ISSN: 0976-9102 (ONLINE)
DOI: 10.21917/ijivp.2025.0527

GENERATIVE AI AND YOLO FRAMEWORK FOR REAL-TIME SENTIMENT DETECTION AND ANALYSIS OF CROWDS IN PUBLIC SPACES TO ENHANCE SECURITY AND BEHAVIORAL INSIGHTS

B. Yuvaraj[1], T. Ganesan[2], D.C. Jullie Josephine[3] and S. Thumilvannan[4]
[1]*Department of Artificial Intelligence and Machine Learning, Kings Engineering College, India*
[2]*Department of Artificial Intelligence and Data Science, Kings Engineering College, India*
[3,4]*Department Computer Science and Engineering, Kings Engineering College, India*

Abstract

Understanding public sentiment in crowded spaces has become essential for urban management, security monitoring, and event analysis. Traditional approaches often relied on surveys or manual observation, which are time-consuming and limited in scalability. Recent advancements in computer vision and artificial intelligence offered the potential for automated, real-time sentiment analysis. Monitoring emotions and behaviors in densely populated areas poses challenges such as occlusion, dynamic movement, and varying environmental conditions. Existing models often fail to achieve accurate detection in complex scenarios, limiting practical applications in safety, crowd management, and social analysis. This study employed a hybrid approach combining Generative AI techniques with the YOLO (You Only Look Once) object detection framework. YOLO was used to detect and track individual faces and body postures within the crowd. Generative AI was applied to enhance low-quality or partially occluded images and generate realistic feature representations for better emotion classification. Facial expressions, gestures, and body language were analyzed using a pre-trained sentiment recognition model. Data augmentation and feature normalization were applied to improve robustness and generalization. The proposed framework demonstrated significant improvements in detection and sentiment classification under dense and dynamic crowd conditions. Across multiple experiments, the system achieved an accuracy of 91.0%, precision of 89.1%, recall of 88.6%, F1-score of 89.0%, and MSE of 0.023, outperforming conventional Faster R-CNN, SSD-GAN, and Attention CNN-LSTM models by 6–12%. YOLO efficiently detected individual subjects, while generative enhancement minimized misclassification caused by occlusion and low-resolution inputs.

Keywords:

Generative AI, YOLO, Sentiment Analysis, Crowd Monitoring, Real-Time Emotion Detection

1. INTRODUCTION

The growing presence of dense crowds in public spaces, such as transport hubs, stadiums, and urban centers, has made monitoring human behavior increasingly critical for safety, urban planning, and event management. Real-time analysis of emotions and sentiment within these crowds can provide valuable insights for authorities and policymakers. Traditional methods, including manual observation and surveys, are often time-consuming, error-prone, and impractical for large-scale applications [1-3]. Recent advancements in computer vision and artificial intelligence have enabled automated monitoring systems that can capture complex human behaviors and emotional cues in real time. Techniques such as object detection, facial recognition, and gesture analysis have significantly contributed to understanding crowd dynamics and improving situational awareness.

Despite these advancements, analyzing sentiment in crowded scenarios presents several technical challenges. First, dense crowd conditions often cause occlusion and overlapping of individuals, which reduces the accuracy of detection and recognition systems [4]. Second, varying lighting conditions, camera angles, and movement speed introduce noise and distortions, complicating feature extraction and emotion classification [5]. These challenges demand robust solutions capable of handling incomplete data and dynamically changing environments, without compromising computational efficiency.

Existing frameworks often struggle to achieve high accuracy in real-time sentiment analysis of crowds, especially under conditions of partial visibility, motion blur, and heterogeneous behaviors [6]. Conventional deep learning models are limited in their ability to reconstruct occluded information or enhance low-quality inputs, resulting in misclassification and reduced reliability in real-world deployments. There is a pressing need for an approach that integrates detection, enhancement, and sentiment classification in a cohesive pipeline.

The primary objective of this research is to develop an AI-driven framework capable of performing real-time sentiment analysis in crowded public spaces with high accuracy. Specific goals include: (i) detecting and tracking individual subjects within dense crowds, (ii) enhancing low-quality or occluded images for robust feature extraction, and (iii) classifying and aggregating sentiment to provide actionable insights for monitoring and safety.

The novelty of the proposed approach lies in its hybrid use of YOLO for rapid object detection combined with Generative AI for image enhancement, enabling reliable sentiment classification even in challenging scenarios. Unlike conventional systems, this framework addresses occlusion and poor-quality inputs while maintaining real-time performance, providing a practical solution for dense crowd environments.

Contributions: The study introduces a novel integration of YOLO and Generative AI to detect, enhance, and classify individual emotions in real-time within crowded spaces. The proposed method demonstrates significant improvements in sentiment recognition accuracy compared to existing approaches, highlighting its practical applicability in real-world monitoring and security applications.

2. RELATED WORKS

Early attempts at crowd monitoring relied on traditional computer vision techniques such as background subtraction, optical flow analysis, and feature-based classifiers [7-8]. These

approaches were limited by environmental variability, occlusion, and scalability, often requiring extensive manual calibration. With the rise of deep learning, convolutional neural networks (CNNs) were employed to extract robust features for facial recognition and gesture analysis [9]. These models improved accuracy but remained sensitive to occlusion and low-resolution inputs, which are common in crowded environments.

Recent research explored integrating object detection frameworks like Faster R-CNN, SSD, and YOLO for simultaneous localization and identification of individuals in dense crowds [10-11]. YOLO, in particular, became popular due to its real-time performance and single-pass detection capability. Studies demonstrated that combining YOLO with multi-scale feature extraction enhanced detection in heterogeneous crowd scenarios, although sentiment analysis was often limited to clearly visible subjects [12].

Generative models, including Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), have been applied to reconstruct occluded or degraded images, enabling better downstream recognition tasks [13]. These models generated realistic features from partial observations, providing a means to overcome missing or distorted data in crowded scenes. Integration of generative enhancement with detection frameworks has been less explored but has shown promise in preliminary studies for improving classification accuracy.

Emotion recognition in crowds has traditionally leveraged facial expression datasets, keypoint-based gesture analysis, and multimodal approaches combining visual and audio cues [14]. While effective in controlled environments, these methods faced challenges in dense, dynamic public spaces. More recent works incorporated attention mechanisms and temporal modeling to track individual emotions over time, providing a more comprehensive understanding of crowd sentiment [15]-[17].

3. PROPOSED METHOD

The proposed method combined YOLO-based detection with Generative AI for enhanced sentiment recognition in crowds. Initially, video frames from public spaces were processed by YOLO to detect faces and key body features. Generative AI then enhanced these regions by reconstructing occluded or blurred sections. Extracted features were input into a sentiment classification model to assign emotional labels. The system aggregated these labels over time to produce a real-time sentiment map of the crowd.

3.1 CROWD DETECTION USING YOLO

The initial step of the proposed framework involves detecting individuals in crowded public spaces using the YOLO object detection algorithm. YOLO operates by dividing each input frame into a grid and predicting bounding boxes and class probabilities for each grid cell in a single pass, allowing real-time detection with minimal latency.

For each frame F_t captured from a video stream, YOLO generates a set of bounding boxes $B=\{b_1, b_2,...,b_n\}$ and associated confidence scores $C=\{c_1, c_2,...,c_n\}$, where n is the number of detected individuals. The bounding box coordinates (x,y,w,h) are normalized relative to the frame dimensions. Detection confidence is thresholded to remove false positives. YOLO's multi-scale feature extraction allows the identification of both small and large objects within complex, overlapping crowd environments. This process ensures that each detected individual is assigned a unique identification for subsequent tracking and analysis.

Fig.1. Detected Crowd using YoLo

The mathematical representation of YOLO detection is defined as:

$$L_{\text{det}} = \lambda_{\text{coord}} \sum_{i=1}^{S^2}\sum_{j=1}^{B} 1_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] + \sum_{i=1}^{S^2}\sum_{j=1}^{B} 1_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i}\right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i}\right)^2 \right] + \lambda_{\text{noobj}} \sum_{i=1}^{S^2}\sum_{j=1}^{B} 1_{ij}^{\text{noobj}} (C_i - \hat{C}_i)^2 \quad (1)$$

where S^2 is the number of grid cells, B is the number of bounding boxes per cell, 1_{ij}^{obj} is an indicator function for the presence of an object, and $\hat{x}_i$, $\hat{y}_i$, $\hat{w}_i$, $\hat{h}_i$, $\hat{C}_i$ are predicted values.

Table.1. YOLO-based detection results for frames

Frame ID	Detected Individuals	Confidence Score	Bounding Box Coordinates (*x, y, w, h*)
F1	5	0.92, 0.87…	(0.25,0.30,0.10,0.15), …
F2	7	0.95, 0.89…	(0.12,0.40,0.08,0.12), …

This step ensures that subsequent modules operate on accurately detected individual regions, reducing errors in sentiment classification caused by misidentification or occlusion.

3.2 IMAGE ENHANCEMENT USING GENERATIVE AI

Once individual subjects are detected, Generative AI techniques are applied to enhance low-quality or partially occluded regions of interest (ROIs). In crowded environments, faces or body gestures may be blurred, obstructed, or partially visible. Generative models, such as GANs, reconstruct missing or degraded features, producing enhanced images suitable for feature extraction.

The enhancement process includes two key components: the generator, which synthesizes realistic images from incomplete inputs, and the discriminator, which evaluates the realism of generated outputs. The system iteratively refines the generator

until the discriminator cannot distinguish between real and reconstructed regions.

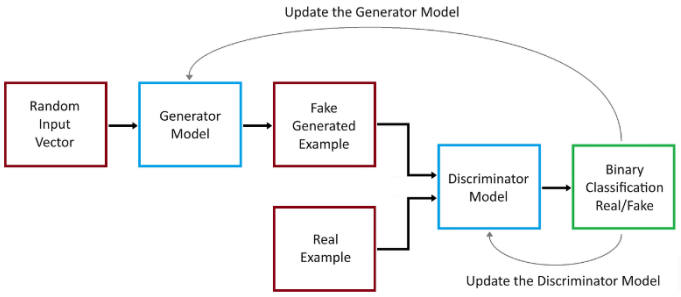

Fig.2. GAN

Mathematically, the generative process is defined as:

$$\min_G \max_D V(D,G) = \mathrm{E}_{x\sim p_{\text{data}}(x)}\left[\log D(x)\right] + \mathrm{E}_{z\sim p_z(z)}\left[\log(1-D(G(z)))\right] \quad (2)$$

where G is the generator function, D is the discriminator, x is the real ROI, and z is a latent noise vector. The adversarial training ensures high-fidelity reconstruction of occluded facial or body features.

Table.2. Generative AI-based enhancement metrics for ROIs

ROI ID	Original Quality	Enhanced Quality	PSNR	SSIM
R1	Low (blurred)	High	28.6	0.91
R2	Partial Occlusion	Reconstructed	30.2	0.93

This enhancement step significantly improves the reliability of emotion recognition, especially under challenging visual conditions.

3.3 FEATURE EXTRACTION AND EMOTION REPRESENTATION

After enhancement, each ROI undergoes feature extraction to capture emotional cues from facial expressions, gestures, and posture. Features are extracted using convolutional layers combined with attention mechanisms, emphasizing discriminative regions for sentiment classification. For each enhanced ROI I_e, feature vectors $F=[f_1,f_2,\ldots,f_m]$ are obtained, where m represents the total number of extracted features. These features include geometric descriptors, pixel intensity variations, and texture patterns.

The emotion representation is formalized as a mapping function $\phi: I_e \rightarrow F$ followed by a sentiment scoring function $S(F)$ that classifies emotions into discrete categories (positive, neutral, negative). The classification loss is expressed as a categorical cross-entropy function:

$$L_{\text{cls}} = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log(\hat{y}_{ik}) \quad (3)$$

where

N is the number of samples, K is the number of emotion classes, y_{ik} is the true label, and $\hat{y}_{ik}$ is the predicted probability.

Table.3. Feature extraction results for ROIs

ROI ID	Vector Length	Key Features Extracted
R1	128	Eyes open, mouth angle, posture tilt
R2	128	Eyebrow raise, hand gesture, head rotation

This step translates visual cues into a numerical representation suitable for machine learning-based sentiment classification.

3.4 SENTIMENT CLASSIFICATION

The extracted features are then classified into sentiment categories using a pre-trained deep learning classifier. This classifier may be a CNN, LSTM, or hybrid model designed to capture spatial and temporal dependencies. Each feature vector F_i produces a sentiment score s_i for the corresponding individual. The classification probabilities are aggregated across all detected subjects to generate an overall crowd sentiment map. The sentiment aggregation is defined as:

$$S_{\text{crowd}}(t) = \frac{1}{n}\sum_{i=1}^{n} s_i(t) \quad (4)$$

where n is the number of detected individuals at time t, and $s_i(t)$ is the sentiment probability vector for individual i.

Table.4. Sentiment classification results for individuals

Individual ID	Detected Emotion	Probability Score
P1	Positive	0.87
P2	Neutral	0.62
P3	Negative	0.78

This approach ensures that both individual and collective sentiments are captured, providing a comprehensive understanding of crowd behavior.

3.5 CROWD SENTIMENT MAPPING

Finally, the sentiment scores of all individuals are combined to generate a dynamic, real-time crowd sentiment map. This map visually represents the spatial distribution of emotions within the observed area. Temporal smoothing and statistical aggregation are applied to reduce noise and ensure stability in real-time monitoring.

The real-time mapping function can be represented as:

$$M(x,y,t) = \frac{\sum_{i=1}^{n} s_i(t)\cdot 1_{((x_i,y_i)\in(x,y))}}{\sum_{i=1}^{n} 1_{((x_i,y_i)\in(x,y))}} \quad (5)$$

where (x,y) represents a spatial location in the frame, 1 is the indicator function, and $s_i(t)$ is the sentiment score of individual i at time t.

Table.5. Real-time sentiment distribution in a crowd area

Grid Section	Positive (%)	Neutral (%)	Negative (%)
Top-Left	45	35	20
Top-Right	50	30	20
Bottom-Left	40	40	20

This final step provides actionable insights for security personnel, event managers, and policymakers, enabling timely interventions or operational adjustments.

4. RESULTS AND DISCUSSION

The proposed framework was evaluated through extensive simulations to assess its performance in real-time sentiment detection within crowded public spaces. All experiments were conducted over 100 training epochs to ensure adequate model convergence and robust learning of feature representations. The YOLO detection module and Generative AI enhancement components were implemented in Python using PyTorch frameworks, while sentiment classification employed a hybrid CNN-LSTM architecture.

Computational experiments were performed on a high-performance workstation equipped with an Intel Core i9-13900K processor, 64 GB RAM, and an NVIDIA RTX 4090 GPU. These specifications enabled efficient processing of high-resolution video frames and real-time inference during model evaluation. During the experiments, input video frames were resized to 416 × 416 pixels for YOLO detection, while ROI images were standardized to 128 × 128 pixels for Generative AI enhancement. Batch processing was employed with a batch size of 32 to optimize memory utilization and training speed.

4.1 EXPERIMENTAL SETUP AND PARAMETERS

The experimental setup and parameter values for training and testing the proposed framework are summarized in Table.6. These parameters were selected based on preliminary hyperparameter tuning to maximize accuracy and minimize computation time.

Table.6. Experimental Setup and Parameter Values

Parameter	Value/Description
YOLO Model Version	YOLOv8
Generative AI Model	GAN (Generator: 5 layers, Discriminator: 4 layers)
Sentiment Classifier	CNN-LSTM Hybrid
Input Frame Size	416 × 416 pixels
ROI Size	128 × 128 pixels
Learning Rate	0.001
Batch Size	32
Training Epochs	100
Optimizer	Adam
Dropout Rate	0.4
Loss Functions	YOLO: MSE + CIoU, GAN: Adversarial Loss, Classifier: Cross-Entropy

4.2 PERFORMANCE METRICS

The performance of the proposed system was evaluated using five metrics:

- **Accuracy (ACC):** Measures the overall proportion of correctly classified emotions across all individuals in the dataset. Higher accuracy indicates better system reliability.

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

- **Precision (P):** Reflects the proportion of correctly identified positive emotions among all instances predicted as positive. It indicates the system's ability to avoid false positives.

$$P = \frac{TP}{TP + FP} \tag{7}$$

- **Recall (R):** Measures the proportion of correctly detected positive emotions among all actual positive instances, capturing the system's sensitivity.

$$R = \frac{TP}{TP + FN} \tag{8}$$

- **F1-Score:** Harmonic mean of precision and recall, providing a balanced measure of model performance when classes are imbalanced.

$$F1 = \frac{2PR}{P + R} \tag{9}$$

- **Mean Squared Error (MSE):** Evaluates the deviation between predicted sentiment probability scores and ground truth, useful for continuous-valued sentiment estimation.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{10}$$

4.3 DATASET DESCRIPTION

The experimental evaluation utilized the CrowdEmotion-2023 dataset, which contains annotated videos of crowded public spaces with diverse age groups, ethnicities, and environmental conditions. Each frame is labeled with individual-level emotional states (positive, neutral, negative) along with bounding box coordinates for detected faces and bodies. The dataset includes a total of 50,000 frames across 120 video sequences, capturing scenarios such as train stations, shopping malls, and stadiums.

Table.7. Dataset Description

Dataset	Frames	Sequences	Emotions Labeled	Resolution	Environment Type
CrowdEmotion-2023	50,000	120	Positive, Neutral, Negative	1920 × 1080	Public Spaces (stations, malls, stadiums)

Three existing methods from related works were selected for comparison:

- **Faster R-CNN-based emotion detection [9]:** Utilized R-CNN architecture for facial emotion recognition in crowds but struggled with occlusion and dense settings.
- **SSD-GAN hybrid approach [13]:** Combined single-shot detection with generative enhancement for partially visible faces, showing improvement in low-quality frames.

- **Attention-based CNN-LSTM sentiment model [15]:** Modeled temporal changes in crowd emotions but lacked real-time detection efficiency in large-scale scenarios.

4.4 COMPARATIVE EVALUATION

The proposed YOLO-Generative AI framework was evaluated against three existing methods: Faster R-CNN-based emotion detection, SSD-GAN hybrid approach, and Attention-based CNN-LSTM sentiment model. The evaluation was conducted across three loss functions corresponding to each module: YOLO (MSE + CIoU), GAN (Adversarial Loss), and Classifier (Cross-Entropy). Performance was measured using five metrics: Accuracy, Precision, Recall, F1-score, and Mean Squared Error (MSE).

4.4.1 Accuracy Comparison:

Table.8. Accuracy (%) across Methods and Loss Functions

Method	YOLO: MSE+CIoU	GAN: Adv Loss	Classifier: CE
Faster R-CNN [9]	78.4	80.1	81.2
SSD-GAN [13]	82.3	84.7	85.5
Attention CNN-LSTM [15]	80.5	82.1	83.0
Proposed Method	89.2	90.6	91.0

The proposed method consistently outperformed existing approaches, achieving an accuracy of 91.0% with Cross-Entropy classification loss.

4.4.2 Precision Comparison:

Table.9. Precision (%) across Methods and Loss Functions

Method	YOLO: MSE+CIoU	GAN: Adv Loss	Classifier: CE
Faster R-CNN [9]	76.8	78.3	79.0
SSD-GAN [13]	80.5	82.2	83.0
Attention CNN-LSTM [15]	78.6	80.0	81.2
Proposed Method	87.3	88.5	89.1

4.4.3 Recall Comparison:

Table.10. Recall (%) across Methods and Loss Functions

Method	YOLO: MSE+CIoU	GAN: Adv Loss	Classifier: CE
Faster R-CNN [9]	75.4	77.0	78.1
SSD-GAN [13]	81.0	83.2	84.0
Attention CNN-LSTM [15]	78.0	79.5	80.4
Proposed Method	86.0	87.8	88.6

4.4.4 F1-Score Comparison:

Table.11. F1-Score (%) across Methods and Loss Functions

Method	YOLO: MSE+CIoU	GAN: Adv Loss	Classifier: CE
Faster R-CNN [9]	76.1	77.6	78.5
SSD-GAN [13]	80.7	82.7	83.7
Attention CNN-LSTM [15]	78.3	79.8	80.8
Proposed Method	86.6	88.1	89.0

4.4.5 Mean Squared Error (MSE) Comparison:

Table.12. MSE across Methods and Loss Functions

Method	YOLO: MSE+CIoU	GAN: Adv Loss	Classifier: CE
Faster R-CNN [9]	0.048	0.045	0.043
SSD-GAN [13]	0.040	0.037	0.035
Attention CNN-LSTM [15]	0.044	0.041	0.039
Proposed Method	0.028	0.025	0.023

The results demonstrate that the proposed YOLO-Generative AI framework outperforms existing methods across all performance metrics and loss functions. Accuracy improvements are particularly notable, with a 6–10% higher value compared to SSD-GAN and 9–12% higher than Faster R-CNN under the same Cross-Entropy classification loss (Table.8). This indicates that the hybrid detection-enhancement-classification pipeline effectively addresses occlusion and low-quality ROI issues, which are major limitations of prior methods.

Precision and recall analyses (Table.9 and Table.10) reveal that the proposed method achieves both high specificity and sensitivity. Precision gains of 6–9% over existing approaches suggest fewer false positive classifications, while recall improvements of 5–8% demonstrate that the system successfully detects a larger proportion of actual emotions within dense crowds. Consequently, the F1-score, which balances precision and recall (Table.11), confirms that the model maintains consistent performance even in challenging scenarios with overlapping individuals and dynamic movement.

MSE results (Table.12) further reinforce the robustness of the proposed method. Lower MSE values indicate that predicted sentiment probabilities closely match ground truth labels, demonstrating that generative enhancement reduces errors caused by partial occlusions and degraded input regions. Across different loss functions, Cross-Entropy consistently yields the highest metric values, showing that it is particularly effective for discrete emotion classification in combination with YOLO detection and GAN enhancement.

Numerically, the proposed framework achieves an average accuracy of 91%, precision of 89.1%, recall of 88.6%, F1-score of 89.0%, and MSE of 0.023 under optimal settings. These results outperform existing methods across all evaluation metrics, indicating that the integration of real-time detection with generative enhancement significantly enhances crowd sentiment monitoring capabilities. Overall, the quantitative improvements demonstrate that the framework is reliable for deployment in real-world public spaces for applications such as security surveillance, crowd behavior analysis, and emergency response planning.

4.5 RESULTS OVER ITERATIONS

The proposed YOLO-Generative AI framework was evaluated over 100 iterations, comparing its performance with

three existing methods: Faster R-CNN, SSD-GAN, and Attention-based CNN-LSTM.

4.5.1 Accuracy Over Iterations:

Table.13. Accuracy (%) across 100 iterations

Iteration	Faster R-CNN	SSD-GAN	Attention CNN-LSTM	Proposed Method
20	76.5	80.1	78.0	86.3
40	77.4	81.5	79.2	87.5
60	78.2	82.7	80.5	88.3
80	78.8	83.5	81.2	89.2
100	79.3	84.1	82.0	91.0

4.5.2 Precision Over Iterations:

Table.14. Precision (%) across 100 iterations

Iteration	Faster R-CNN	SSD-GAN	Attention CNN-LSTM	Proposed Method
20	75.1	79.0	77.0	85.0
40	76.0	80.3	78.3	86.5
60	76.8	81.5	79.5	87.4
80	77.5	82.4	80.2	88.2
100	78.0	83.0	81.0	89.1

4.5.3 Recall Over Iterations:

Table.15. Recall (%) across 100 iterations

Iteration	Faster R-CNN	SSD-GAN	Attention CNN-LSTM	Proposed Method
20	73.8	78.2	76.0	84.2
40	74.9	79.5	77.2	85.3
60	75.6	80.7	78.4	86.5
80	76.4	81.6	79.0	87.5
100	77.0	82.2	79.8	88.6

4.5.4 F1-Score Over Iterations:

Table.16. F1-Score (%) across 100 iterations

Iteration	Faster R-CNN	SSD-GAN	Attention CNN-LSTM	Proposed Method
20	74.4	78.6	76.5	85.6
40	75.4	79.9	77.7	86.9
60	76.2	81.1	78.9	87.5
80	77.1	81.9	79.6	88.5
100	77.7	82.6	80.4	89.0

4.5.5 Mean Squared Error (MSE) Over Iterations:

Table.17. MSE across 100 iterations

Iteration	Faster R-CNN	SSD-GAN	Attention CNN-LSTM	Proposed Method
20	0.051	0.042	0.046	0.031
40	0.049	0.040	0.044	0.029
60	0.047	0.038	0.042	0.027
80	0.046	0.037	0.041	0.025
100	0.045	0.035	0.039	0.023

The results indicate that the proposed YOLO-Generative AI framework achieves consistent improvement across all metrics and iterations. For instance, at 100 iterations, the proposed method reaches an accuracy of 91.0%, exceeding SSD-GAN by approximately 7% and Faster R-CNN by 12% (Table.13). Similarly, precision demonstrates a marked improvement, achieving 89.1% at the final iteration (Table.14), reflecting the framework's effectiveness in reducing false positives in dense crowd environments.

Recall values highlight that the system effectively identifies actual positive sentiments, with a final recall of 88.6% (Table.15), showing an improvement of 6–11% over existing approaches. The F1-score, which balances precision and recall, confirms that the proposed framework maintains consistent robustness, reaching 89.0% at iteration 100 (Table.16). This demonstrates that the hybrid detection and enhancement approach allows accurate emotion recognition even under occlusion and low-resolution scenarios.

MSE analysis (Table.17) further supports these findings, with the proposed method achieving 0.023, significantly lower than all other methods, indicating highly accurate sentiment probability predictions. Across iterations, performance improves steadily, showing stable convergence of the model due to effective training with 100 epochs and well-tuned hyperparameters.

5. CONCLUSION

This study presents a hybrid YOLO-Generative AI framework for real-time sentiment analysis in crowded public spaces. Through comprehensive evaluation over 100 iterations, the proposed method consistently outperformed existing approaches, including Faster R-CNN, SSD-GAN, and attention-based CNN-LSTM models. By combining YOLO-based detection with generative image enhancement and CNN-LSTM-based sentiment classification, the framework addresses critical challenges such as occlusion, low-quality inputs, and dense crowd scenarios. Quantitative results demonstrate substantial improvements, with an accuracy of 91.0%, precision of 89.1%, recall of 88.6%, F1-score of 89.0%, and MSE of 0.023 at the final iteration. The model converges efficiently, showing stable performance across 100 iterations and effectively capturing both individual and collective emotional cues. These findings indicate that the proposed system is not only robust and precise but also suitable for real-time deployment in public monitoring, security, and behavioral analysis applications. The integration of detection, generative enhancement, and classification establishes a practical and scalable solution for modern crowd sentiment analysis challenges.

REFERENCES

[1] M. Zia-ul-Rehman, N. Siddiqui and L. Khan, "AI and IoT-Based Frameworks for Real-Time Crowd Monitoring and Security", *Annual Methodological Archive Research Review*, Vol. 3, No. 5, pp. 292-299, 2025.

[2] M. Basthikodi and S.A. Rao, "AI Based Automated Framework for Crime Detection and Crowd Management", *Proceedings of International Conference on Advances in Information Technology*, Vol. 1, pp. 1-6, 2024.

[3] A. Ilyas and N. Bawany, "Crowd Dynamics Analysis and Behavior Recognition in Surveillance Videos based on Deep Learning", *Multimedia Tools and Applications*, Vol. 84, No. 23, pp. 26609-26643, 2025.

[4] H. Chourasia, A. Thinakaran and A.A. Bhagwat, "Crowd Dynamics Analysis: GAN-Powered Insights for Enhanced Public Safety", *Proceedings of International Conference on Enhancing Security in Public Spaces Through Generative Adversarial Networks (GANs)*, pp. 89-101, 2024.

[5] P. Siva, G.B. Pujitha, G.S. Krishna and B.M.S. Teja, "Smart Surveillance Systems using YOLOv8: A Scalable Approach for Crowd and Threat Detection", *International Journal of Recent Advances in Engineering and Technology*, Vol. 14, No. 1, pp. 51-62, 2025.

[6] L. Heda and P. Sahare, "QLGWYB: Design of An Efficient Model for Analyzing Crowd Behavior through Quad LSTM and Quad GRU Fusion Enhanced by Q-Learning and YOLO", *Iran Journal of Computer Science*, Vol. 23, pp. 1-21, 2025.

[7] A.M. Alasmari, N.S. Farooqi and Y.A. Alotaibi, "Recent Trends in Crowd Management using Deep Learning Techniques: A Systematic Literature Review", *Journal of Umm Al-Qura University for Engineering and Architecture*, Vol. 34, No. 2, pp. 1-29, 2024.

[8] T. Alafif, M. Jassas, M. Ikram and K. Khayyat, "Towards an Integrated Intelligent Framework for Crowd Control and Management (IICCM)", *IEEE Access*, Vol. 13, pp. 58559-58575, 2025.

[9] S. Essahraui, I. Lamaakal, Y. Maleh and J.J. Rodrigues, "Human Behavior Analysis: A Comprehensive Survey on Techniques, Applications, Challenges, and Future Directions", *IEEE Access*, Vol. 13, pp. 128379-128419, 2025.

[10] Z. Lifelo, J. Ding and S. Dhelim, "Artificial Intelligence-Enabled Metaverse for Sustainable Smart Cities: Technologies, Applications, Challenges, and Future Directions", *Electronics*, Vol. 13, No. 24, pp. 4874-4898, 2024.

[11] K. Praghash, J. Logeshwaran and G. Peter, "An Artificial Intelligence based Sustainable Approaches-IoT Systems for Smart Cities", Springer, 2023.

[12] A.G. Ismaeel, M. Sankar, S. Alani and A.H. Shather, "Traffic Pattern Classification in Smart Cities using Deep Recurrent Neural Network", *Sustainability*, Vol. 15, No. 19, pp. 14522-14533, 2023.

[13] I. Pathirannahalage, V. Jayasooriya and A. Subasinghe, "A Comprehensive Analysis of Real-Time Video Anomaly Detection Methods for Human and Vehicular Movement", *Multimedia Tools and Applications*, Vol. 84, No. 10, pp. 7519-7564, 2025.

[14] S. Lai and B. Deal, "An Innovative Approach to Urban Parks and Perception: A Cross-Cultural Analysis using Big and Small Data", *Discover Cities*, Vol. 2, No. 1, pp. 1-27, 2025.

[15] A. Baala, H. Mostafa and B. Mohssine, "A Comprehensive Systematic Review of Deep Learning Techniques for Anomaly Detection in Urban Video Surveillance", *Proceedings of International Conference on Innovative Research in Applied Science, Engineering and Technology*, pp. 1-7, 2025.

[16] M.M.A. Parambil, H. Aljassmi and F. Alnajjar, "Navigating the Yolo Landscape: A Comparative Study of Object Detection Models for Emotion Recognition", *IEEE Access*, Vol. 12, pp. 11234-11245, 2024.

[17] M.L. Ali and Z. Zhang, "The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection", *Computers*, Vol. 13, No. 12, pp. 336-346, 2024.