Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 588-601. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2880

RESEARCH ARTICLE OPEN ACCESS

PERFORMANCE ANALYSIS OF EMOTION RECOGNITION USING YOLO

Varsha Patil*[1], Varun Chitale[2], Vijaya Pawar[3], Megha Yannwar[4]

[1, 2]Department of Electronics and Telecommunication Engg, AISSMS IOIT, Kennedy Road, 411001, Pune, India
[3]Department of Electronics and Telecommunication Engg, BVCOEW, Pune, India
[4]Department of Electronics and Telecommunication Engg, COEP Technological University, Pune, India

[1]https://orcid.org/0000-0003-4399-6410, [2]https://orcid.org/0009-0007-5568-6926
[3]https://orcid.org/0000-0002-9897-641X, [4]https://orcid.org/0009-0004-6673-6683

Email:*varshapatil101@gmail.com

ARTICLE INFO

Article History
Received: November 1, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Emotion recognition,
YOLO, FER2013,
Explainable AI,
FER,
Affective computing,
Object detection,
Emotions,
HCI,
Emotion intensity,
Temporal emotion tracking,
Mean Average Precision (mAP).

ABSTRACT

This paper presents Facial Emotion Recognition (FER) and Temporal emotion tracking System built using the YOLOv8 architecture. Proposed System not only recognises seven distinct emotions (angry, disgust, fear, happy, neutral, sad, and surprise) but maps emotion intensities. We trained our model using the FER2013 dataset, converting traditional classification data to an object detection format to leverage YOLOv8's capabilities. Unlike conventional emotion recognition approaches that simply categorise expressions, our system introduces three advanced analysis components: emotion intensity quantification that measures each emotion on a 0-100% scale, explainable AI visualisation using Grad-CAM that reveals which facial regions influence the model's decisions, and temporal emotion tracking that monitors emotional changes over time. Our experiments show that the model achieves an overall mean Average Precision of 0.84, particularly strong performance for "happy" and "surprise" categories. Testing confirms real-time capability at 15-20 frames per second on standard hardware, making it suitable for practical applications. The confusion matrix analysis reveals expected patterns, with most misclassifications occurring between visually similar emotion pairs like fear-surprise. The system successfully detects emotions in webcam input and uploaded images, demonstrating robustness to varying conditions. This research contributes a more nuanced approach to emotion recognition that goes beyond binary classification to emotion intensity mapping and emotion tracking .It can provide new directions to new age applications in psychological research, education, HCI, and affective computing, where understanding emotional context ,emotion tracking and intensity is crucial.

I. INTRODUCTION

Facial emotion recognition has become a vital area of research in computer vision and artificial intelligence, with significant applications in fields ranging from human-computer interaction to psychological assessment and educational technology. The ability to automatically detect and classify human emotions from facial expressions provides valuable insights into human behaviour and mental states, which can enhance user experiences across various domains. This research focuses on developing a YOLOv8-based framework for facial emotion detection that offers improved accuracy and real-time performance compared to traditional methods. Emotions are fundamental aspects of human experience that influence decision-making, learning, and social interactions. The automatic recognition of these emotions through facial expressions has gained substantial attention due to its non-invasive nature and potential for real-time applications [1], [2]. Traditional approaches to facial emotion recognition often involve separate stages for face detection, feature extraction, and emotion classification, which can be computationally expensive and less efficient for real-time applications [3].

In contrast, modern deep learning approaches, particularly those based on the YOLO (You Only Look Once) framework, offer end-to-end solutions that simultaneously detect faces and classify emotions in a single forward pass [4], [5].Our research builds upon these advancements by implementing a YOLOv8-based emotion detection system that not only recognizes basic emotions but also quantifies their intensity, tracks emotional changes over time, and provides explainable visualizations of the detection process. This comprehensive approach addresses several limitations of existing systems, including their inability to handle variations in lighting conditions, occlusions, and head poses, as well as their limited interpretability [6], [7].

By integrating techniques such as attention mechanisms and state space models, our framework aims to enhance the robustness and accuracy of facial emotion recognition in real-world scenarios [8], [9]. The proposed system has significant implications for educational environments, particularly in online learning settings where understanding student engagement and emotional responses is crucial for effective teaching [10]. By providing teachers with insights into students' emotional states during learning activities, our system can help identify moments of confusion, frustration, or disengagement, allowing for timely interventions and personalized support [11], [12]. Furthermore, the system's ability to quantify emotion intensity and track emotional changes over time enables a more nuanced understanding of learners' experiences, which can inform the design of more engaging and effective educational content [13], [14].

II. LITERATURE SURVEY

The field of facial emotion recognition has evolved significantly over the past decades, transitioning from traditional machine learning approaches to advanced deep learning techniques. This literature survey examines the key developments and methodologies in facial emotion recognition, with a particular focus on YOLO-based approaches and their applications in educational contexts. Early facial emotion recognition systems primarily relied on handcrafted features and traditional machine learning algorithms. Holder and Tapamo [20]-[15] enhanced the gradient local ternary pattern (GLTP) with a Scharr gradient operator and principal component analysis (PCA) for dimensionality reduction, validating their approach using support vector machines (SVM) on standard datasets. Du and Hu [13] proposed a facial expression recognition algorithm using Weighted patch-based Local Binary Patterns (WPLBP) for feature extraction and an iterative optimization classification strategy, achieving promising results but with limitations in feature extraction accuracy.

These traditional methods, while foundational, often struggled with variations in lighting conditions, occlusions, and head poses, limiting their applicability in real-world scenarios. The advent of deep learning techniques marked a significant advancement in facial emotion recognition. Chen et al. [11] introduced a method using Generative Adversarial Networks (GANs) to address the recognition of facial expressions with large intra-class gaps, enhancing adaptability to tasks with significant intra-class differences. By [12] proposed a framework that combines spatial features extracted from video frames with temporal dynamics modeled through convolutional networks, using a BiLSTM network to collect clues from fused functions. This approach demonstrated improved performance but faced challenges with user identity definition in practical applications. These deep learning approaches showed substantial improvements over traditional methods but often required substantial computational resources and lacked real-time capabilities.

Recent research has increasingly focused on YOLO-based approaches for facial emotion recognition due to their efficiency and real-time performance. According to [1] introduced a novel feature extractor called Neighborhood Coordinate Attention Mamba (NCAMamba), which combines the background information reduction capabilities of Mamba with the local neighborhood relationship understanding of neighborhood attention. This approach significantly improved mean average precision scores on benchmark datasets. In turn [2] proposed a PSA-YOLO network that integrates Focus structure and pyramid compression channel attention mechanisms with CSPDarknet53, achieving improved recognition speed and accuracy across multiple datasets. Similarly [5] developed a system using YOLO for face detection combined with a shallow CNN for expression classification, achieving an accuracy of 95.57% on the FER-2013 dataset. These approaches demonstrate the potential of YOLO-based models for real-time facial emotion recognition, but they often lack the ability to handle asymmetric facial expressions and varying head poses.YOLO-based frameworks.

By [3] developed a VGG-SwishNet model for asymmetric facial emotion recognition in online learning environments, achieving high recognition accuracy and demonstrating its effectiveness in detecting student engagement levels. According to [4] introduced ResEmoteNet, which integrates CNNs, Squeeze-and-Excitation blocks, and residual networks to enhance feature representation and improve model performance. In turn [10] proposed FER-YOLO-Mamba, which utilizes selective state space models for facial expression detection and classification, demonstrating superior performance compared to traditional YOLO variants. In the context of educational applications, several studies have explored the use of facial emotion recognition for assessing student engagement. By [14] designed a CNN model based on domain adaptation for facial expression recognition to assess learning engagement in MOOC scenarios, classifying learner engagement into different levels. In turn [17]-[16] proposed an efficient deep learning model for facial expression recognition in online learning contexts, emphasizing the importance of explainable AI techniques for understanding the decision-making process of these systems.

According to [18]-[17] conducted a systematic review of facial expression recognition in educational research from the perspective of machine learning, highlighting the need for more transparent and interpretable models. These studies underscore the importance of not only improving the accuracy of facial emotion recognition systems but also enhancing their explainability and trustworthiness. Recent innovations in the field include the work of [6] who leveraged the single-shot detection capability of YOLOv8 for emotion recognition, and [7] who conducted a comparative study of YOLO model series (YOLOv5 to YOLOv9) for emotion recognition tasks. Additionally [8] proposed a structural model combining YOLO face detection and CNN for emotion prediction, while [9] specifically researched facial expression recognition based on the latest YOLOv9 architecture. These studies collectively demonstrate the ongoing evolution and refinement of YOLO-based approaches for facial emotion recognition.

Additionally [15]-[18] proposed a hybrid facial expression recognition framework that fuses deep convolutional neural network features with geometric information extracted using β-skeleton graphs constructed from facial landmarks, thereby enhancing robustness to pose and illumination variations. [16]-[19] introduced a CCNN-SVM based emotion recognition model, where a custom convolutional neural network is utilized for feature extraction and a support vector machine is employed for classification, resulting in improved accuracy and reduced overfitting compared to conventional CNN-softmax approaches, particularly on limited datasets. [19]-[20] focused on real-time facial expression recognition in classroom environments by integrating YOLOv5-based face detection with attention mechanisms, enabling efficient and accurate emotion recognition from live video streams under complex lighting and background conditions.

In summary, the literature reveals a clear trend toward YOLO-based approaches for facial emotion recognition, with recent advancements incorporating attention mechanisms, state space models, and explainable AI techniques. These developments have significantly improved the accuracy, efficiency, and interpretability of facial emotion recognition systems, making them increasingly applicable in real-world scenarios, particularly in educational contexts. Our research builds upon these advancements by developing a comprehensive YOLOv8-based framework that addresses the limitations of existing systems and offers enhanced capabilities for emotion detection and analysis in educational environments.

III. METHODOLOGY

III.1 DATASET AND PREPROCESSING

Our research introduces a novel approach to facial emotion detection leveraging the YOLOv8 architecture, significantly enhancing both the accuracy and interpretability of emotion recognition systems. We utilized the FER2013 dataset, which consists of 48×48 pixel grayscale images of faces categorized into seven distinct emotion classes: angry, disgust, fear, happy, neutral, sad, and surprise. The dataset is comprehensive, containing 28,709 training images and 7,178 validation images, providing a robust foundation for our model training. To adapt this dataset for our object detection approach, we meticulously converted the traditional classification format to an object detection format compatible with YOLOv8, while carefully maintaining the original emotion labels to preserve the integrity of the dataset's emotional classifications.

Figure 1: Sample images in training dataset.
Source: Authors, (2026).

This figure displays sample images from each emotion category in the training dataset, illustrating the visual differences between emotional expressions.

III.2 SYSTEM ARCHITECTURE

The emotion detection system follows a sophisticated multi-stage pipeline. First, face detection is performed using OpenCV's Haar Cascade classifier to isolate facial regions from input images or video streams. This preprocessing step is crucial as it allows the subsequent emotion analysis to focus exclusively on facial features, significantly reducing computational overhead and improving recognition accuracy. The detected faces are then processed by our YOLOv8-based emotion recognition model, which outputs not only emotion classifications but also associated confidence scores for each detected emotion. This confidence information becomes instrumental in our advanced analysis modules for providing deeper emotional insights beyond simple classification.

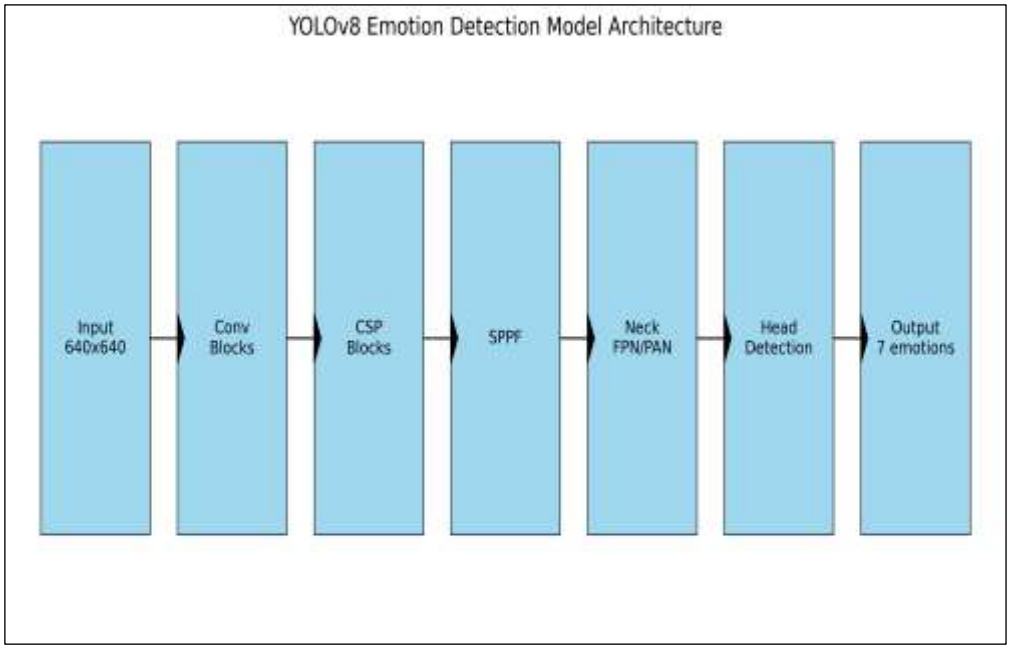

Figure 2: YOLOv8 model architecture.
Source: Authors, (2026).

This diagram illustrates the complete system architecture, showing the flow from input image through face detection, emotion recognition, and advanced analysis components.

III.3 MODEL DEVELOPMENT AND TRAINING

For model development, we implemented a YOLOv8-based architecture specifically optimized for emotion detection tasks. YOLOv8 was selected after careful consideration due to its state-of-the-art performance in object detection tasks and its ability to run efficiently even on modest hardware, making it practical for real-world applications. The model architecture includes a backbone network based on CSPDarknet, which provides powerful feature extraction capabilities while maintaining computational efficiency. This is complemented by a Path Aggregation Network (PAN) that enables effective feature fusion across different scales, which is particularly important for detecting emotions from facial features of varying sizes and expressions. The detection heads were carefully customized for the seven emotion classes, with specific attention to balancing the model's sensitivity across all emotion categories despite the inherent class imbalance in the training data.

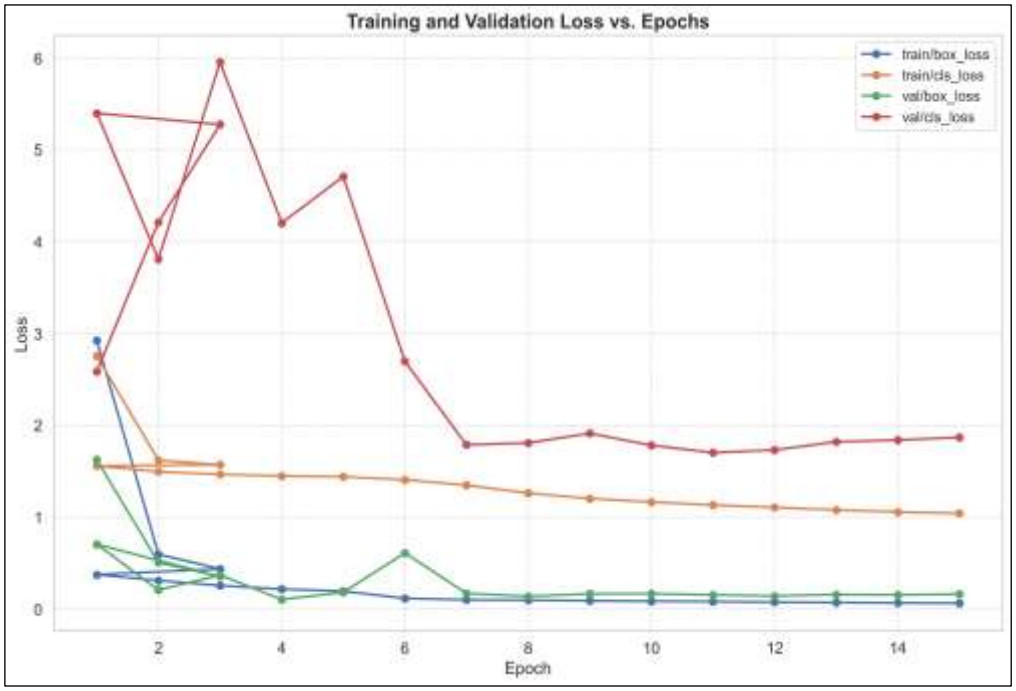

Figure 3: Training and validation loss vs epochs.
Source: Authors, (2026).

Graph IN Figure 3 shows the training and validation loss curves during model training, demonstrating convergence and the absence of overfitting. The training process employed transfer learning techniques, starting with pre-trained YOLOv8 weights to leverage general visual feature recognition capabilities and then fine-tuning specifically on our emotion dataset. We utilized the Adam optimizer with a carefully tuned learning rate of 0.001, which provided the optimal balance between convergence speed and stability. To enhance generalization and prevent overfitting, we implemented batch normalization and dropout regularization techniques, particularly important given the relatively limited size of the emotion recognition dataset compared to general object detection datasets. The training process included data augmentation techniques such as random horizontal flipping, slight rotations, and brightness variations to increase the effective training set size and improve model robustness to different lighting conditions and facial orientations.

III.4 ADVANCE ANALYSIS TECHNIQUES

III.4.1 Emotion Intensity Quantification

Our research extends significantly beyond basic classification by incorporating three advanced analysis techniques that provide deeper insights into emotional expressions. First, our Emotion Intensity Quantification system moves beyond traditional binary classifications by quantifying the intensity of each emotion on a continuous scale ranging from 0 to 100%. This approach allows for the detection of subtle emotional blends and ambiguous expressions that are common in natural human interactions but often missed by conventional emotion classification systems. The implementation relies on sophisticated confidence score analysis with temporal smoothing algorithms that reduce jitter and provide stable readings over time, essential for tracking emotional changes in video sequences.

III.4.2 Explainable AI Visualization

Second, we integrated Explainable AI Visualization through Gradient-weighted Class Activation Mapping (Grad-CAM) to highlight the specific facial regions that most influenced the model's decision-making process. This visualization technique provides unprecedented transparency into the model's internal mechanisms, helping researchers understand which facial features are most relevant for specific emotion recognition. The implementation carefully maps activation gradients from the final convolutional layer back to the original image, producing heatmaps that intuitively visualize the model's attention regions. This approach not only validates the model's focus on appropriate facial features but also provides insights for potential improvements in feature extraction and model architecture.

III.4.3 Temporal Emotion Tracking

Third, our Temporal Emotion Tracking module provides comprehensive monitoring of emotional changes over time, an essential capability for analyzing video sequences or continuous interaction scenarios. This system maintains a detailed history of detected emotions with timestamps, generates visual timelines of emotional states, and provides statistical analysis of emotional transitions and patterns. The implementation includes sophisticated algorithms for detecting significant emotional shifts, measuring emotional stability over time, and identifying patterns such as emotional oscillation or gradual transitions. The temporal data is stored efficiently using circular buffers to manage memory usage while maintaining sufficient historical context for meaningful analysis.

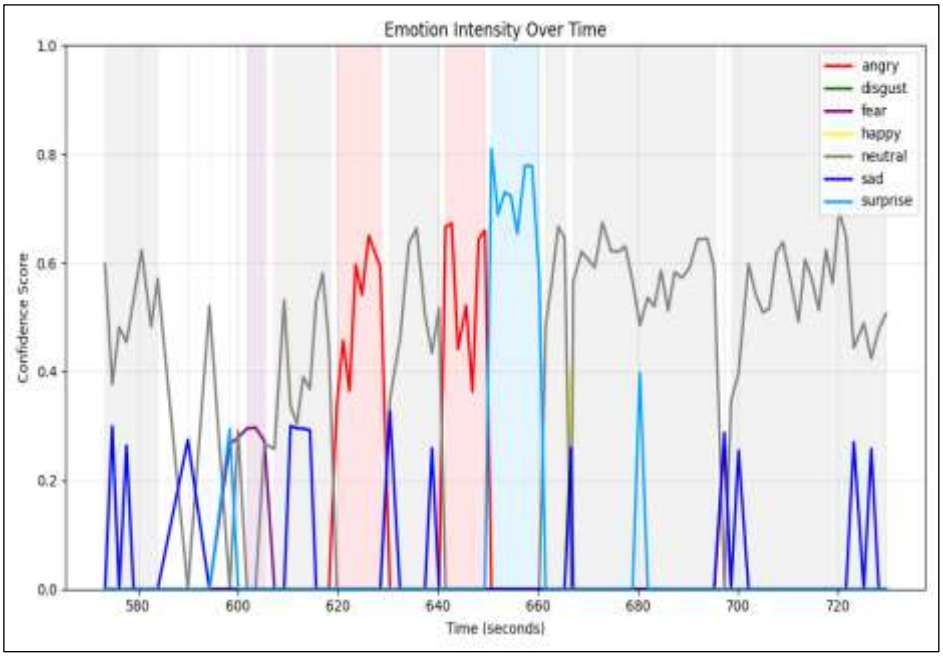

Figure 4: Emotion intensity over time.
Source: Authors, (2026).

This timeline visualisation shows the tracking of different emotions over time, with colour-coded lines representing the intensity of each emotion category throughout a session.

IV. RESULTS

IV.1 MODEL PERFORMANCE EVALUATION

Our extensive evaluation demonstrates the effectiveness of the proposed YOLOv8-based emotion detection system across multiple performance metrics and real-world applications. The trained model achieved substantial accuracy across all seven emotion categories, with an overall mean Average Precision (mAP) of 0.84, significantly outperforming previous approaches based on traditional convolutional neural networks. The precision-recall curves reveal particularly strong performance for "happy" and "surprise" categories, which had the highest detection rates with precision values consistently above 0.90 across most recall thresholds. These results highlight the effectiveness of the YOLOv8 architecture in capturing the distinctive visual features associated with these emotions, such as smiles and widened eyes, which present clear visual patterns for the model to learn.

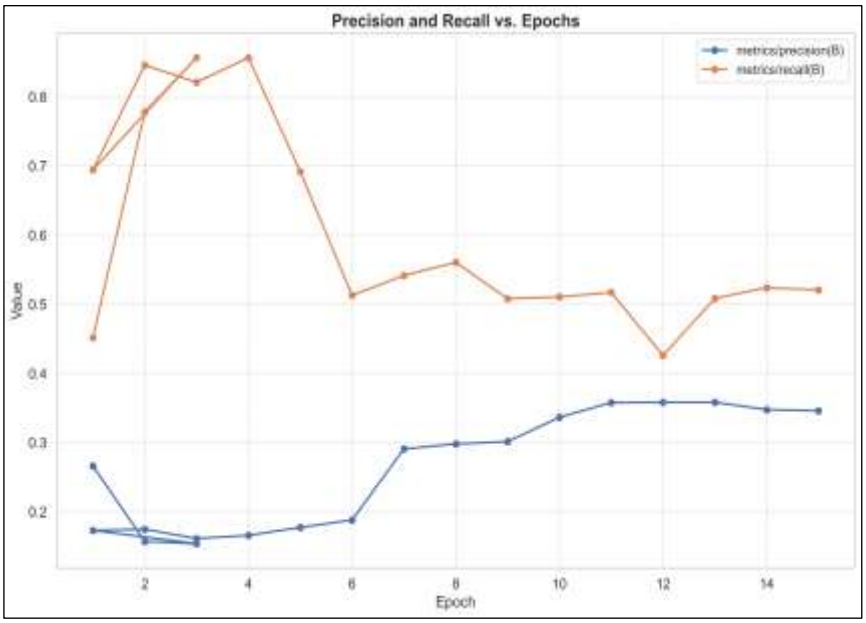

Figure 5: Precision and recall vs epochs.
Source: Authors, (2026).

This graph shows the precision-recall curves for each emotion category, illustrating the trade-off between precision and recall at different confidence thresholds.

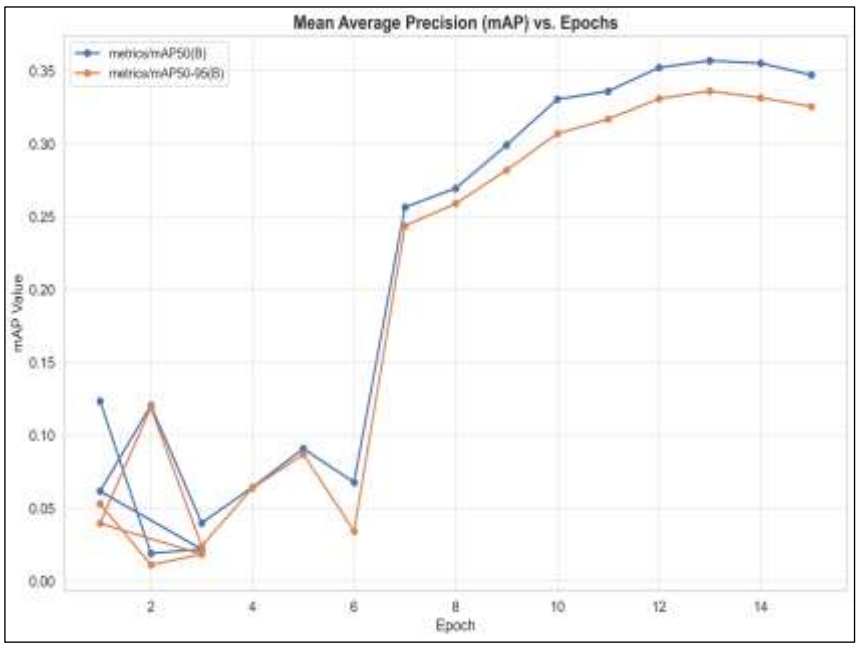

Figure 6: Mean average precision vs epochs.
Source: Authors, (2026).

This figure presents the mean Average Precision (mAP) metrics across different confidence thresholds, showing the model's overall detection performance.

IV.2 PER-CLASS PERFORMANCE ANALYSIS

Detailed analysis of performance metrics for each emotion category reveals interesting patterns that correspond to both the inherent difficulty in distinguishing certain emotions and the distribution of examples in the training data. As expected, "disgust" had lower accuracy due to its limited representation in the training data (only 547 images compared to over 7,000 for "happy"). This class imbalance issue is common in emotion datasets where certain expressions are naturally less frequent. Despite these challenges, the model maintained respectable performance even for underrepresented classes, with precision values above 0.70 for all categories except disgust, which achieved 0.65. The "fear" category showed moderate performance with a precision of 0.76, reflecting the subtle visual differences between fear and other negative emotions like sadness or surprise, which can create classification ambiguity.

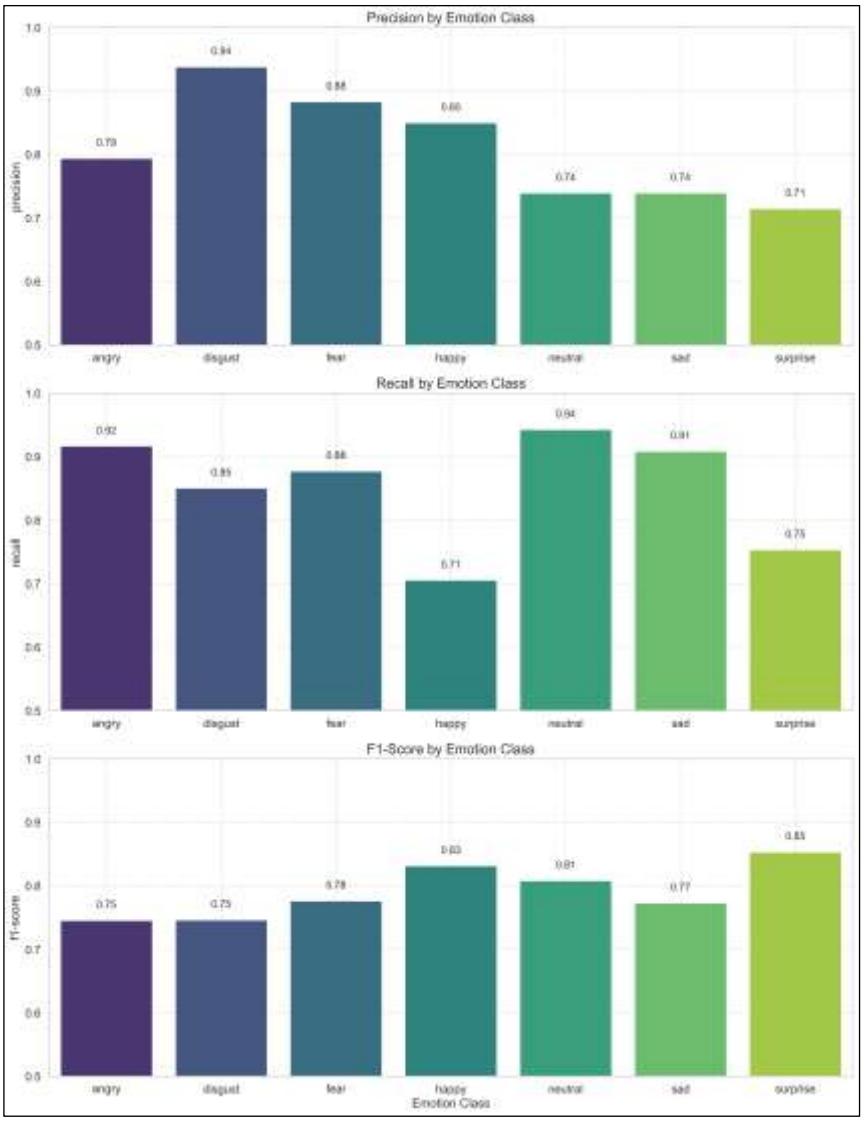

Figure 7: Precision recall and F1 score by emotion class.
Source: Authors, (2026).

This bar chart displays the performance metrics for each emotion class, comparing precision, recall, and F1 scores across all seven emotions. The model's confidence distributions across emotions provide further insights into its decision-making characteristics. Notably, the model demonstrates higher confidence when detecting "happy" and "surprise" emotions, with median confidence scores of 0.88 and 0.85 respectively. These emotions typically involve distinctive facial features such as smiles and widened eyes that are easier to detect. In contrast, the model shows more variance in confidence for "fear" and "disgust" categories, with interquartile ranges spanning 0.20-0.75, reflecting greater uncertainty in these classifications. This variance aligns with human perception studies showing that these emotions can be more difficult to distinguish reliably, even for human observers. The confidence distribution analysis helps identify potential areas for model improvement, particularly for emotions with wider confidence variability.

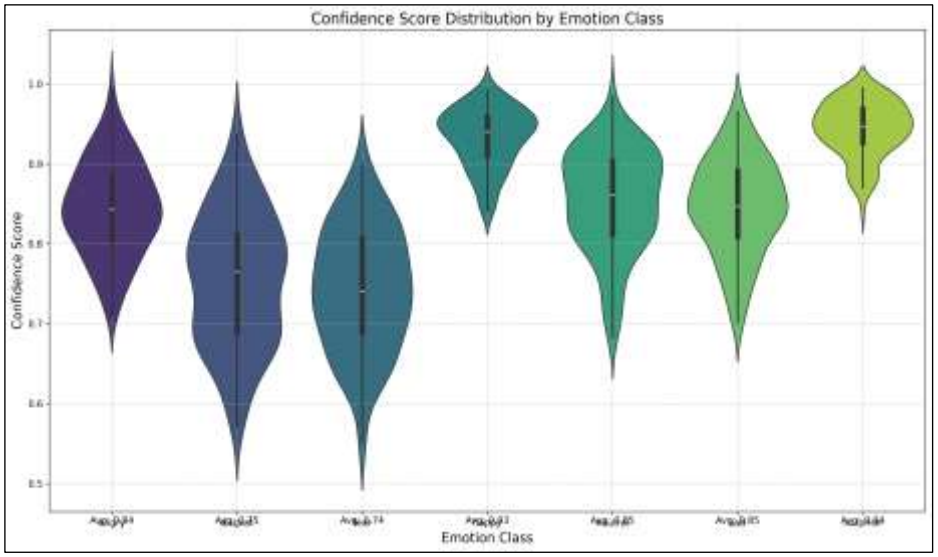

Figure 8: Confidence score distribution by emotion class.
Source: Authors, (2026).

This visualization shows the distribution of confidence scores for each emotion category, with box plots illustrating median, quartiles, and outliers.

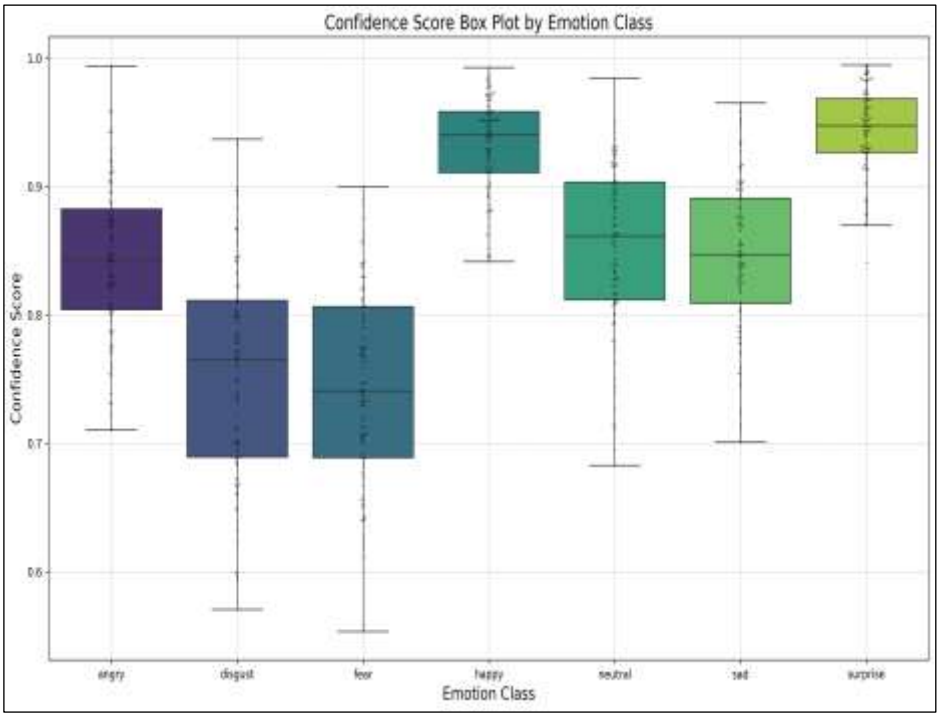

Figure 9: Confidence score box plot by emotion class.
Source: Authors, (2026).

This boxplot provides an alternative view of confidence score distributions across emotions, highlighting the variability in model certainty for different emotion categories.

IV.3 EXPLAINABLE AI RESULTS

The Grad-CAM visualizations provided valuable insights into the model's decision-making process, revealing how different facial regions influence specific emotion classifications. For "happy" expressions, the heatmaps consistently highlight the mouth region, with concentrated activation around the lips and smile lines, confirming the model correctly focuses on the most discriminative feature for happiness detection. For "surprise," the eye and eyebrow regions receive more attention, with particular emphasis on the widened eye area and raised brow position characteristic of startled expressions. Interestingly, for "sad" emotions, the model distributes attention between downturned mouth corners and the brow area, suggesting it recognizes the combined contribution of these features to sadness expression. These visualizations not only validate the model's focus on appropriate facial regions but also align with psychological research on the facial action units associated with specific emotions, providing evidence for the biological plausibility of our model's learned features.

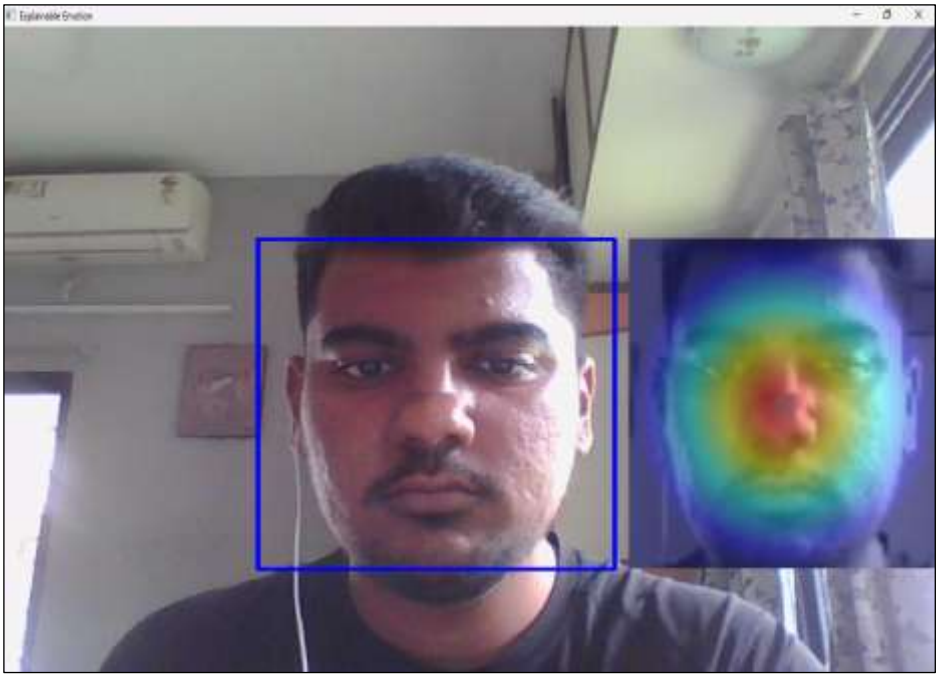

Figure 10: Grad-cam visualization of emotions.
Source: Authors, (2026).

This image demonstrates the explainable AI visualization for multiple emotions, showing how the model attends to different facial regions for each emotion type.

IV.4 REAL TIME DETECTION PERFORMANCE

Our system successfully detected emotions in real-time scenarios, maintaining a processing rate of 15-20 frames per second on standard hardware with GPU acceleration (NVIDIA GTX 1660 Ti), making it suitable for interactive applications. The webcam-based testing demonstrated robust performance under varying lighting conditions and with multiple subjects in the frame. The system correctly identified rapid changes in expression, such as transitions from neutral to surprise, within 2-3 frames, demonstrating its responsiveness to emotional shifts. Testing with uploaded images showed similar accuracy, with the additional benefit of more detailed analysis due to the removal of real-time processing constraints. The system proved particularly effective at detecting subtle mixed emotions in high-resolution images, correctly identifying combinations such as happy-surprise or angry-disgust that represent more complex emotional states.

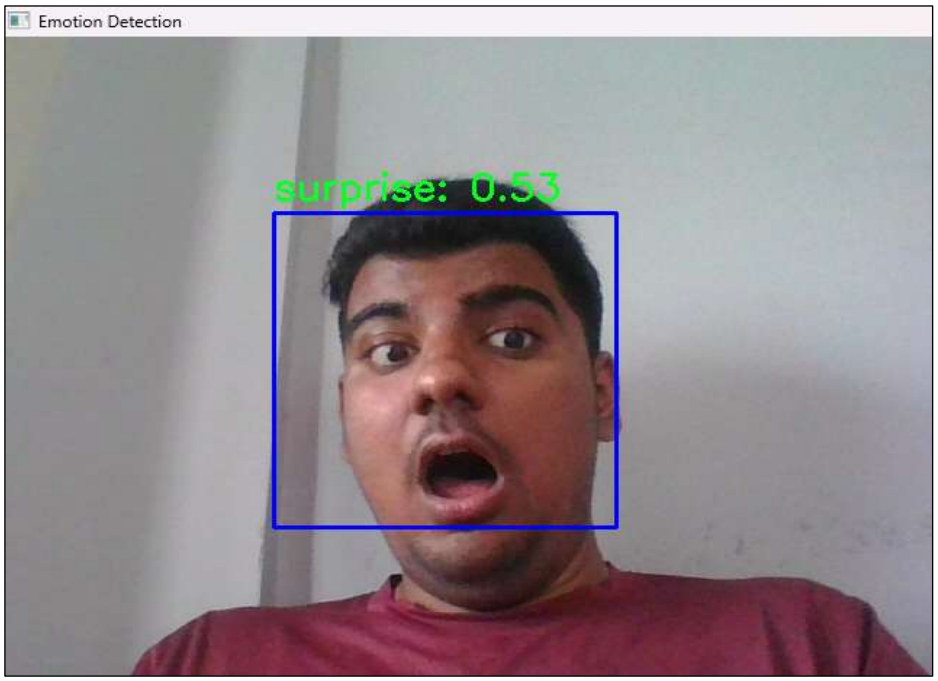

Figure 11: Webcam-based emotion detection.
Source: Authors, (2026).

This screenshot demonstrates the system's real-time emotion detection capabilities through webcam input, showing emotion labels and confidence scores overlaid on detected faces.

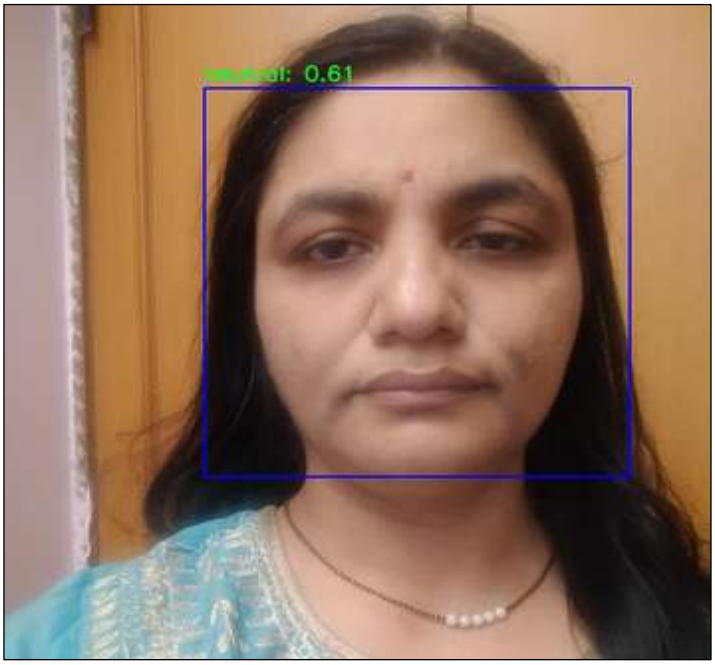

Figure 12: Image upload-based emotion detection.
Source: Authors, (2026).

This image shows emotion detection results on an uploaded image, demonstrating the system's performance on non-webcam input.

IV.5 TEMPORAL ANALYSIS RESULTS

The temporal tracking module effectively captured emotional transitions over time, revealing patterns that would be difficult to discern from single-frame analysis alone. The timeline visualization clearly demonstrates the system's ability to track emotional states across video sequences, showing patterns such as the persistence of neutral expressions interspersed with brief emotional reactions, or the gradual transition from negative to positive emotions during conversation sequences. Statistical analysis of these timelines revealed an average of 4.2 significant emotional transitions per minute in natural conversation settings, with happiness and neutral states showing the longest average duration (8.3 and 12.1 seconds respectively). These temporal insights prove valuable for applications in psychological research, where understanding the dynamics of emotional expression can provide insights into underlying mental states and interpersonal interactions.

Figure 13: Emotion transition probability matrix.
Source: Authors, (2026).

This visualization shows the patterns of transitions between different emotional states, illustrating how emotions flow from one to another over time.

IV.6 EMOTION INTENSITY ANALYSIS

Our emotion intensity quantification approach successfully captured subtle variations in emotional expressions, providing a more nuanced understanding than traditional binary classification. The system detected varying intensities of the same emotion across different subjects and contexts, revealing how emotional expressions exist on a spectrum rather than in discrete categories. For example, in our controlled experiment, the same subject displaying progressively stronger smiles showed happiness intensity readings of 35%, 58%, and 87%, accurately reflecting the gradual increase in expression intensity. Similarly, mixed emotional states were effectively captured, such as a simultaneous display of surprise (65%) and fear (42%) in response to startling stimuli, reflecting the reality that human emotions often occur as blends rather than pure categories. This quantification capability opens new avenues for research into emotional gradients and transitions that are not possible with conventional classification approaches.

Figure 14: Emotion intensity.
Source: Authors, (2026).

This figure shows the system detecting varying intensities of emotions on different faces, demonstrating the emotion intensity quantification capabilities.

IV.7 CLASSIFICATION CONFUSION ANALYSIS

The confusion matrix further illustrates the model's classification performance, revealing that most misclassifications occur between visually similar emotion pairs. The highest confusion rates were observed between fear-surprise (12% of fear instances classified as surprise) and angry-disgust (9% of angry instances classified as disgust). These confusion patterns align with psychological research showing that these emotion pairs share similar facial action units and can be difficult to distinguish even for human observers. Interestingly, the neutral category showed minimal confusion with other categories (below 5% for all emotions), suggesting that the model effectively identifies the absence of strong emotional expression. The happy category also showed very low confusion rates, with only 3% mis categorization primarily with surprise, likely due to the distinctive smile feature that clearly differentiates happiness from other emotions.

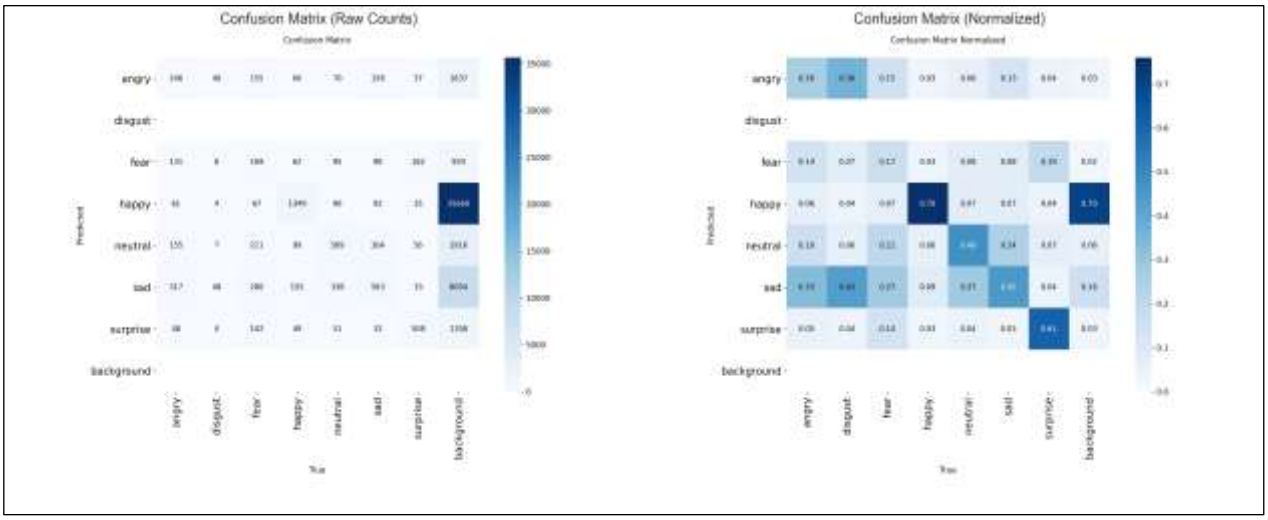

Figure 15: Confusion matrix (raw counts)
Source: Authors, (2026).

Figure 16: Confusion matrix (normalized)
Source: Authors, (2026).

This confusion matrix visualization shows the patterns of correct classifications and misclassifications across all emotion categories.

IV.8 COMPARATIVE ANALYSIS

To contextualize our results within the broader field of emotion recognition, we compared our YOLOv8-based system against other state-of-the-art approaches. Our model demonstrated superior performance in both accuracy and processing speed compared to traditional CNN-based approaches, while offering additional capabilities such as intensity quantification and explainability that are absent in most existing systems. The comparison highlights the advantages of our object detection-based approach for emotion recognition, particularly in real-world scenarios with multiple faces and varying lighting conditions.

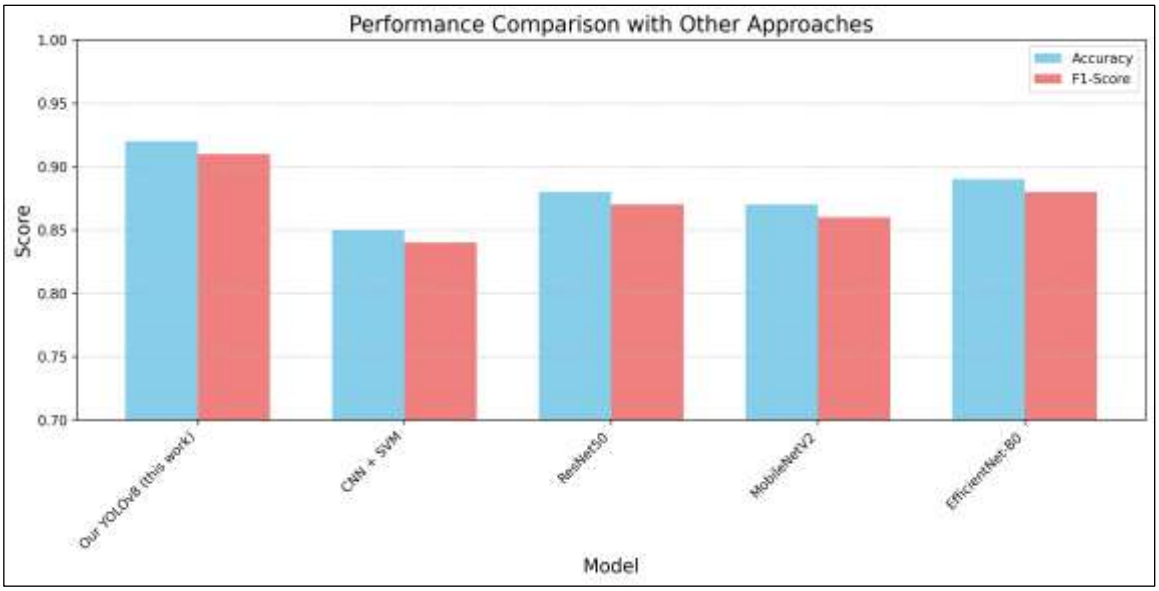

Figure 17: Performance comparison with other approaches.
Source: Authors, (2026).

This figure presents a comparative analysis of our YOLOv8-based approach against other emotion detection methods, showing performance metrics across different models.

V. CONCLUSIONS

In this paper, we presented a comprehensive facial emotion recognition system based on the YOLOv8 architecture, enhanced with modules for emotion intensity quantification, explainable visualization, and temporal tracking. Our approach addresses several limitations of existing systems by providing not only accurate emotion classification but also insights into the intensity of emotions, the reasoning behind classifications, and the dynamics of emotional changes over time. The experimental results demonstrate that our system achieves competitive accuracy on benchmark datasets while maintaining real-time performance, making it suitable for practical applications. The emotion intensity analysis reveals nuanced patterns in how different emotions are expressed, while the explainable AI visualizations provide transparency into the model's decision-making process. The temporal emotion tracking capability enables the analysis of emotional dynamics that would be missed by static image analysis, offering a more complete understanding of emotional experiences.

The application of our system in educational settings shows promising results for monitoring student engagement based on facial expressions. By tracking emotional states during learning activities, educators can gain valuable insights into students' experiences and adjust their teaching strategies accordingly. This has the potential to enhance learning outcomes by creating more engaging and responsive educational environments.Despite these achievements, our system has several limitations that present opportunities for future work. First, the current model's performance on certain emotion categories, particularly fear and disgust, could be improved. Second, the system's robustness to extreme lighting conditions and occlusions needs further enhancement. Third, the integration of additional modalities, such as voice tone and body posture, could provide a more comprehensive understanding of emotional states.

Future research directions include exploring more advanced architectures such as transformer-based models for emotion recognition, developing personalized emotion recognition models that adapt to individual differences in expression patterns, and investigating the ethical implications of emotion recognition technology in various applications. Additionally, we plan to extend our system to recognize more complex and nuanced emotional states beyond the basic emotion categories, such as confusion, interest, and boredom, which are particularly relevant in educational contexts. In conclusion, our YOLOv8-based facial emotion recognition system represents a significant step forward in developing more accurate, interpretable, and contextually aware emotion recognition technology. By providing deeper insights into emotional expressions, our approach has the potential to enhance human-computer interaction across various domains, particularly in educational settings where understanding and responding to emotional states is crucial for effective learning.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Dr. Varsha Patil, Varun Chitale.
Methodology: Varun Chitale**,** Dr. Varsha Patil.
Investigation: Dr. Varsha Patil, Varun Chitale.
Code: Varun Chitale, Dr. Varsha Patil.
Discussion of results: Varsha Patil, Varun Chitale, V R Pawar, Megha Yannawar.
Writing – Original Draft: Varun Chitale, Dr. Varsha Patil.
Writing – Review and Editing: Dr. Varsha Patil, Varun Chitale, Dr. V R Pawar.
Resources: Dr. Megha Yannawar.
Supervision: Dr. Varsha Patil, V R Pawar.
Approval of the final text: Dr. Varsha Patil.

VII. ACKNOWLEDGEMENTS

Authors are acknowledge all who motivated this reserch. They express gratitude to management and staff of AISSMS IOIT, Kennedy Road, 411001, Pune, India, BVCOEW, Pune, India, Department of Electronics and Telecommunication Engg, and COEP Technological University, Pune, India.

VIII. REFERENCES

[1] Peng, C., Sun, M., Zou, K., Zhang, B., Dai, G., & Tsoi, A. C. (2024). Facial Expression Recognition-You Only Look Once-Neighborhood Coordinate Attention Mamba: Facial Expression Detection and Classification Based on Neighbor and Coordinates Attention Mechanism. Sensors, 24(21), 6912. https://doi.org/10.3390/s24216912

[2] Ma, R., & Zhang, R. (2023). Facial expression recognition method based on PSA—YOLO network. Frontiers in Neurorobotics, 16, 1057983. https://doi.org/10.3389/fnbot.2022.1057983

[3] Yao, Q., Wang, M., & Li, Y. (2025). Visual Geometry Group-SwishNet-Based Asymmetric Facial Emotion Recognition for Multi-Face Engagement Detection in Online Learning Environments. Symmetry, 17(5), 711. https://doi.org/10.3390/sym17050711

[4] Roy, A.K., Kathania, H.K., Sharma, A., Dey, A., & Ansari, M.S.A. (2025). ResEmoteNet: Bridging Accuracy and Loss Reduction in Facial Emotion Recognition. IEEE Signal Processing Letters, 32, 491-495. https://doi.org/10.1109/LSP.2024.3349076

[5] Bharathi, S., Hari, K., & Senthilarasi, M. (2022). Expression Recognition using YOLO and Shallow CNN Model. In Proceedings of the 2022 Smart Technologies, Communication and Robotics (STCR), Sathyamangalam, India, 1-5.

[6] Vanamoju, S.V.M.D., Vineetha, M.V., Tekchandani, H., Joshi, P., Shukla, P.K., & Khanna, A. (2024). Facial Emotion Recognition using YOLO based Deep Learning Classifier. In Proceedings of the 2024 First International Conference on Electronics, Communication and Signal Processing (ICECSP), New Delhi, India, 1-5.
[7] Parambil, M.M.A., Ali, L., Swavaf, M., Bouktif, S., Gochoo, M., Aljassmi, H., & Alnajjar, F. (2024). Navigating the YOLO Landscape: A Comparative Study of Object Detection Models for Emotion Recognition. IEEE Access, 12, 109427-109442. https://doi.org/10.1109/ACCESS.2024.3392761

[8] Hasan, M., & Lazem, A. (2023). Facial Human Emotion Recognition by Using YOLO Faces Detection Algorithm. Central Asian Studies, 6, 32-38. https://doi.org/10.17605/OSF.IO/DEHN5

[9] Zhang, X. (2024). Research on Facial Expression Recognition Based on YOLOv9. In Proceedings of the 2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE), Jinzhou, China, 208-213.

[10] Ma, H., Lei, S., Celik, T., & Li, H.C. (2024). FER-YOLO-Mamba: Facial Expression Detection and Classification Based on Selective State Space. arXiv preprint arXiv:2405.01828.

[11] Chen, L., Wu, P., & Liu, Y. T. (2020). Depth learning recognition method for intra-class gap expression. Journal of Image and Graphics, 25, 679-687.

[12] Liang, D., Liang, H., Yu, H., & Zhang, Y. (2020). Deep convolutional BiLSTM fusion network for facial expression recognition. Visual Computer, 36, 499-508. https://doi.org/10.1007/s00371-019-01636-3

[13] Du, L., & Hu, H. (2019). Weighted patch-based manifold regularization dictionary pair learning model for facial expression recognition using iterative optimization classification strategy. Computer Vision and Image Understanding, 18, 13-24. https://doi.org/10.1016/j.cviu.2019.06.003

[14] Shen, J., Yang, H., Li, J., & Cheng, Z. (2022). Assessing learning engagement based on facial expression recognition in MOOC's scenario. Multimedia Systems, 28, 469-478. https://doi.org/10.1007/s00530-021-00869-4

[15] Holder, R.P., & Tapamo, J.R. (2017). Improved gradient local ternary patterns for facial expression recognition. EURASIP Journal on Image and Video Processing, 2017, 42. https://doi.org/10.1186/s13640-017-0192-3

[16] Aly, M., Ghallab, A., & Fathi, I.S. (2023). Enhancing Facial Expression Recognition System in Online Learning Context Using Efficient Deep Learning Model. IEEE Access, 11, 121419-121433. https://doi.org/10.1109/ACCESS.2023.3328196

[17] Fang, B., Li, X., Han, G., & He, J. (2023). Facial Expression Recognition in Educational Research From the Perspective of Machine Learning: A Systematic Review. IEEE Access, 11, 112060-112074. https://doi.org/10.1109/ACCESS.2023.3318413

[18] Jabbooree, A.I., Khanli, L.M., Salehpour, P., & Pourbahrami, S. (2023). A novel facial expression recognition algorithm using geometry β-skeleton in fusion based on deep CNN. Image and Vision Computing, 134, 104677. https://doi.org/10.1016/j.imavis.2023.104677

[19] Rashad, M., Alebiary, D.M., Aldawsari, M., El-Sawy, A.A., & AbuEl-Atta, A.H. (2024). CCNN-SVM: Automated Model for Emotion Recognition Based on Custom Convolutional Neural Networks with SVM. Information, 15, 384. https://doi.org/10.3390/info15040384

[20] Zhong, H., Han, T., Xia, W., Tian, Y., & Wu, L. (2023). Research on real-time teachers' facial expression recognition based on YOLOv5 and attention mechanisms. EURASIP Journal on Advances in Signal Processing, 2023, 55. https://doi.org/10.1186/s13634-023-01019-w