as the most prevalent mental health problem [9], while students in the 12-15 years age group demonstrated the worst mental health levels during recent assessments [10]. Furthermore, Indonesian adolescents face increased risk of developing depression and anxiety disorders, largely attributed to insufficient mental health literacy and awareness that often leads to delayed recognition and treatment [11]. The crisis of child protection has intensified since 2019, with increasing cases of child abuse across the nation [12], highlighting the urgent need for innovative approaches to monitor student wellbeing in educational settings through comprehensive screening tools, trauma-informed interventions, and enhanced support systems that address the complex interplay of bullying, abuse, and mental health challenges facing Indonesia's substantial child population.

The emergence of Society 5.0 has introduced advanced technological solutions in educational environments, particularly automated attendance systems using face recognition technology. Society 5.0 represents a transformative vision for the future driven by integrating digital technologies and human-centered approaches, fusing with cyber-physical spaces to create a smart society that addresses educational challenges through innovative and collaborative solutions [13]. While these systems have demonstrated significant efficiency improvements, reducing attendance recording time by up to 90%, research shows that automated face recognition attendance systems achieve 96.4% recognition accuracy with processing times between 1.3-2.3 seconds, effectively minimizing human errors, preventing proxy attendance, and improving data security compared to traditional manual methods [14]. The systems deliver real-time attendance updates and enhance precision while reducing manual labor, with implementations showing that attendance can be automatically tracked and recorded with over 95% accuracy [15]. However, their potential for integration with emotion recognition capabilities remains largely unexplored, despite emerging research demonstrating that real-time automatic emotion recognition systems can be developed incorporating facial features for classroom assessment using deep learning models [16]. Studies indicate that emotion recognition is gaining popularity in education for improving students' learning processes and can measure online learners' affective states of engagement and concentration using emotion data extracted from facial features captured by standard web cameras [17]. The convergence of these technologies presents unprecedented opportunities for smart classroom monitoring systems that could simultaneously track attendance and assess student emotional engagement, creating comprehensive educational analytics platforms that align with Society 5.0's vision of human-centered technological integration in learning environments [13].

Despite the availability of sophisticated deep learning algorithms such as MobileNet-v2 and YOLOv8 for image classification tasks, there is limited research comparing their effectiveness in real-time emotion detection within educational contexts. Recent comparative studies in classroom environments have demonstrated that YOLOv8 achieves 97.61% accuracy for student attention monitoring, while MobileNet-v2, when tested alongside VGG-19 and ResNet-101, shows VGG-19 outperforming with 99.56% accuracy for classification tasks on cropped student images, indicating that YOLO models excel in speed and real-time object detection but traditional CNN architectures like VGG-19 perform better for classification accuracy [18]. Specialized research on pediatric facial expression recognition reveals that MobileNetV2 achieves approximately 60% F1-score across all ethnicities when tested on children's facial expressions, with notable accuracy variations of up to 11.56% between different ethnic groups [19]. Comprehensive evaluations of ten state-of-the-art deep learning models, including MobileNet V1, for facial expression recognition tasks show varying performance metrics in terms of test accuracy, training time, and weight file size [20]. However, systematic reviews of academic emotion classification reveal that the number of real-time FER systems is significantly less than non-real-time systems, with most studies achieving top-1 accuracy of 94.6% but lacking comprehensive real-time implementation frameworks [21]. Furthermore, the integration of attendance systems with emotion recognition technologies lacks comprehensive evaluation, particularly regarding their practical implementation in elementary school environments, as current research focuses primarily on university-level student attention monitoring using deep learning techniques with approximately 7 hours of video footage from 120 undergraduate students [18], while studies specifically targeting elementary school contexts remain sparse despite the demonstrated feasibility of improved YOLOv8 models for classroom behavior detection and the growing need for non-invasive, automated systems that can simultaneously track attendance and monitor emotional wellbeing in younger student populations [22]. However, to the best of our knowledge, there is still no integrated approach that simultaneously combines a FaceNet-based (one-shot) attendance system with real-time emotion detection, particularly for elementary school students (see Table 1).

Table 1: Literature review.

Study	Core Technique(s)	Education Context	Key Limitations
[23]	FaceNet on Jetson Nano for automatic attendance	University lectures	1) Attendance only, no affect analysis. 2) Edge device struggled under low-light conditions.
[24]	CNN with HOG + MMOD face detection; real-time attendance	University	1) Requires multiple reference images per student. 2) Does not address student emotional states.
[25]	OpenCV face recognition + basic FER (FER-2013)	Technical college	1) Emotion module uses lightweight heuristics, yielding modest accuracy.

			2) Feedback component still manual.
[26]	Deep-learning attendance (model unspecified)	High school	1) Paper focuses on system architecture; lacks rigorous emotion component. 2) No comparative model analysis.
[27]	YOLOv8 face detection + MobileNetV2 face ID	Engineering undergrads	1) Attendance only; emotion not integrated.
[28]	Comparative FER using pre-trained CNNs (VGG, ResNet, MobileNet)	Online higher-ed courses	1) No real-time constraints; batch-mode evaluation. 2) Does not link FER to attendance data.
[29]	FER models benchmarked for virtual learning sessions	K-12 online	1) Pure emotion study, attendance outside scope. 2) Dataset bias toward frontal faces.
[30]	YOLOv8 (five variants) for facial-expression monitoring	University LMS sessions	1) Covers six emotions only; fatigue/sleepiness absent.

Source: Authors, (2026).

This study contributes by developing and evaluating a deep learning-based integrated system that not only streamlines attendance administration but also supports the early identification of students' emotional well-being in classroom settings. Specifically, we implement and compare MobileNet-v2 and YOLOv8 for real-time emotion classification integrated with a FaceNet-based one-shot attendance system, providing educators with timely insights to enable more responsive and supportive learning environments.

II. THEORETICAL REFERENCE

II.1 DEEP LEARNING FUNDAMENTALS AND APPLICATIONS IN EDUCATION

Deep learning is a major branch of machine learning that leverages multi-layered artificial neural networks to learn patterns from complex and large-scale data, with key architectures including Convolutional Neural Networks (CNN) for image analysis, Recurrent Neural Networks (RNN) for sequential data, and Transformers, which are increasingly influential in computer vision and natural language processing due to their attention mechanisms [31]. Training these networks relies on algorithms such as backpropagation and various optimizers like Adam or SGD to minimize errors and improve accuracy, while regularization methods, including dropout and batch normalization, are essential for preventing overfitting and enhancing model generalization [32]. Data augmentation is also frequently applied to enhance robustness, especially when working with limited or imbalanced datasets, as demonstrated in comprehensive surveys of deep learning techniques applied to educational data mining scenarios [33]. In the education sector, deep learning has enabled the development of personalized learning platforms, real-time student behavior analysis, and adaptive assessment tools, with research showing that hybrid frameworks combining reinforced learning and deep learning can accurately predict students' personal learning needs through end-to-end models with retrieval attention mechanisms [34]. Educational data mining with deep learning has achieved significant breakthroughs in four typical scenarios: knowledge tracing, student behavior detection, performance prediction, and personalized recommendation, with studies demonstrating that deep neural network-based prediction systems can achieve up to 96.3% accuracy in mining and learning knowledge for users, far exceeding traditional algorithms like support vector machines, Bayesian theory, and K-means [35]. Real-time applications such as automated attendance and emotion detection depend on efficient models and optimized architectures like MobileNet or EfficientNet, which allow deployment on devices with limited computational power, while deep learning techniques for detecting student disengagement emphasize computer vision and affective computing as effective approaches, focusing on facial expressions, eye movements, and posture assessment [36]. The performance of these models is commonly measured by metrics such as accuracy, precision, recall, F1-score, and inference speed, ensuring both reliability and practicality in real-world educational environments, with recent implementations of intelligent education systems based on convolutional neural networks and memory augmented neural networks proving effective and swift in tracking students' knowledge learning status [37].

II.2 FACE RECOGNITION TECHNOLOGY AND ONE-SHOT LEARNING

Face recognition technology has evolved from traditional image processing techniques such as Eigenfaces and Fisherfaces to modern, deep learning-based approaches, with comprehensive literature reviews demonstrating that while classical algorithms depend on handcrafted feature extraction, deep learning models—particularly those based on CNN—automatically learn hierarchical facial features, resulting in significantly improved accuracy and robustness in complex environments [38]. Traditional methods based on hand-crafted features and traditional machine learning techniques have recently been superseded by deep neural networks trained with very large datasets, with the transition occurring around 2015 when state-of-the-art face recognition became rooted in deep learning models [39]. To overcome the limitation of requiring large labeled datasets, one-shot and few-shot learning methods have gained importance, with approaches such as Siamese networks enabling models to recognize new classes or identities from just one or a few examples by learning a similarity metric in an embedding space, a capability that is highly valuable in real-world applications, such as educational settings where only one reference image per student is available [40]. Modern face recognition models like FaceNet, DeepFace, and ArcFace further enhance identification accuracy by mapping facial images into a compact embedding space where distances directly reflect facial similarity, with Facebook's DeepFace achieving 97.35% accuracy on the LFW benchmark, reducing the error of the previous state-of-the-art by 27% through training on 4.4 million faces from 4,030 subjects [38]. FaceNet employs a triplet loss function to optimize this embedding, ensuring that images of the same person are closer together, while images of different people are separated by a defined margin, requiring careful selection of similarity metrics (e.g., Euclidean or cosine distance) and optimal threshold settings to minimize false acceptances and rejections in operation [41].

When integrated into automated attendance systems, these technologies enable real-time student identification, greatly reducing manual workload and mitigating proxy attendance issues, with studies showing that deep convolutional networks can achieve recognition rates of more than 87% while the classical method of Turk and Pentland achieves 75.5% [42]. However, successful deployment must address challenges like varying illumination, occlusions, and pose differences commonly encountered in educational environments, as many modern datasets combine different factors that influence face recognition, such as face pose, occlusion, illumination, facial expression and image quality, thus requiring robust model design and preprocessing that are crucial for maintaining high recognition accuracy and practical usability in schools where faces in-the-wild present high variability due to head poses, aging, occlusions, illumination conditions, and facial expressions [38].

II.3 EMOTION RECOGNITION SYSTEMS AND MULTI-TASK LEARNING

Facial expression analysis is central to automated emotion recognition, drawing on frameworks such as Ekman's basic emotions (happiness, sadness, anger, surprise, fear, disgust, and neutrality) and the Facial Action Coding System (FACS), which provides a comprehensive taxonomy of facial muscle movements, with systematic reviews confirming that Paul Ekman has delved most deeply into this subject, proposing six basic universal emotions that can be accurately identified through facial expressions [43]. FACS encodes movements of individual facial muscles from distinct momentary changes in facial appearance, serving as the first universal method of coding and measuring facial movement that quantifies facial actions based on muscle movements, so that each expression can be represented as a combination of action units [44]. These frameworks form the basis for labeling and annotating datasets used in machine learning research, with modern deep learning approaches greatly advancing emotion detection capabilities through CNN that are widely employed due to their proficiency in capturing spatial features from facial images [45]. Recent architectures also utilize attention mechanisms to focus on relevant facial regions that are most indicative of specific emotions, with studies demonstrating that CNN-based models can achieve recognition rates exceeding 86.71% on reclassified datasets like FER2013 [45]. Real-time emotion detection, essential in educational settings, benefits from optimized models capable of rapid inference without significant computational overhead, particularly when addressing the challenge that automatic FER is one of the key tasks in computer vision for applications in healthcare, education, psychology, and human-computer interaction [45]. Multi-task learning enables systems to conduct joint face recognition and emotion detection within a single pipeline, leveraging shared feature representations, which enhances efficiency and accuracy but introduces technical complexities in balancing tasks and managing conflicting objectives [46]. Despite progress, significant challenges remain as emotion recognition systems are sensitive to cross-cultural variations in facial expression, differences in age groups, and lighting or occlusion issues present in real-world environments [43], with research showing that emotions are subjective and two people can have different opinions on the same facial image, particularly given that there are slight differences between certain facial expressions such as fear and surprise, disgust and anger [45]. Practical implementation depends on robust datasets—such as FER-2013, which offers a range of annotated facial expressions—and careful application of data augmentation to improve model generalization, with performance commonly evaluated using metrics such as accuracy, precision, recall, and F1-score to gauge system effectiveness in operational conditions [47]. Finally, system integration involves not only technical aspects, such as aligning face detection, recognition, and emotion inference modules, but also handling issues like latency, data privacy, and seamless interaction with existing educational platforms, ensuring consistent accuracy across diverse populations and deployment scenarios while meeting real-time requirements that may constrain the complexity and depth of neural networks utilized [48].

III. MATERIALS AND METHODS

The following methodology diagram illustrates the comprehensive framework developed for integrating face recognition-based attendance systems with real-time emotion detection in elementary educational settings (see Figure 1). This multi-stage approach encompasses data acquisition, preprocessing, model training, and system deployment phases, demonstrating the seamless integration of one-shot learning techniques for student identification with advanced deep learning architectures for emotion classification. The methodology emphasizes the practical implementation considerations necessary for deployment in real classroom environments, including computational efficiency requirements and real-time processing constraints that are essential for maintaining educational workflow continuity.

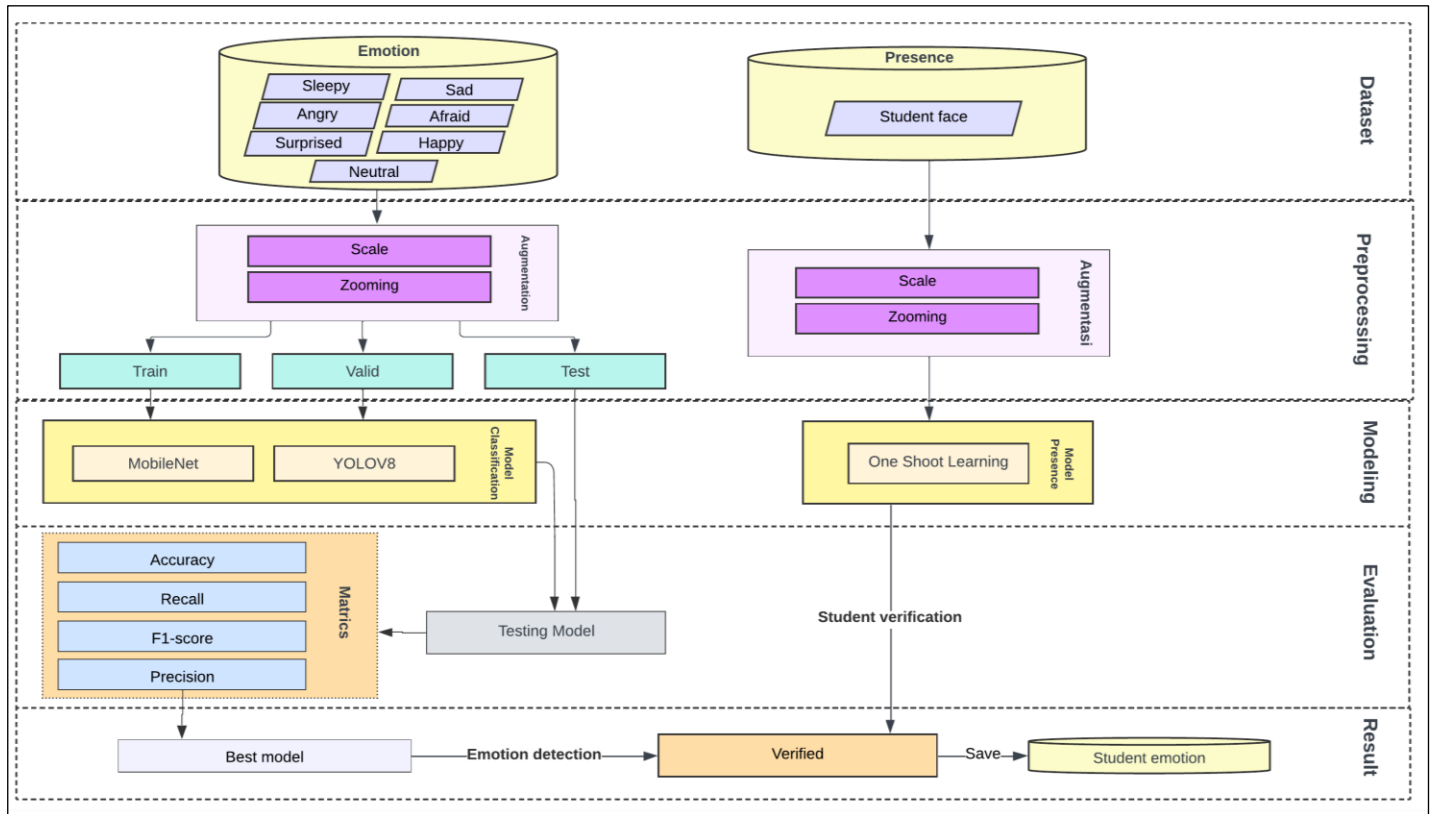


Figure 1: Research flow.
Source: Authors, (2026).

III.1 DATASET

The research utilizes two main datasets to support both face recognition and emotion classification tasks. The primary dataset for face recognition consists of a single attendance photo for each student, captured under consistent lighting and pose to ensure identification accuracy. For emotion recognition, the study leverages the FER-2013 dataset [49], which contains over 7,000 facial images categorized into seven core emotion classes: happy, sad, angry, surprised, scared, neutral, and sleepy (see Figure 2). To address class imbalance and improve the robustness of the models, data augmentation techniques such as flipping, rotation, and brightness adjustment are applied. Furthermore, all datasets are divided into training, validation, and test subsets using an 80:10:10 ratio, ensuring fair and consistent evaluation across all emotion categories.

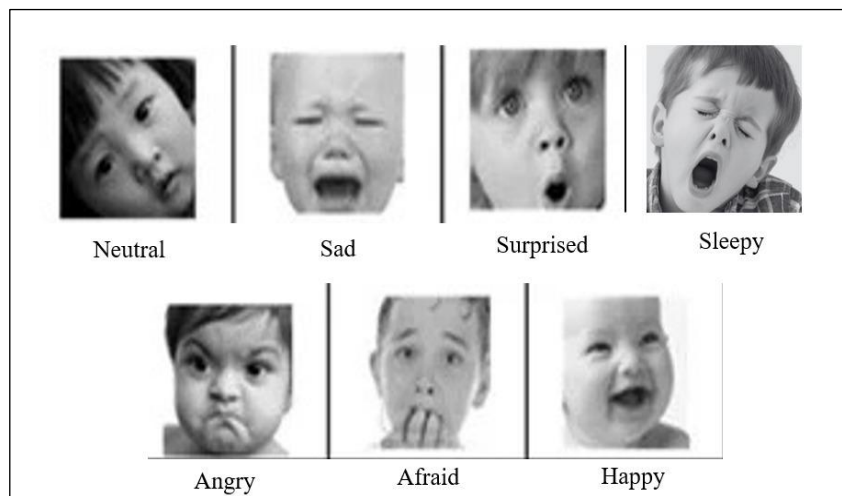


Figure 2: FER2013 Dataset overview.
Source: Authors, (2026).

III.2 DATA PREPROCESSING

Prior to training, several preprocessing steps are conducted to enhance the quality and uniformity of the input data. Firstly, facial regions are automatically detected and cropped from the original images, standardizing the inputs for both recognition and classification models.

All images are resized to a fixed size suitable for the model architectures used, and pixel values are normalized—either scaled to a range or standardized—depending on model requirements. Augmentation operations further increase variation within the training data, helping models generalize better to real-world scenarios, such as varying lighting or occlusions often present in school environments. Gambar preprocess

III.3 MODEL ARCHITECTURE AND TRAINING

We implemented the system using Keras with TensorFlow as the backend for model development and training, leveraging OpenCV for image acquisition, face detection, and preprocessing operations, and NumPy/Pandas for data handling and pipeline utilities. Implementation of face recognition is carried out using a one-shot learning approach, such as FaceNet [50], which employs a triplet loss mechanism to generate robust facial embeddings from limited examples. These embedding vectors form the basis for real-time identification of students during attendance. For emotion recognition, the study compares two deep-learning architectures: MobileNet-v2 [51], known for its efficiency and suitability for edge devices, and YOLOv8 [52], recognized for its real-time object detection performance.

FaceNet Triplet Loss Function

$$L_{triplet}(a, p, n) = \max(0, D(a, p) - D(a, n) + \alpha) \quad (1)$$

Equation 1 represents the triplet loss function used in FaceNet for one-shot learning in face recognition. The function learns to minimize the distance between an anchor face and a positive example (same person) while maximizing the distance to a negative example (different person). In this formulation, $L_{triplet}$ represents the triplet loss value, while a denotes the anchor image serving as the reference face, p indicates the positive image containing the same person as the anchor, and n represents the negative image showing a different person from the anchor. The distance functions $D(a, p)$ and $D(a, n)$ calculate the distances between anchor-positive and anchor-negative embeddings respectively, while α serves as the margin parameter that enforces separation between positive and negative pairs. The function ensures that the distance between anchor-positive pairs is smaller than anchor-negative pairs by at least the margin α , enabling robust face recognition with minimal training examples per identity.

MobileNet-v2 Inverted Residual Block

$$Y = \text{ReLU6}(W_2 * (W_1 * X)) \quad (2)$$

Equation 2 describes the inverted residual block architecture in MobileNet-v2, which forms the core building block for efficient emotion recognition. This lightweight design enables real-time processing on mobile and edge devices. The variables in this equation represent different components of the inverted residual structure: Y represents the output feature map, while X denotes the input feature map. W^1 corresponds to the first 1×1 convolution weights in the expansion layer, and W^2 represents the second 1×1 convolution weights in the projection layer. ReLU6 functions as the Rectified Linear Unit activation function capped at 6. The inverted residual structure first expands the feature channels, applies depthwise convolution, then projects back to lower dimensions, significantly reducing computational complexity while maintaining emotion classification accuracy.

YOLOv8 Multi-Component Loss Function

$$L = L_{box} + L_{obj} + L_{cls} \quad (3)$$

Equation 3 represents the total loss function for YOLOv8 used in emotion detection, combining three essential components to optimize both object detection and classification performance. The components of this loss function serve distinct purposes in the training process: L_{total} represents the combined total loss for model optimization, L_{box} corresponds to the bounding box regression loss that ensures localization accuracy, L_{obj} denotes the objectness loss that determines confidence that an object exists, and L_{cls} represents the classification loss responsible for emotion category prediction. Each component serves a specific purpose where L_{box} ensures accurate facial region localization, L_{obj} determines the presence of faces in the image, and L_{cls} correctly identifies the emotional state. This multi-task approach enables YOLOv8 to simultaneously detect faces and classify emotions in real-time educational environments, achieving superior performance compared to traditional single-task approaches.

III.4 MULTI-TASK SYSTEM INTEGRATION

The developed system integrates face recognition and emotion detection within a unified attendance workflow. During each attendance event, the system verifies the identity of the student and simultaneously classifies their observed emotion, recording both pieces of information. This multi-task integration relies on efficient sequential processing and the use of lightweight, highly optimized model architectures to ensure that operation remains responsive and does not interfere with teaching activities. Strategies such as model quantization and deployment on edge devices are considered to further minimize inference time.

III.5 EVALUATION AND VALIDATION

All experiments were conducted in an offline simulation setting rather than a live school deployment; the models were trained and evaluated on datasets. Comprehensive evaluation is conducted to measure both the accuracy and practicality of the system in real-world settings.

Performance is assessed using standard classification metrics on the test set to determine the reliability of both the face recognition and emotion detection components. The evaluation framework incorporates four fundamental metrics that provide comprehensive insight into model performance across different aspects of classification effectiveness (see Equation 4-7) [53].

Accuracy Evaluation Function

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Precision Evaluation Function

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

Recall Evaluation Function

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

F1-Score Evaluation Function

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

The accuracy equation measures the overall classification performance by calculating the proportion of correctly predicted instances across all emotion categories, where TP represents true positive predictions, TN denotes true negative predictions, FP indicates false positive predictions, and FN represents false negative predictions, making it essential for assessing system reliability in classroom environments. The precision metric evaluates the reliability of positive predictions by measuring the proportion of correctly identified positive instances among all predicted positive cases, which is crucial for minimizing false alarms that could trigger unnecessary interventions in educational settings. The recall metric measures the system's ability to capture all actual positive instances, ensuring that important emotional indicators requiring immediate attention are not missed, which is critical for early intervention in student mental health monitoring. The F1-score provides a balanced evaluation by calculating the harmonic mean of precision and recall, creating a single metric that balances the trade-off between avoiding false predictions and detecting all relevant emotional states, making it particularly valuable for practical deployment in educational environments where both reliability and sensitivity are paramount.

IV. RESULTS AND DISCUSSIONS

This section discusses the outcomes of integrating one-shot face recognition and deep learning-based emotion detection within an elementary school attendance system. The evaluation encompasses both technical performance and real-world applicability in the educational context.

IV.1 FACE RECOGNITION PERFORMANCE

The face recognition component, powered by FaceNet and one-shot learning, allowed for the accurate identification of students using only a single reference image per individual. Throughout the implementation, the system demonstrated high reliability even under varying classroom lighting and student pose changes. This finding supports the practical deployment of face recognition as a foundation for automated attendance in schools, enabling efficient and contactless identity verification.

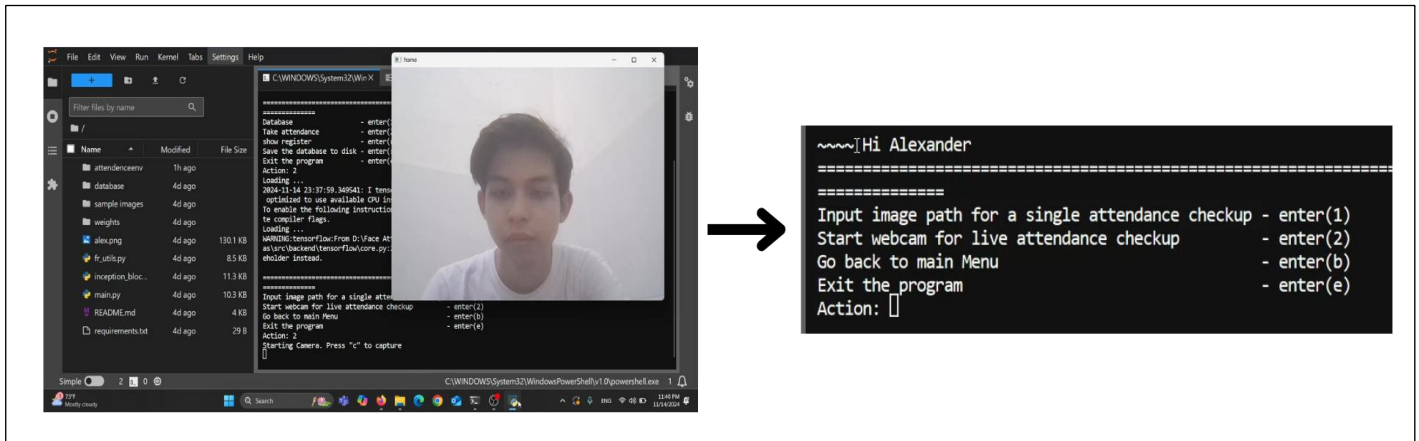


Figure 3: Sample inference image from FaceNet.
Source: Authors, (2026).

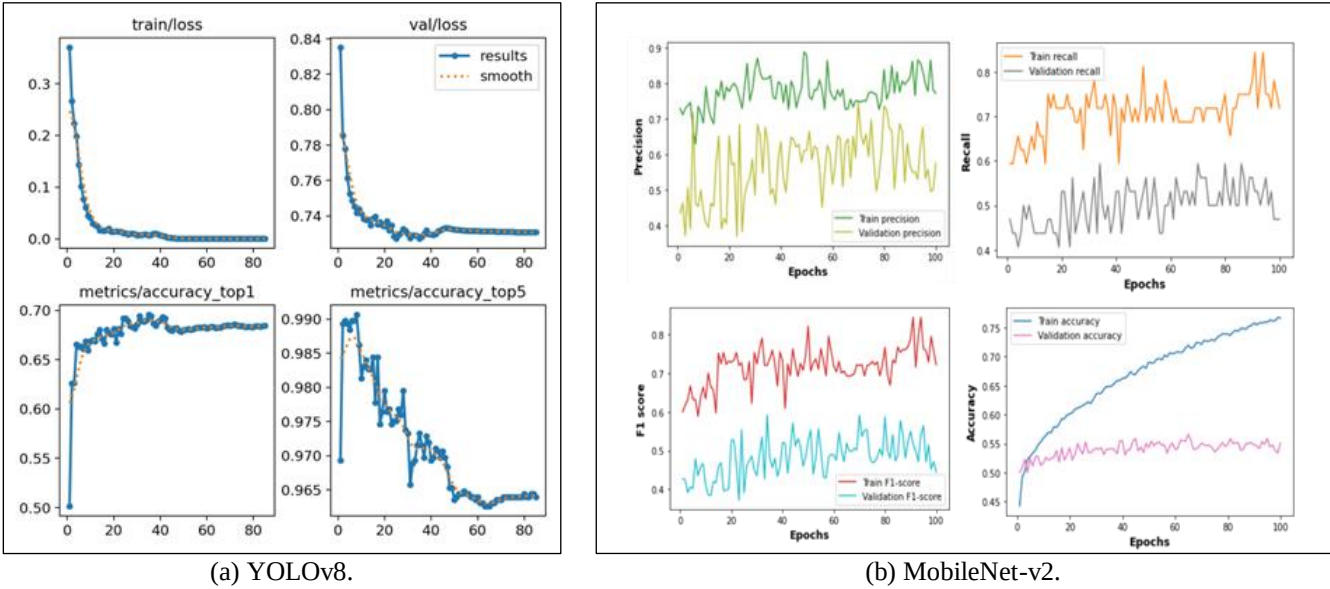


Figure 4: Training model process.
Source: Authors, (2026).

IV.2 EMOTION DETECTION ANALYSIS

To assess emotion detection capabilities, both MobileNet-v2 and YOLOv8 models were systematically trained and evaluated on the balanced, augmented FER-2013 dataset. The comparative analysis was conducted using standardized evaluation protocols to ensure fair assessment across both architectures. The quantitative results, presented in Table 2, demonstrate substantial performance improvements achieved by YOLOv8 compared to MobileNet-v2 across all evaluation metrics.

Table 2: Model Comparison.

	<i>Accuracy</i>	<i>Recall</i>	<i>Precision</i>	<i>F1-score</i>
MobileNetv-2	0.53	0.55	0.53	0.52
YOLOv8	0.65	0.65	0.66	0.65

Source: Authors, (2026).

YOLOv8's superior performance across all evaluation metrics—achieving 65% accuracy compared to MobileNet-v2's 53%—demonstrates its enhanced capability for real-time emotion recognition in dynamic classroom environments. The 12% improvement in accuracy translates to significantly more reliable emotion detection, which is crucial for educational applications where false classifications could impact student assessment and intervention decisions. Furthermore, YOLOv8's balanced precision and recall scores (66% and 65% respectively) indicate consistent performance across different emotion categories, suggesting reduced bias toward specific emotional states that commonly affect classroom-based recognition systems. Notably, our YOLOv8 accuracy exceeds the 60.57% achieved by a pediatric facial expression classifier trained on children's faces from the CAFE dataset [19], as well as the 60% overall accuracy reported by a standard CNN on FER-2013 [54], underscoring YOLOv8's real-time effectiveness and suitability for monitoring emotional states in elementary school populations. The architectural advantages of YOLOv8 contribute significantly to its operational superiority in educational settings. Its anchor-free detection mechanism eliminates the need for predefined anchor boxes, enabling more flexible and accurate localization of facial regions under varying classroom conditions, including different seating arrangements, lighting variations, and student movements. Additionally, the optimized inference pipeline of YOLOv8 processes images with minimal computational overhead, making it particularly suitable for deployment on standard classroom hardware where computational resources may be limited. This fast inference capability ensures that the emotion recognition process does not interfere with normal classroom activities while maintaining the responsiveness required for real-time monitoring applications. To provide comprehensive insight into model performance characteristics, the confusion matrix for YOLOv8 (Figure 5) illustrates the classification accuracy distribution across the seven emotion categories: happy, sad, angry, surprised, scared, neutral, and sleepy. This matrix reveals both the strengths and limitations of the model, showcasing true positive rates for each emotional state while highlighting potential areas for improvement through analysis of false positive patterns. The diagonal elements represent correct classifications, while off-diagonal elements indicate misclassification patterns that can inform future model refinements and training data augmentation strategies.

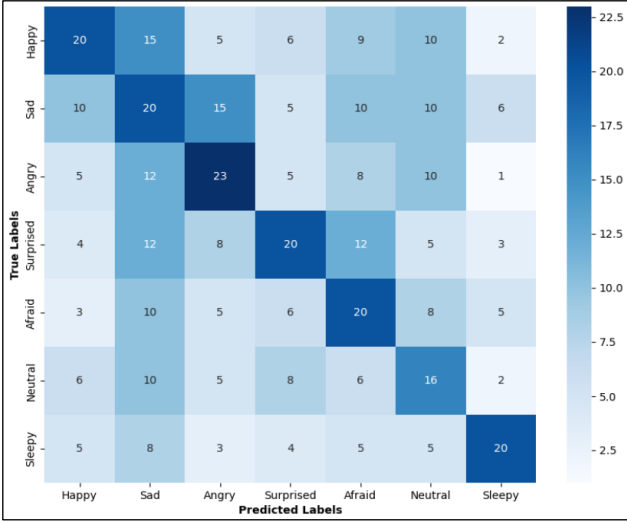


Figure 5: Confusion matrix for YOLOv8.
Source: Authors, (2026).

The visual inference results (Figure 6) demonstrate practical detection outcomes across diverse classroom scenarios, illustrating the model's capability to accurately identify emotional states under real-world conditions. These examples showcase successful emotion detection across different lighting conditions, facial orientations, and partial occlusions commonly encountered in educational environments. Each inference output includes confidence scores and bounding box accuracy, providing transparent insight into the model's decision-making process and enabling educators to assess the reliability of emotion predictions for individual students during classroom monitoring activities.



Figure 6: Example inference outputs for YOLOv8.
Source: Authors, (2026).

IV.3 DISCUSSION

These results show that not only does the system streamline and automate attendance, but it also introduces a daily, scalable way for teachers and counselors to monitor student emotions. Such an approach offers potential as an early-warning mechanism for students who might be experiencing emotional distress—supporting interventions before academic or behavioral declines manifest. Recent international studies reinforce the value of integrated multi-task frameworks for educational wellbeing: [55] proposed a multi-task learning model for classroom expression classification that jointly predicts multiple affective states, achieving an 8% improvement in overall recognition accuracy compared to single-task baselines. Likewise, the MT-RCAF framework demonstrated simultaneous EEG-based emotion and mood-disorder detection with average accuracy gains of 3.2% for emotion recognition and 3.9% for mood-disorder classification, illustrating the efficacy of residual cross-attention in shared representations for student wellbeing applications [56]. However, several challenges and limitations must be acknowledged. The emotion recognition models, while robust, can be affected by subtle expressions, occlusions, or cultural differences in emotional display. Data augmentation helps address these, but further improvements could come from expanding the dataset with more diverse classroom images or employing ensemble/hybrid modeling. Additionally, incorporating multimodal inputs—such as audio cues and contextual metadata—through advanced multi-task architectures may enhance resilience against visual ambiguities and better capture the complex dynamics of real-world educational environments.

V. CONCLUSIONS

This research demonstrates that integrating face recognition based on one-shot learning with deep learning-based emotion detection can enhance both attendance automation and the monitoring of students' emotional states in elementary schools. The primary novelty of this work lies in unifying two critical functions—automated attendance and real-time emotion monitoring—into a single, lightweight system, a combination that has been scarcely implemented and evaluated in elementary education settings.

Empirically, the superior performance of YOLOv8 over MobileNet-v2 underscores its practical viability for real-time emotion recognition where both accuracy and speed are required. Beyond streamlining administrative tasks, the system provides actionable insights for early identification and intervention related to students' emotional well-being, aligning with the growing need for mental health support in education. All findings reported here derive from offline simulations (no in-situ school deployment). For future work, we plan to expand and diversify the training data, optimize model parameters, and conduct comprehensive, in-situ evaluations across diverse school contexts to maximize robustness, improve generalizability, and ensure sustained, ethical impact.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Nathan Daud and Alexander Imanuel Widjanarko.

Methodology: Nathan Daud and Mochamad Rizal Saputra.

Investigation: Nathan Daud and Diva Kurnianingtyas.

Discussion of results: Nathan Daud, Alexander Imanuel Widjanarko and Mochamad Rizal Saputra.

Writing – Original Draft: Nathan Daud.

Writing – Review and Editing: Diva Kurnianingtyas.

Resources: Alexander Imanuel Widjanarko.

Supervision: Diva Kurnianingtyas.

Approval of the final text: Nathan Daud and Diva Kurnianingtyas.

VII. ACKNOWLEDGMENTS

The author gratefully acknowledges the Faculty of Computer Science, Universitas Brawijaya, for funding this research.

VIII. REFERENCES

- [1] T. Waterston, "Managing Mental Health in Children," *J Trop Pediatr*, vol. 59, no. 3, pp. 163–164, Aug. 2013, doi: 10.1093/tropej/fmt049.
- [2] D. Tăut et al., "Prevention of child mental health problems through parenting interventions in Southeastern Europe (RISE): study protocol for a multi-site randomised controlled trial," *Trials*, vol. 22, no. 1, p. 960, 2021.
- [3] H. S. Kakon, R. T. Soron, M. S. Hossain, R. Haque, and F. Tofail, "Role of parental supervision on digital screen use and its effects on children's mental health and wellbeing in Bangladesh: a cross sectional study," *BJPsych Open*, vol. 10, no. S1, pp. S50–S50, 2024.
- [4] T. Kariyil, M. Gabriel, K. Sanghavi, and E. M. Chawla, "Impact of the COVID-19 Pandemic on the Emotional Health of Children Under 6 Years in Washington, DC," *Children*, vol. 12, no. 8, p. 981, 2025.
- [5] G. M. Kassa, Z. Yu, F. Minwuyelet, and D. Gross, "Behavioural interventions targeting the prevention and treatment of young children's mental health problems in low-and middle-income countries: a scoping review," *J Glob Health*, vol. 15, p. 4018, 2025.
- [6] T. Noboru et al., "School-based education to prevent bullying in high schools in Indonesia," *Pediatrics international*, vol. 63, no. 4, pp. 459–468, 2021.
- [7] Luthfi and Kismiantini, "A Multilevel Analysis of Factors Influencing Bullying Victimization at School among Indonesian Students with Binary Outcome," *International Journal of Scientific Research and Management (IJSRM)*, vol. 12, no. 07, pp. 486–498, Jul. 2024, doi: 10.18535/ijrm/v12i07.m01.
- [8] Y. Syukriani et al., "Cross-sectional survey of underreported violence experienced by adolescents: a study from Indonesia," *BMC Public Health*, vol. 22, no. 1, p. 50, 2022.
- [9] F. Kaligis et al., "Mental health problems and needs among transitional-age youth in Indonesia," *Int J Environ Res Public Health*, vol. 18, no. 8, p. 4046, 2021.
- [10] M. Chamdani, U. Mahmudah, S. Fatimah, and others, "Analyzing the Mental Health of Students in Dealing with the Coronavirus Disease Pandemic (Covid-19)," *JPI (Jurnal Pendidikan Indonesia)*, vol. 10, no. 4, pp. 667–676, 2021.
- [11] D. I. Yani, J. Y. X. Chua, J. C. M. Wong, M. Pikkarainen, Y. S. S. Goh, and S. Shorey, "Perceptions of Mental Health Challenges and Needs of Indonesian Adolescents: A Descriptive Qualitative Study," *Int J Ment Health Nurs*, vol. 34, no. 1, p. e13505, 2025.
- [12] R. Y. Putri and S. Musrifah, "Assessment Of Child Violence: Protection Policies And The Impact Of Violence On Children," *MIKIA: Mimbar Ilmiah Kesehatan Ibu Dan Anak (Maternal and Neonatal Health Journal)*, pp. 128–138, 2022.
- [13] G. P. Laura Icela, R. M. María Soledad, and E. G. Juan Antonio, "Education 4.0 Maturity Models for Society 5.0: Systematic literature review," 2023. doi: 10.1080/23311975.2023.2256095.
- [14] N. V. Prasada and I. Ikrimach, "Employee attendance system using face recognition and GPS using local binary pattern histogram," *International Journal of Science, Technology, Engineering and Mathematics*, vol. 4, no. 4, pp. 83–107, 2024.
- [15] M. Shamila, B. Prakash, A. Ahmed, P. Prajeet, and R. Pant, "Smart Attendance Automation System," in *E3S Web of Conferences*, 2023. doi: 10.1051/e3sconf/202343001019.
- [16] S. Fakhar et al., "Smart Classroom Monitoring Using Novel Real-Time Facial Expression Recognition System," *Applied Sciences (Switzerland)*, vol. 12, no. 23, 2022, doi: 10.3390/app122312134.
- [17] M. N. Hasnine, H. T. Nguyen, T. T. T. Tran, H. T. T. Bui, G. Akçapınar, and H. Ueda, "A Real-Time Learning Analytics Dashboard for Automatic Detection of Online Learners' Affective States," *Sensors*, vol. 23, no. 9, 2023, doi: 10.3390/s23094243.
- [18] A. Nag, J. Ferdous, S. Uddin, and R. Debnath, "Real-Time and Automated Student Attention Monitoring System in Classroom Using Deep Learning Techniques," Aug. 2024. doi: 10.1109/ICCIT64611.2024.11021871.

- [19] A. Banerjee, O. C. Mutlu, A. Kline, P. Washington, D. Wall, and S. Surabhi, "Training and Profiling a Pediatric Facial Expression Classifier for Children on Mobile Devices: Machine Learning Study," *JMIR Form Res*, vol. 7, Aug. 2023, doi: 10.2196/39917.
- [20] C. Qian, J. A. L. Marques, A. R. de Alexandria, and S. J. Fong, "Application of Multiple Deep Learning Architectures for Emotion Classification Based on Facial Expressions," *Sensors (Basel)*, vol. 25, no. 5, p. 1478, 2025.
- [21] J. Lek and J. Teo, "Academic Emotion Classification Using FER: A Systematic Review," *Hum Behav Emerg Technol*, vol. 2023, pp. 1–27, Aug. 2023, doi: 10.1155/2023/9790005.
- [22] H. Chen, G. Zhou, and H. Jiang, "Student Behavior Detection in the Classroom Based on Improved YOLOv8," *Sensors*, vol. 23, p. 8385, Aug. 2023, doi: 10.3390/s23208385.
- [23] S. M and S. S. G, "Automatic Attendance Monitoring System Using Deep Learning," *EAI*, Aug. 2021. doi: 10.4108/eai.7-12-2021.2314601.
- [24] S. Aryal, R. Singh, A. Sood, and G. Thapa, "Automatic attendance system using deep learning," in *Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM)*, Amity University Rajasthan, Jaipur-India, 2019.
- [25] G. V. S. Suhas, T. D. Reddy, K. J. Bhoyania, L. Karthik, and M. Patil, "Real-time Student Attendance System with Emotion and Feedback Analysis," *Indiana Journal of Multidisciplinary Research*, vol. 4, no. 3, pp. 36–42, 2024.
- [26] A. Joseph, A. Nair, A. B. Koshy, M. Nallaballi, and N. Jayapandian, "Deep Learning Based Face Recognized Attendance Management System using Convolutional Neural Network," in *2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG)*, 2023, pp. 1–7. doi: 10.1109/ICTBIG59752.2023.10456032.
- [27] I. Sahputra, M. Fikry, and K. Kurniawati, "A Robust Approach to Student Attendance Using Web-Based Facial Recognition," *Proceedings of International Conference on Multidisciplinary Engineering (ICOMDEN)*, vol. 2, p. 78, Jan. 2025, [Online]. Available: <https://proceedings.unimal.ac.id/icomden/article/view/806>
- [28] G. Bhoomika, V. D. Pujitha, M. Sindusha, C. S. Rekha, and B. Suvama, "Facial Emotion Recognition: A Comparative Study of Pre-trained Deep Learning Models," in *2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC)*, 2024, pp. 377–382. doi: 10.1109/AIC61668.2024.10730808.
- [29] A. S. Ahmad, R. Hassan, N. H. Zakaria, and S. H. Moi, "Comparative studies of facial emotion detection in online learning," in *AIP Conference Proceedings*, 2023, p. 30009.
- [30] D. Tresnawati, S. Nurhidayanti, and N. Lestari, "A Comparison of YOLOv8 Series Performance in Student Facial Expressions Detection on Online Learning," *Jurnal Online Informatika*, vol. 10, no. 1, pp. 93–104, 2025.
- [31] A. Parmar, "Human Emotion Recognition Using Multi-modal Deep Learning: A Review of Methods, Datasets, and Challenges," *International Journal Of Scientific Research In Engineering And Management*, vol. 09, pp. 1–9, Aug. 2025, doi: 10.55041/IJSREM47033.
- [32] A. Imran, J. Li, and A. Alshammari, "AI-driven educational transformation in ICT: Improving adaptability, sentiment, and academic performance with advanced machine learning," *PLoS One*, vol. 20, Aug. 2025, doi: 10.1371/journal.pone.0317519.
- [33] Y. Lin, H. Chen, W. Xia, F. Lin, Z. Wang, and Y. Liu, "A Comprehensive Survey on Deep Learning Techniques in Educational Data Mining," *Data Sci Eng*, pp. 1–27, Aug. 2025, doi: 10.1007/s41019-025-00303-z.
- [34] Y. Sun, W. Huang, Z. Wang, X. Xu, M. Wen, and P. Wu, "Smart Teaching Systems: A Hybrid Framework of Reinforced Learning and Deep Learning," *International Journal of Emerging Technologies in Learning (IJET)*, vol. 18, pp. 37–50, Aug. 2023, doi: 10.3991/ijet.v18i20.44217.
- [35] Z. Li, "Network Intelligent Education System Based on the Deep Learning Algorithm," *Security and Communication Networks*, vol. 2022, pp. 1–7, Aug. 2022, doi: 10.1155/2022/5677089.
- [36] A. Mohamed, M. Ali, S. Ahmed, N. Hani, M. Hisham, and M. Mahmoud, "Detecting Student Disengagement in Online Classes Using Deep Learning: A Review," *arXiv preprint arXiv:2411.10464*, 2024.
- [37] X. Zhang and Z. Cao, "A Framework of an Intelligent Education System for Higher Education Based on Deep Learning," *International Journal of Emerging Technologies in Learning*, vol. 16, no. 7, 2021, doi: 10.3991/ijet.v16i07.22123.
- [38] D. Saez Trigueros, L. Meng, and M. Hartnett, "Face Recognition: From Traditional to Deep Learning Methods," *ArXiv*, Aug. 2018, doi: 10.48550/arXiv.1811.00116.
- [39] T. Pereira, D. Schmidli, Y. Linghu, X. Zhang, S. Marcel, and M. Günther, "Eight Years of Face Recognition Research: Reproducibility, Achievements and Open Issues," *ArXiv*, Aug. 2022.
- [40] S. Almadby and L. Elrefaei, "Deep Convolutional Neural Network-Based Approaches for Face Recognition," *Applied Sciences*, vol. 9, p. 4397, Aug. 2019, doi: 10.3390/app9204397.
- [41] K. Tirana and E. Bejleri, "FROM ORIGINS TO INNOVATIONS: AI'S ROLE AND THE COST IMPACT ON COMPUTER VISION," *AGORA INTERNATIONAL JOURNAL OF ECONOMICAL SCIENCES*, vol. 18, pp. 320–333, Aug. 2024, doi: 10.15837/aijes.v18i2.6960.
- [42] V. Espinosa-Duró and M. Faundez-Zanuy, "Face identification by means of a neural net classifier," Aug. 1999, pp. 182–186. doi: 10.1109/CCST.1999.797910.
- [43] F. Salinas, G. Ortiz, and P. Martínez, "Recognition of emotions through facial expression analysis," *South Florida Journal of Development*, vol. 2, pp. 2102–2118, Aug. 2021, doi: 10.46932/sfjdv2n2-076.
- [44] A. R. Livingston, *An efficient boosted classifier tree-based feature point tracking system for facial expression analysis*. Old Dominion University, 2012.
- [45] C. Mejia-Escobar, M. Cazorla, and E. Martinez-Martin, "Towards a Better Performance in Facial Expression Recognition: A Data-Centric Approach," *Comput Intell Neurosci*, vol. 2023, no. 1, 2023, doi: 10.1155/2023/1394882.
- [46] S. Kumar, B. Ghoraani, and A. Savakis, "Joint and discriminative dictionary learning for facial expression recognition," in *IS and T International Symposium on Electronic Imaging Science and Technology*, 2016. doi: 10.2352/ISSN.2470-1173.2016.11.IMAWM-455.
- [47] A. Muazu, E. Manzi, and J. Aminu, "A systematic review of methods of emotion recognition by facial expressions," *Int J Adv Res*, vol. 9, pp. 1141–1152, 2021.

- [48] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma, "Automated Facial Action Coding System for dynamic analysis of facial expressions in neuropsychiatric disorders," *J Neurosci Methods*, vol. 200, no. 2, 2011, doi: 10.1016/j.jneumeth.2011.06.023.
- [49] M. S. Ambare, "FER2013: Facial Expression Recognition Dataset," 2013, Kaggle. [Online]. Available: <https://www.kaggle.com/datasets/msambare/fer2013>
- [50] R. Bankar, N. Bargat, I. Hanmante, and Prof. H. Dakore, "Face Recognition Using Facenet Deep Learning Network for Attendance System," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2022, doi: 10.32628/cseit228630.
- [51] D. Sinha and M. El-Sharkawy, "Thin MobileNet: An Enhanced MobileNet Architecture," in 2019 IEEE 10th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2019, 2019, doi: 10.1109/UEMCON47517.2019.8993089.
- [52] B. Gašparović, G. Mauša, J. Rukavina, and J. Lerga, "Evaluating YOLOV5, YOLOV6, YOLOV7, and YOLOV8 in Underwater Environment: Is There Real Improvement?," in 2023 8th International Conference on Smart and Sustainable Technologies, SpliTech 2023, 2023, doi: 10.23919/SpliTech58164.2023.10193505.
- [53] N. Daud, A. I. Widjanarko, N. Sitanggang, and D. Kurnianingtyas, "Face Expression Recognition on Ojek Online for Safety Improvement Using Deep Learning," in 2024 11th International Conference on Soft Computing & Machine Intelligence (ISCMI), 2024, pp. 332–337, doi: 10.1109/ISCMI63661.2024.10851582.
- [54] E. S. Agung, A. P. Rifai, and T. Wijayanto, "Image-based facial emotion recognition using convolutional neural network on emognition dataset," *Sci Rep*, vol. 14, no. 1, p. 14429, 2024.
- [55] J. He, J. Zhou, J. Gan, D. Wu, and X. Wen, "Classroom Expression Classification Model Based on Multitask Learning," *Journal of Applied Sciences*, vol. 42, no. 6, p. 947, 2024, doi: 10.3969/j.issn.0255-8297.2024.06.005.
- [56] X. Kong, Y. Guo, Y. Ouyang, W. Cheng, M. Tao, and H. Zeng, "MT-RCAF: A Multi-Task Residual Cross Attention Framework for EEG-based emotion recognition and mood disorder detection," *Comput Methods Programs Biomed*, vol. 268, p. 108835, 2025.