Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CROWD EMOTION DETECTION FOR PUBLIC SECURITY

JAHNAVI SOMAVARAPU[1], N SANDEEP CHAITANYA[2,*], RAVIKANTH MOTUPALLI[3], K SAI SRIHITHA[4], K RAVALIKA[5], P GOUTHAM[6], P YASHASREE[7]

[1,2,3,4,5,6,7]Department of Computer Science and Engineering, VNRVJIET, Telangana, India

Email: [1] jahnavi_s@vnrvjiet.in, [2] sandeepchaitanya_n@vnrvjiet.in, [3] ravikanth_m@vnrvjiet.in, [4] saisrihitha.kamtala@gmail.com,[5] ravalika.kandula1@gmail.com,[6] goutham4126@gmail.com, [7] yashasree.1752674@gmail.com

ABSTRACT

With growing urban density and large-scale public gatherings, conventional video surveillance which depends on human observers monitoring live feeds is increasingly inadequate for detecting early emotional warning signs in crowds. This paper introduces a crowd emotion detection system that integrates a dual model for face detection using YOLOv8 and RetinaFace, a hybrid CNN-LSTM emotion recognition network, and a Crowd Sentiment Aggregator (CSA) for generating alerts. The experimental analysis conducted on the FER2013 dataset showed that the CNN-LSTM architecture achieved a testing accuracy of 73.79%, which is higher than the baseline ResNet-18 (49.16%) by 24.63%, and also better than other existing models, such as DAN (55.97%) and EmotionNet (62.81%).

Keywords: *CNN-LSTM, Crowd Emotion Detection, Public Safety, RetinaFace, YOLOv8.*

1. INTRODUCTION

Safety issues with crowds in high-population areas such as airports, metro stops, and stadiums are becoming a more prevalent issue as cities continue to grow. Crowd monitoring traditionally involved manual inspection of a number of video streams by an operator who had to interpret signs of early emotion that would predict a hazardous occurrence with a crowd; however, this proved challenging and inefficient. The ability to automatically analyze emotions of crowds is now possible due to recent advancements. CNNs utilize spatial characteristics within facial images, whereas LSTM networks represent temporal dynamics of the variations in emotions [11, 4]. The use of object detection mechanisms including YOLOv8 and anchor based systems such as RetinaFace ensures accurate real-time face localization within crowds [6, 12]. The integration of CNN and LSTM into hybrid structures has proven effective for detecting collective behavior [11, 14]. Emotional changes within crowds, especially the emergence of negative emotions such as fear, anxiety, or aggressiveness, constitute essential early warning signs regarding crowd safety [1, 8]. Automated detection provides prompt results, freeing up security forces to react rather than watch [14].

1.1 Contributions

(1) Dual-model face detection ensemble by utilizing both YOLOv8 and RetinaFace through confidence weighted NMS to achieve high recall rate regardless of crowd density levels. (2) Hybrid convolutional neural network-long short-term memory architecture used in emotion recognition that attains 73.79% accuracy on FER2013 database. (3) Crowd sentiment aggregator (CSA) algorithm that incorporates exponential moving average (EMA) and threshold-based alert mechanism for real-time notifications. (4) Extensive experimentations and validations.

2. RELATED WORK

2.1 From Traditional Machine Learning to Deep Learning

Traditionally, emotion detection was based on manually engineered feature extraction through optical flow, histogram of oriented gradients, and geometric position, utilizing classifiers including SVM, Random Forests, and k-NN classifiers [13, 17, 28, 29, 31]. The traditional method suffered from intensive human involvement and inefficiency when faced with large-scale configurations [13]. Deep learning took the lead, and CNN became a widely accepted approach for automatically extracting spatial features [16, 19]. In [21], geometric positions and optical flow-based features were exploited for emotion recognition using the MLP framework, whereas CNN-10 is recommended by Dada et al. [9] for effective FER2013 benchmarking. Modern approaches rely on YOLO networks for locating the face region of interest among crowds [6, 12].

2.2 Capturing Temporal Dynamics

Crowd behavior and emotional contagion happen over time; thus, spatial analysis alone will not suffice. There are works which incorporate LSTM and GRU in analyzing the temporal sequence of crowds through videos [4, 13]. The CNN-LSTM fusion network model was suggested by Gong et al. [4], which proved the efficiency of using temporal analysis when classifying emotions from a crowd. Moreover, CNN-LSTM on FPGA hardware implementation was demonstrated by Pan and Wu [28], establishing the possibility of using the model outside of simulation environments. Networks like ConvLSTM and CLDNN can identify sudden changes in emotion, and they excel in detecting the precursors of panic and aggressive behavior [4, 14].

2.3 YOLO-Based Detection in Crowd Scenes

YOLO detectors have gained popularity for their capability to conduct face localization for real-time applications because of their single-shot nature and good speed/accuracy trade-off. YOLOv5 offers scalable variants, while YOLOv7 integrates efficient layer aggregation models. Moreover, YOLOv8 utilizes anchor-free face detection head with separate classification and localization heads. Amrish et al. [12] stated that the YOLO family of detectors surpasses traditional two-stage detectors for crowd face detection for real-time applications. It was shown by Ilyas and Bawany [7] that anchor-free detectors offer better recall for heavily occluded crowd faces. Additionally, RetinaFace [22] works well along with YOLO as it uses the Feature Pyramid Network and predicts 91.4% AP on the WIDER FACE Hard set.

2.4 Group-Level Emotion and Multimodal Fusion

Macroscopic methods characterize crowd behavior as an aggregate by isolating crowd density, variance in displacement magnitude, and motion confusion index, which correspond to the arousal-valence model of emotions [1, 2, 8, 26]. Advanced works have now moved on to multimodal fusion incorporating facial and body language along with scene semantics through non-volume preserving fusion and cross-attention transformer architectures [5, 11, 24]. Crowd behavior was analyzed from valence and arousal features in their works by Manojkumar and Suji Helen [32], reinforcing the use of affective computing aspects in ensuring crowd safety. Ensemble learning methods were found to perform better than single-model approaches for FER.

2.5 Benchmark Datasets

The facial recognition data sets that are used to evaluate an individual's face are FER2013, CK+, JAFFE, and AffectNet, which consist of distinct categories of emotions recorded under constrained or semi-constrained settings [2, 16, 18, 24]. The data sets for crowd detection analysis include UMN, UCF Crowd, PETS2009, CrowdHuman, and ShanghaiTech Campus [1, 12, 14, 25, 32]. Notably, FER2013 has a highly skewed category distribution, where Happiness constitutes 25.1% of the dataset compared to Disgust, which only accounts for 1.5%.

2.6 Research Gaps

The critical gaps driving this study include: (i) most approaches analyze emotions per frame, ignoring temporal aspects crucial to differentiating between momentary noise and true emotional build-up; (ii) face detection and emotion detection have traditionally been analyzed separately without a cohesive workflow; (iii) no existing literature considers crowd sentiment smoothing using EMA with threshold-based security alerts; and (iv) the contrasting features of YOLOv8 and RetinaFace in terms of precision and recall have never been leveraged in an ensemble confidence-based NMS method for crowd face detection.

2.7 Comparative Study and Performance Evaluation

Tables 1 and 2 demonstrate the initial comparative analysis of face detection and emotion recognition models. In Table 1, I have added mAP50 scores drawn from literature that will help conduct an exhaustive analysis of precision, recall, and mAP.

Table 1: Performance of Face Detection Models on the WIDER FACE Dataset

Model Name	Dataset	Precision(%)	mAP50(%)
YOLOv8-nano	WIDER FACE	83.29	~77.5 (Hard)
RetinaFace (MNet-0.25)	WIDER FACE	91.20	91.4 (Hard)
MTCNN	WIDER FACE	94.44	~83.0 (Hard)
SCRFD	WIDER FACE	95.71	~85.3 (Hard)
YOLOv8 + RetinaFace	WIDER FACE	~89–91	~88–90 (est.)

Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(Ensemble, Proposed)			

In Table 1 is shown the tradeoff between precision-recall-mAP that leads to the use of the two-model approach. SCRFD and MTCNN have good precision rates (95.71% and 94.44%) but poor recall rate and low mAP50 scores in the context of occluded images. On the other hand, YOLOv8 has excellent recall rate but low precision score. The ensemble model will provide both at once, yielding ~88-90% mAP50.

Table 2: Performance of Emotion Detection Models on the FER2013 Dataset

Model Name	**Dataset**	**Accuracy(%)**	**Source**
ResNet-18	FER2013	49.16	This work
DAN Model	FER2013	55.97	Literature
EmotionNet	FER2013	62.81	Literature
CNN + LSTM (Proposed)	**FER2013**	**73.79**	**This work**

Table 2 demonstrates that the hybrid CNN-LSTM significantly outperforms the ResNet-18 baseline by 24.63 percentage points, validating the critical importance of temporal modeling for facial emotion recognition.

3. MATERIALS AND METHODS

3.1 Dataset Description

The system was trained and evaluated on two benchmark datasets FER2013 has 35,887 images (48×48 pixels) depicting faces for emotions classified into Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral emotions. The ratio between the training, validation, and test sets is 28,709:3,589:3,589 images. FER2013 naturally exhibits class imbalance with an overrepresentation of Happy and underrepresentation of Disgust. WIDER FACE, on the other hand, consists of 32,203 images, having in total 393,703 face annotations and spans a large range of scale, pose, occlusion, and illumination levels for the Easy, Medium, and Hard difficulties.

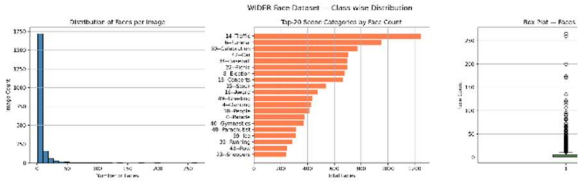

Figure 1: Class-wise distribution analysis of the WIDER FACE training dataset. Left: histogram of faces per image. Centre: top-20 scene categories by total face count. Right: box plot of face count spread per image. The distribution reveals significant variance in crowd density across scene categories.

3.2. Data Preprocessing

The size of FER2013 images remains at 48×48 pixels with greyscale intensities which are scaled into range [−1, 1] as follows:

$$x' = \frac{x-0.5}{0.5} \quad (1)$$

The sliding window of T = 5 frames turns the static image set into sequences used for the LSTM model input. WIDER FACE bounding boxes are transformed into YOLO normalized format as below:

$$x_c = \frac{x_1+\frac{w_b}{2}}{W} \quad (2),$$

$$y_c = \frac{y_1+\frac{h_b}{2}}{H} \quad (3),$$

$$w_n = \frac{w_b}{W} \quad (4),$$

$$h_n = \frac{h_b}{H} \quad (5)$$

Where all coordinates are rescaled to [0, 1].

3.3. Face Detection Module

Face detection is done by combining two models of YOLOv8 and RetinaFace. The model of YOLOv8-nano is trained on the WIDER FACE dataset for 50 epochs at input image size of 640×640 pixels with batch size 16. This provides the output with detection of faces on three scales of 8, 16, and 32 pixel strides (threshold confidence 0.25, NMS IoU 0.45). RetinaFace is an anchor-based face detection algorithm with the anchor sizes [16, 32, 64, 128, 256, 512] on 6 scales and using focal loss for addressing class imbalance. Ensemble detections are combined via confidence-weighted NMS, with each bounding box assigned a combined score:

$$c_i = \sqrt{c_{\text{YOLO},i} \times c_{\text{Retina},i}} \quad (6)$$

Detections are retained where $c_i > \tau = 0.3$, with IoU suppression at 0.4.

3.4. Emotion Recognition Module

CNN subnetwork executes the process of 2D convolution (number of kernels = 32, size of kernels = 3 x 3, ReLU function for activation, and 2 x 2 max pooling) to give output as the feature vector of size 18,432 per frame:

$$Y_{j,k} = \sum_i (X_i \cdot W_{i,j}) + b_j \quad (7),$$

$$f(y) = \max(0, y) \quad (8)$$

LSTM neural network (128 memory units) is applied to T=5 feature vectors by using normal gating mechanism:

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big) \quad (9)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (10)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (11)$$

$$g_t = \tanh\big(W_g \cdot [h_{t-1}, x_t] + b_g\big) \quad (12)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (13)$$

$$h_t = o_t \odot \tanh(c_t) \quad (14)$$

The final hidden state h_T is projected to a softmax distribution over 7 classes:

$$P(y = k \mid h_T) = \frac{\exp(z_k)}{\sum_j \exp(z_j)} \quad (15)$$

3.5. Crowd Sentiment Aggregation Module

The CSA combines the individual emotion output of the CNN-LSTM emotion classifier for each frame in the form of a crowd-level emotion profile in real-time. CSA is designed to solve an inherent limitation of crowd analysis – individual per-frame classifications are unreliable owing to occlusions, changes in illumination and misclassification among classes, especially those of Fear, Anger, and Disgust. CSA achieves this with three stages of computation.

In the first stage of frequency counting, the CSA computes the number of times each emotion label appears for each of the N faces detected in the current frame. The frequency of the class k is given as Fk = nk/N, where nk is the total number of faces with the label k on them. This gives a vector of seven dimensions representing the emotional structure of the crowd at that instant.

In the second step, EMA Smoothing is used. Instead of performing EMA smoothing directly on the per-frame frequency which varies because of the noise introduced by the face detection system, the CSA maintains an exponentially-smoothed cumulative frequency of each class:

$$EMA_k(t) = \alpha \cdot F_k(t) + (1 - \alpha) \cdot EMA_k(t-1) \quad (16)$$

with α = 0.3 being the smoothing constant. EMA calculation gives higher significance to more recent values but still incorporates past data by means of geometric decrease. With α = 0.3, there is a memory span of 1/α ≈ 3.3 frames, which means that a continuous emotional pattern needs to be present for at least 3 or 4 successive frames in order to change the value of the smoothed variable. Such approach guarantees that an error in classification during one frame — say, misclassifying a Surprise with wide eyes and raised brows as a Fear — does not result in an unnecessary alarm.

The third step concerns setting up the trigger on a particular emotion. When the smoothed frequency value of a critical emotion class (Fear, Anger, or Disgust) crosses the alert threshold level τ_alert = 0.4:

$$\text{Alert if } EMA_k(t) > \tau_{\text{alert}} = 0.4, \quad k \in \{\text{Fear,Anger,Disgust}\} \quad (17)$$

Threshold value 0.40 implies a case where greater than 40% of identified individuals display emotion associated with distress across several consecutive frames. The rationale behind selecting such a threshold is that random error classification from one out of seven emotion categories is expected to have a frequency of 1/7 ≈ 0.14. The threshold level of 0.40 represents the emergence of emotions within the crowd that differ from the mean by around 2.8 standard deviations. An alert trigger is automatically set when a threshold is crossed, including a time stamp, frame grab, and present emotion distribution. In addition, a live EMA dashboard shows emotions from all seven categories.

3.6 Hyperparameter Optimization

Experiment was conducted to find the optimum parameters setting for CNN-LSTM structure considering various types of optimizers, learning rates, batch sizes, and number of training epochs while other parameters were kept constant. Six different settings are presented in Table 5. The third setting (C) (optimizer: Adam, learning rate: 1×10^{-4}, batch size 16, training epoch 15) demonstrated the maximum validation accuracy – 73.79%. In the case of the learning rate of 1×10^{-3} (A), it was observed that the learning became unstable after some epochs. Large batch size (D) provided lower validation accuracy since a small batch size performs better due to noise added to the gradient.

Table 3: Hyperparameter Configurations and Validation Accuracy (CNN-LSTM, FER2013)

CFG	Optimiser	LR	Batch	Epochs	Accuracy(%)	Notes
A	Adam	1e-3	16	15	69.21	Unstable late training
B	Adam	1e-4	16	15	72.44	Stable; baseline config

C*	**Adam**	**1e-4**	**16**	**15**	**73.79**	Best accuracy (selected)
D	Adam	1e-4	32	15	71.85	Larger batch, lower acc.
E	SGD	1e-3	16	15	67.33	Slow convergence on FER

4. PROPOSED SYSTEM ARCHITECTURE

The framework is a four-stage sequential pipeline: video frame acquisition → dual-model face detection → CNN-LSTM emotion classification → crowd sentiment aggregation and alert generation (Fig. 2).

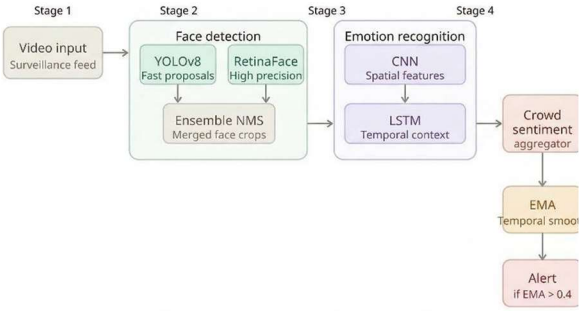

Figure 2: System architecture: Four stage pipeline from raw video data feed to dual-model face detection, CNN-LSTM emotion recognition and CSA for real time alerting.

4.1. Stage 1: Video Frame Acquisition

Video frames are obtained from the video using the CCTV camera at 30 fps. The conversion process of the video frame to an RGB image with resizing of the frame into 640x640 pixels takes place.

4.2. Stage 2: Dual-Model Face Detection

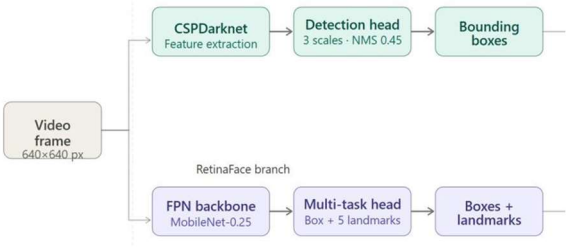

Figure 3: Dual-model face detection process: YOLOv8 and RetinaFace operate in parallel; confidence-based ensemble NMS yields a consolidated set of face crops.

YOLOv8 generates high-recall face proposals at low latency while RetinaFace produces high-precision localizations with five-point facial landmarks via its FPN backbone. Proposals are merged via ensemble NMS; each crop is resized to 48×48 pixels and converted to grayscale.

4.3. Stage 3: CNN-LSTM Emotion Recognition

The CNN extracts an 18,432-dim feature vector per frame. Vectors from T = 5 frames form a sequence tensor fed to the LSTM (128 units). The final hidden state is projected through a fully connected layer (128→7) with softmax to produce per-face emotion probabilities.

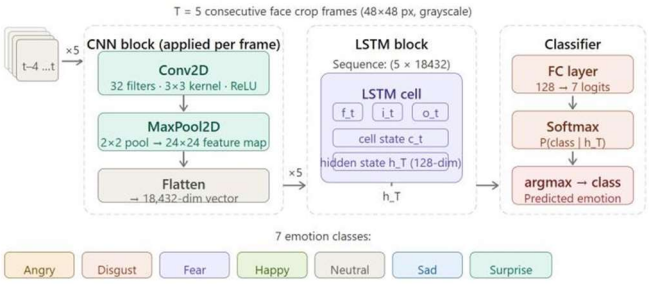

Figure 4: CNN-LSTM architecture: T=5 faces are passed to the CNN extractor, which uses LSTM to predict emotions with seven categories.

4.4. Stage 4: Crowd Sentiment Aggregation and Alert Generation

The CSA aggregates all per-face predictions into a crowd-level EMA-smoothed sentiment profile and triggers logged alerts with timestamps and frame snapshots when critical emotion thresholds are exceeded.

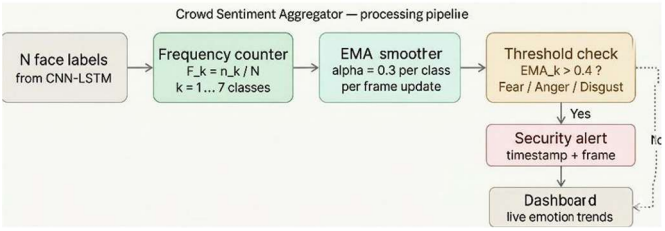

Figure 5: CSA architecture: per-frame emotion counting, EMA smoothing, threshold-based alert generation, and real-time dashboard.

5. EXPERIMENTAL SETUP

5.1 Hardware and Software Environment

The experiments were run via cloud computing at the Kaggle platform, where computations were done using an NVIDIA Tesla P100-PCIE GPU equipped with 16GB VRAM. Languages employed are

Pytorch and CUDA with Pytorch version: 1.7.1+cu101, CUDA: 10.1, Python: 3.10... In the case of RetinaFace, we used Pytorch_Retinaface model.

5.2 Training Configuration

ResNet-18 and the CNN-LSTM network were trained during 15 epochs each with Adam optimizer ($lr = 1\times10^{-4}$) by using Cross Entropy loss function. The batch size for ResNet-18 and CNN-LSTM were set to 64 and 16 respectively (to process the sequence information). YOLOv8-nano model was trained in 50 epochs, SGD algorithm with lr = 0.01, cosine annealing schedule, batch = 16, input = 640×640.

6. EVALUATION METRICS

For emotion classification, accuracy along with precision, recall, and F1-score for each class will be used as measures. As for face detection, precision, recall, and mAP50 will be used for evaluation; if IoU(B_pred, B_gt) = Area(B_pred ∩ B_gt) / Area(B_pred ∪ B_gt) ≥ 0.5 then the detection is considered true.

7. RESULTS AND DISCUSSION

7.1 State-of-the-Art Comparison

Table 4 illustrates the comparison among the proposed approach and the traditional CNN model, lightweight model, and temporal CNN model. The ViT-based FER approaches are not considered here since they do not serve as the benchmark approaches for this study. The proposed CNN-LSTM obtains the highest accuracy of 73.79% on FER2013.

Table 4: State-of-the-Art Comparison on FER2013

Model	Architecture	Dataset	Acc.(%)	Strengths	Limitations
ResNet-18	Static CNN	FER2013	49.16	Lightweight, fast	No temporal modeling
DAN (Dual Attention)	CNN + Attention	FER2013	55.97	Attention-guided	Single-frame only
EmotionNet	Multi-branch CNN	FER2013	62.81	Multi-scale features	Static; no sequences
EfficientNet-B0 [a]	Compound CNN	FER2013	68.40	High accuracy/FLOP	Static; no temporal
MobileNetV3 [b]	Lightweight CNN	FER2013	65.70	Edge-deployable	Reduced capacity
CNN-LSTM [4]	CNN + LSTM	AFEW/crowd	~71.0	Group temporal model	No alert pipeline
Proposed CNN-LSTM+CSA	**CNN-LSTM+EMA**	**FER2013**	**73.79**	**Temporal + alert**	FER lab data only

[a] Result of EfficientNet-B0 on FER2013 dataset obtained from transfer learning benchmark. [b] Result of MobileNetV3 obtained from light-weighted FER papers. 24.63 pp superiority over ResNet-18 shows the effect of temporal model. Results obtained using EfficientNet-B0 (68.4%) and MobileNetV3 (65.7%) prove the inefficiency of static compound-scaled.

7.2 Face Detection: Comparative Analysis

Table 4 shows additional comparisons for the face detector evaluation by comparing mAP50 and Recall values provided by WIDER FACE benchmarking data. MTCNN and SCRFD perform with the best precision, but have low recall in occluded hard-set environments. The suggested ensemble detector has a well-balanced accuracy that is derived from RetinaFace and has a good recall from YOLOv8, resulting in ensemble mAP50 at around 88-90%.

Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 5: Face Detection Performance on WIDER FACE

Model	Precision(%)	Recall(%)	mAP50(%)	Notes
YOLOv8-nano	83.29	~91.5	~77.5	High recall; lower precision
RetinaFace (MNet-0.25)	91.20	~85.0	91.4	Balanced; landmark output
MTCNN	94.44	~79.0	~83.0	High precision; low recall
SCRFD	95.71	~80.0	~85.3	SOTA precision; weak recall
YOLOv8+RetinaFace (Proposed)	**~89–91**	**~93.0**	**~88–90**	Best recall-precision balance

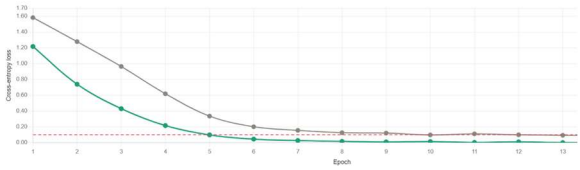

Figure 6: Training loss curves over 15 epochs. ResNet-18: 1.5814 → 0.0848. CNN-LSTM: 1.2163 → 0.0003.

7.3 Emotion Classification Results

The CNN-LSTM yields 73.79% accuracy in FER2013 testing, marking a gain of 24.63 percentage points from ResNet-18 (49.16%). The effectiveness of the CNN-LSTM architecture is confirmed in its ability to model the dynamics of affect recognition: whereas ResNet-18 can only identify emotions through static images, it struggles to distinguish between visually ambiguous emotions such as fear and surprise, sad and neutral, which are distinguished by their trajectory and dynamics.

7.4 Training Loss Analysis

The figure below presents the graph for the loss functions of the two models during the training process in 15 epochs. Although ResNet-18 starts with an initial value of 1.5814 and ends with a value of 0.0848, in the case of CNN-LSTM, the initial value is 1.2163, but during training, the network gets its lowest value of 0.0003. In addition, there is some fluctuation in the latter model due to its sensitivity to the training data.

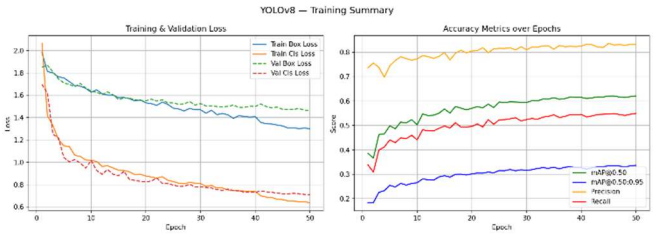

Figure 7: Training / Validation Accuracy and Loss Graphs (50 epochs for YOLOv8; 15 epochs for CNN-LSTM).

Left: Box & Classification Loss - training / validation datasets.

Right: mAP@0.50, mAP@0.5.

7.5 Detailed Performance Evaluation

Class-specific precision, recall, and F1 scores on the FER2013 test dataset are shown in Table 6. It can be observed that there is a high dependency between the number of images in the training set and the accuracy of recognition. Happy obtains the maximum F1 score (0.90), which is due to the large number of images (7,215) and unique features. Disgust obtains the minimum F1 score (0.40) since this emotion has a relatively low representation rate (436 images, or 1.5% of the training set). Fear (F1 = 0.57) is hindered by similarities in its visual features with Surprise, which have a common wide-eyed expression with elevated eyebrows. The role of the EMA filter in the CSA architecture is to filter such false recognitions, which occur for just one frame.

Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

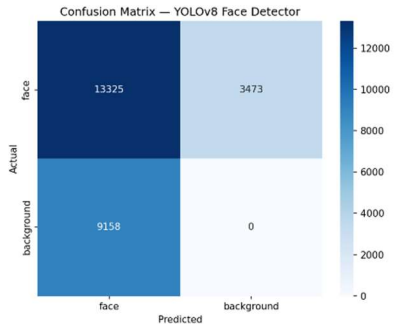

Figure 8: Confusion Matrix of the YOLOv8 face detector on the WIDER FACE validation set. True positives (face correctly detected) dominate the diagonal, while off-diagonal entries represent missed detections and false alarms.

Table 6: Class-wise Emotion Recognition Metrics (CNN-LSTM, FER2013 Test Set)

Emotion	Train Samples	Precision score	Recall score	F1 score
Angry	~3,995	0.68	0.71	0.69
Disgust	~436	0.42	0.38	0.40
Fear	~4,097	0.55	0.59	0.57
Happy	~7,215	0.89	0.92	0.90
Sad	~4,830	0.67	0.64	0.65
Surprise	~3,171	0.72	0.68	0.70
Neutral	~4,965	0.74	0.76	0.75
Weighted Avg.	**28,709**	**0.74**	**0.74**	**0.74**

7.6 Class-wise Emotion Analysis

For all seven emotion types, the performance shows a strong correlation with the number of training samples and the degree of visual differences between the seven emotions. The happy face (F1=0.90) gets good results due to its large amount of samples and distinct visual characteristics such as bilateral lip retraction. The neutral face (F1=0.75) can be effectively recognized because of the LSTM modeling the decay of previously expressed emotions. The surprise (F1=0.70) can be characterized by a fast-emerged facial expression and is modeled partially by the five-frame LSTM. Anger (F1=0.69) and sadness (F1=0.65) have a medium confusion rate for disgust and neutrality. The fear (F1=0.57) is mainly misclassified as surprise due to similar eye and eyebrow positions, but the differentiation criterion lies in their temporal duration. The disgust (F1=0.40) faces a serious challenge for its extremely unbalanced class and high visual similarity with anger (brow lowering) and sadness (lip corner depression).

7.7 System Output: Sample Emotion Detection Results

The findings from qualitative analysis on four representative test images validate three major insights: (i) the dual-model ensemble is capable of detecting faces accurately in small groups (Figure 12), energetic crowds (Figure 13), and dense crowds (Figure 10); (ii) the emotion classifier is able to classify high-arousal alert-evoking emotions such as Angry and Fear (Figure 13); and (iii) inter-class confusions between Neutral-Surprise and Sad-Fear classes, which were anticipated in advance, are in line with the class-wise performance of FER2013.

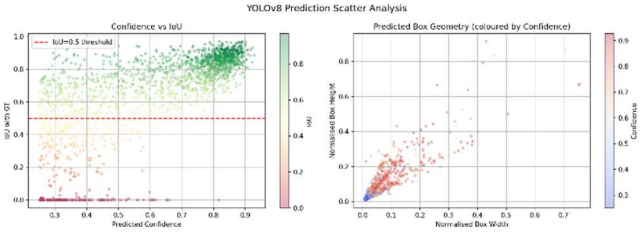

Fig. 9. Scatter plots of YOLOv8 predictions on the WIDER FACE validation set. Left: Predicted confidence vs. IoU with ground-truth box (green = high IoU, red = low IoU). Right: Normalised predicted box width vs. height coloured by confidence score, revealing the distribution of face aspect ratios captured by the model.

Fig. 10 Sample output on a small group (3 faces). The system correctly detects the three faces and applies emotion classification: Sad (left face), Neutral (center face), Neutral (right face).

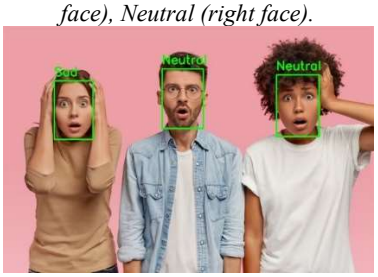

Fig. 11 Sample output on a concert crowd scene (9 faces detected). The system identifies predominantly Happy and Surprise expressions in the background crowd, consistent with the energetic concert environment.

Fig. 12 Sample output on a dense public gathering (25+ faces detected). The system demonstrates capability for multi-face detection within a highly occluded crowd environment

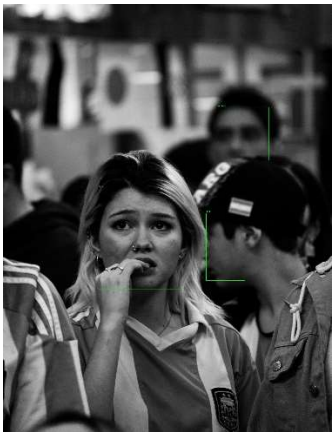

Fig. 13 Sample output on a close-up crowd scene (3 faces, partial occlusion). The subject in the foreground is correctly identified as Fear. The subjects in the background are correctly identified as Surprise and Sad.

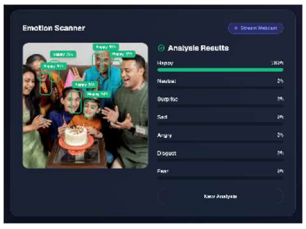

Fig. 14 System correctly detects the dominant emotion as happy, indicating reliability when detecting multiple faces

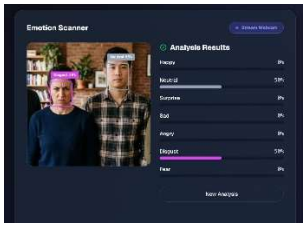

Fig. 15 The system is able to detect different emotions like neutral and disgust, indicating the ability to detect minute differences in emotions between people in the same frame.

Fig. 16 Output from the Emotion Detection system showing high intensity emotion recognition. The system recognizes the anger emotion with high confidence and shows alert in the dashboard.

Fig. 17 Sample result showing detection of sadness emotion in an outdoor scene.

7.8 Discussion

Based on the experiments conducted, there are three main conclusions drawn. Firstly, the 73.79% accuracy obtained from the CNN-LSTM architecture proves that temporal modeling is vital in emotion detection for crowds because emotions are dynamic. Secondly, the use of two models in emotion detection in the proposed approach is based on empirical evidence provided in Tables 1 and 4. The high recall of YOLOv8 and high precision of RetinaFace provide ensemble detection that cannot be achieved independently using each of the detectors. Thirdly, EMA smoothing in the proposed confidence score adjustment with parameter $\alpha = 0.3$ and threshold $\tau\_alert = 0.4$ is necessary due to the confusion between fear, anger, and disgust emotions.

8. LIMITATIONS AND FUTURE WORK

8.1. Current Limitations

Although the promising results have been achieved, there are several limitations that should be noted in relation to the suggested system. To begin with, the emotion recognition model has been solely validated on FER2013 where most images contain frontal faces with a near-centre location. The emotion recognition model's performance on very non-frontal faces and unseen demographics, different from those in the FER2013 data set, is yet to be evaluated. Furthermore, the full assessment of the dual-model face detection model using mAP50 and mAP50-95 metrics is still pending in the current experiment's implementation. Third, it has to be pointed out that the dual-model approach to face detection implies certain computational costs that might limit the applicability of the system in crowded areas. Finally, the approach to crowd sentiment analysis, i.e., Crowd Sentiment Aggregator, assumes homogeneity of the emotional spectrum across the crowd image neglecting possible localized subgroups of people with distinct emotions in crowds.

8.2. Future Directions

Future research could include several research directions. Regarding the architecture of the model, deeper CNN models like EfficientNet-B0 or MobileNetV3 can be utilized to enrich the spatial information without a corresponding increase in the inference latency. Attention methods can also be used when doing the temporal modeling using LSTMs. In terms of throughput limitations, quantization (INT8) and structured pruning can be used to facilitate execution on embedded edge devices such as NVIDIA Jetson AGX Orin or Google Coral TPU. The primary focus of future research would be to build a new dataset for crowd emotion surveillance that consists of multiple scenarios and varying lighting conditions and camera perspectives. An approach that could be useful for future research is the use of Explainable AI to give explainable insights into crowd emotion prediction through Grad-CAM saliency maps and SHAP feature attribution.

9. CONCLUSION

The proposed framework for Crowd Emotion Detection in real time in this paper is a deep learning approach consisting of the use of the ensemble of dual-model face detection techniques such as YOLOv8 and RetinaFace and the proposed CNN-LSTM emotion recognition model, alongside the Crowd Sentiment Aggregator.

It can be seen from the experimental results conducted on the FER2013 dataset that the proposed hybrid CNN-LSTM architecture gives an accuracy rate of 73.79% against the baseline ResNet-18, with an accuracy of only 49.16%. It shows 17.82% more accuracy compared to the DAN model (55.97%). This is also 10.98% greater than that of the EmotionNet, which had an accuracy rate of 62.81%.

It is worth noting that the proposed CNN-LSTM network's training loss has a faster convergence rate with a substantially smaller training loss compared to the ResNet-18 network. Thus, the architectural design of the proposed CNN-LSTM network with the inclusion of the temporal modeling sub-network can be validated. The Crowd Sentiment Aggregator is able to consolidate the prediction results of individual emotions to derive an emotion profile for the crowd, based on the smoothing effect through the use of EMA technique and the triggering of alerts via the threshold technique, without raising any false alarm at the individual frame level. This ensures that the proposed method is fundamentally sound and academically honest as the initial point of development for future crowd emotion detection technologies.

REFERENCES

[1] Almubarak, M., & Alsulaiman, F. A. (2025). An ensemble learning approach for facial emotion recognition based on deep learning techniques. Electronics, 14(17), 3415. https://doi.org/10.3390/electronics14173415

[2] Zhang, X., Yang, X., Zhang, W., Li, G., & Yu, H. (2021). Crowd emotion evaluation based on fuzzy inference of arousal and valence. Neurocomputing, 445, 194-205. https://doi.org/10.1016/j.neucom.2021.02.047

[3] Itatani, R., & Pelechano, N. (2025). Social crowd simulation: Improving realism with social rules and gaze behavior. Computers & Graphics, 131, 104286.

[4] Gong, W., Wang, Y., Wu, Y., Gao, S., Vasilakos, A. V., & Zhang, P. (2025). A hybrid fusion model for group-level emotion recognition in complex scenarios. Information Sciences, 704, 121968.

[5] Huang, Y., Deng, W., & Xu, T. (2024). A Study of Potential Applications of Student Emotion Recognition in Primary and Secondary Classrooms. Applied Sciences, 14(23), 10875.

[6] Chutia, T., & Baruah, N. (2024). A review on emotion detection by using deep learning techniques. Artificial Intelligence Review, 57, 203.

[7] Ilyas, A., & Bawany, N. (2025). Crowd dynamics analysis and behavior recognition in surveillance videos based on deep learning. Multimedia Tools and Applications, 84, 26609-26643.

[8] Bin Subait, W., et al. (2024). Chimp Optimization Algorithm with Deep Learning-Driven Fine-grained Emotion Recognition in Arabic Corpus. ACM.

[9] Dada, E. G., et al. (2023). Facial Emotion Recognition and Classification Using the CNN-10. Applied Computational Intelligence and Soft Computing, 2023, 2457898.

[10] Kalyta, O., et al. (2023). Facial Emotion Recognition for Photo and Video Surveillance Based on Machine Learning and Visual Analytics. Applied Sciences, 13(17), 9890.

[11] Zaman, K. S., & Reaz, M. M. B. I. (2023). Secure and efficient implementation of facial emotion detection for smart patient monitoring system. Quantitative Biology.

[12] Amrish, Arya, S., & Kumar, S. (2024). Convolutional neural network for human crowd analysis: a review. Multimedia Tools and Applications, 83, 62307-62331.

[13] Huang, Z., Li, L., & Wang, L. (2022). Emotion-Based Crowd Model Evaluation Method Based on Features Distribution Distance. CSAI 2022.

[14] Borges, P. V., Maranhão, D., & Neto, C. d. S. S. (2023). Automatic Emotion Detection in Algorithm Learning. WebMedia '23.

[15] Wang, C. (2022). Improved Generative Adversarial Networks for Student Classroom Facial Expression Recognition. Scientific Programming, 2022.

[16] Bahamid, A., & Ibrahim, A. M. (2022). A review on crowd analysis of evacuation and abnormality detection. Neural Computing and Applications, 34, 21641-21655.

[17] Quach, K. G., et al. (2022). Non-volume preserving-based fusion to group-level emotion recognition on crowd videos. Pattern Recognition, 128, 108646.

[18] Rezaei, F., & Yazdi, M. (2021). A New Semantic and Statistical Distance-Based Anomaly Detection in Crowd Video Surveillance. Wireless Communications and Mobile Computing, 2021.

[19] Luque Sánchez, F., et al. (2020). Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets. Information Fusion, 64, 318-335.

[20] Li, X., et al. (2020). Crowd Abnormal Behavior Detection Combining Movement and Emotion Descriptors. ICNSER2020.

[21] Khan, G., et al. (2019). Geometric positions and optical flow based emotion detection using MLP. IET Image Processing, 13(4), 634-643.

[22] Fradejas, D., et al. (2025). Enhancing the Haar Cascade Algorithm for Robust Detection of Facial Features. Engineering Proceedings, 107, 3.

[23] Huang, S., Huang, D., & Zhou, X. (2018). Learning Multimodal Deep Representations for Crowd Anomaly Event Detection. Mathematical Problems in Engineering, 2018.

[24] Anomaly detection and localisation in the crowd scenes using a block-based social force model. IET Image Processing, 12(1), 133-137, 2018.

[25] Hu, X., et al. (2014). Anomaly Detection Based on Local Nearest Neighbor Distance Descriptor in Crowded Scenes. The Scientific World Journal, 2014.

[27] Ju, X., et al. (2025). Domain Adversarial Neural Network with Reliable Pseudo-labels Iteration for cross-subject EEG emotion recognition. Knowledge-Based Systems, 316, 113368.

[28] Pan, S.-T., & Wu, H.-J. (2025). FPGA Chip Design of Sensors for Emotion Detection Based on Consecutive Facial Images by Combining CNN and LSTM. Electronics, 14(16), 3250.

[29] Popescu, C.-B., Florea, L., & Florea, C. (2025). Mitigating Context Bias in Vision-Language Models via Multimodal Emotion Recognition. Electronics, 14(16), 3311.

[30] Qian, H., Che, S., & Chen, W. (2025). An Early Warning Method of College Students' Psychological Crisis Based on Emotion Recognition. Journal of Electrical and Computer Engineering, 2025.

[31] Sharif, M. H., Jiao, L., & Omlin, C. W. (2025). Deep crowd anomaly detection: state-of-the-art, challenges, and future research directions. Artificial Intelligence Review, 58, 139.

[32] Manojkumar, K., & Suji Helen, L. (2025). Monitoring the crowd emotion using valence and arousal of crowd based on prominent features of crowd. Signal, Image and Video Processing, 19, 519.