Computer Systems
Science and Engineering

Tech Science Press

https://doi.org/10.32604/csse.2026.081278

ARTICLE

SAFE: A Semantic Audio-Visual Fusion Engine for Interpretable and Real-Time Crowd Anomaly Detection

Ravi Saharan, Akrisht Singh and Prakash Choudhary*

Department of Computer Science and Engineering, Central University of Rajasthan, Ajmer, India

*Corresponding Author: Prakash Choudhary. Email: prakash.choudhary@curaj.ac.in

Received: 27 February 2026; Accepted: 03 September 2026; Published: 21 September 2026

ABSTRACT: The task of monitoring crowds for safety is a critical challenge, yet traditional surveillance systems are often visual only, error-prone, and lack interpretability. This paper presents SAFE (Semantic Audio-Visual Fusion Engine), a real-time, multi-modal framework that delivers interpretable, operator-facing alerts by fusing complementary audio-visual cues. The visual pipeline couples SSD–ResNet face detection and Hungarian tracking with facial emotion recognition and DBSCAN-based clustering of negative affect to compute a semantic visual anomaly score that includes an explicit overcrowding signal. In parallel, the audio pipeline extracts MFCCs (Mel-Frequency Cepstral Coefficients) and employs a lightweight one-dimensional convolutional neural network (CNN) to estimate the probability of an audio anomaly. A weighted late-fusion mechanism yields a unified alert score and an annotated display that includes person IDs, emotions, and alert banners. Component-wise, the audio model attains an area under the ROC curve (AUC) of 0.99 on UrbanSound8K and 0.82 on ESC-50 public datasets. End-to-end, the system runs on a standard CPU with real-time throughput of 19.37–28.40 FPS across scenarios with 32.93–48.78 ms latency across sparse, dense, and high-motion scenes, consistently triggering fused, overcrowding, and emotion cluster alerts. Compared with single modality baselines, the proposed framework demonstrates higher robustness under low light and noisy acoustic conditions while preserving person-level interpretability, supporting practical deployment in public spaces.

KEYWORDS: Crowd analysis; anomaly detection; multi-modal fusion; affective computing; edge computing; emotion-aware systems

1 Introduction

Ensuring public safety in dynamic crowd environments remains a persistent challenge [1], as traditional video-centric surveillance systems are prone to operator fatigue and inherently miss critical non-visual precursors to distress, such as screams or acoustic anomalies [2,3]. To address these limitations, recent advances in deep learning have matured sufficiently to support integrated multi-modal frameworks that capture both visual and acoustic signals of collective behavior. On the visual front, the combination of real-time Single Shot Detectors (SSD) [4] and robust assignment-based tracking [5] can now be effectively augmented with Facial Emotion Recognition (FER) [6,7] and density-based clustering [8] to quantify spatial patterns of negative affect. In parallel, the efficiency of Mel-frequency cepstral coefficients (MFCCs) [9] paired with convolutional architectures [10] enables reliable acoustic event classification. Orchestrating these complementary components offers a viable path toward the interpretable, operationally feasible audio-visual fusion engines previously envisioned in affective computing studies [11].

Nevertheless, current research on anomaly detection in video surveillance continues to face significant bottlenecks. Deep learning has markedly improved detection accuracy using weak or self-supervised paradigms [12–14], yet these successes often hinge on powerful GPU-based infrastructures and deep spatiotemporal backbones that hinder deployment in real-time edge environments. Furthermore, most methods produce opaque segment- or clip-level predictions that cannot be directly interpreted by human operators monitoring live feeds. As multimodal datasets such as XD-Violence [15] expand, the potential value of complementary cues becomes evident. However, many reported approaches remain computationally heavy, relying on high-dimensional feature pooling or transformer fusion, limiting both transparency and CPU-grade real-time adoption [16]. This tension between high accuracy and practical deployability underscores an unmet need for lightweight but interpretable multimodal systems capable of bridging research prototypes and real-world deployment.

To address these gaps, we propose a practical, real-time audio–visual anomaly detection framework explicitly designed around three guiding principles: computational efficiency, interpretability, and operator usability. The visual pipeline employs a fast SSD–ResNet face detector and a Hungarian-based multi-object tracker to maintain individual identities across frames. Building upon these tracked entities, we incorporate facial emotion recognition (FER) to infer per-person affective states and identify negative emotions such as anger, fear, and sadness. These high-dimensional emotional cues are aggregated spatially through a DBSCAN clustering algorithm to reveal local regions of collective distress or agitation, forming a rich semantic representation of the scene.

In parallel, the audio pipeline processes the ambient sound stream using lightweight signal transformations and a compact 1D CNN applied over extracted MFCCs to detect acoustic anomalies including screams, alarms, or crowd spikes that may signal early-stage emergencies. A weighted late fusion scheme integrates both modalities into a unified anomaly score that is human-interpretable and timely. This score directly drives visual overlays and alert banners in an operator-facing interface, providing actionable semantics such as *Overcrowding*, *Emotion Cluster*, or *Fused Alert* in real time. Unlike complex models that operate as "black boxes," our system aims to supply transparent, situationally contextual information that supports rapid assessment and intervention.

The efficacy of our framework is rigorously validated on a suite of benchmark datasets spanning both audio and visual domains: FER-2013 for emotion recognition [17], UrbanSound8K for sound classification [18], and ESC-50 for generalized acoustic event detection [19]. With all components implemented in Python using accessible libraries such as OpenCV, DeepFace, TensorFlow, and librosa [20], the system achieves high discriminative accuracy while maintaining end-to-end real-time throughput on standard multi-core CPUs. The results demonstrate that effective multimodal crowd understanding does not require computationally intensive architectures but rather well-engineered synergy between perceptual modalities, optimized fusion, and transparent visualization. The main contributions of this work are:

- We propose a user-centric, CPU-based framework capable of generating real-time, interpretable person-level alerts via spatial emotion clustering, thereby significantly enhancing human monitoring efficiency.
- We develop a lightweight fusion strategy that integrates a 1D CNN-based audio anomaly detector with visual cues using a weighted late-fusion mechanism, ensuring system robustness across varying acoustic and visual conditions.
- We perform a comprehensive evaluation across multiple datasets to validate our proposed framework's accuracy, interpretability, and real-time performance, confirming its suitability for practical deployment in resource-constrained security environments.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the proposed methodology in detail; Section 4 presents implementation aspects; Section 5 reports experimental results and discussion; and Section 6 concludes the paper with insights on future directions.

2 Related Work

Research on crowd anomaly detection and multimodal surveillance spans several complementary domains: visual understanding, acoustic analysis, and cross-modal fusion. In this section, we summarize representative contributions in each domain and identify the limitations that motivate us for our proposed framework.

2.1 Visual Anomaly Detection and Crowd Analysis

Early visual crowd analysis techniques sought to model global scene dynamics through optical flow and motion patterns [2,3], though these holistic approaches often failed to localize individuals responsible for abnormal behaviors. The subsequent shift toward deep convolutional networks enabled finer object-level discrimination necessary for real-time deployment [21]. To achieve the required efficiency and stability in crowded scenes, modern pipelines increasingly utilize single-shot detectors like SSD [4] alongside assignment-based tracking algorithms such as SORT [5] to maintain continuous identity traces.

To enhance interpretability, recent research has integrated semantic cues like facial emotion recognition (FER) to bridge raw vision with human perception. Robust toolkits such as DeepFace [7] and benchmarks like FER-2013 [17] allow systems to quantify emotional distress, while density-based clustering algorithms like DBSCAN [8] help identify spatial regions of collective agitation. Surveys of CNN-based crowd counting and density estimation further emphasize the need for computationally efficient crowd analysis in public-safety applications [22]. Despite these improvements, visual-only methods remain vulnerable to environmental factors such as poor lighting or occlusion and often produce anomaly scores lacking semantic grounding, highlighting the need for multimodal approaches.

2.2 Audio Event Detection and Acoustic Modeling

Environmental audio carries rich contextual cues that often precede visible anomalies, such as screams, explosions, or equipment malfunctions. Compact frequency-domain representations, particularly Mel-frequency cepstral coefficients (MFCCs), remain a cornerstone for modelling acoustic texture in environmental sound recognition [9]. With the advent of deep learning, convolutional architectures have achieved state-of-the-art results for large-scale audio classification benchmarks [10].

Datasets like UrbanSound8K [18] and ESC-50 [19] facilitate systematic evaluation of environmental sounds across a diverse range of acoustic settings from city traffic to human exclamations, while libraries such as `librosa` [20] enable unified and reproducible feature extraction pipelines. Beyond classification, acoustic characteristics such as sound energy, pitch, and spectral features have been explored for affective-state recognition, providing complementary information to visual evidence [11]. However, many robust audio models are computationally expensive or trained on domain-specific data, limiting their straightforward adoption in edge-deployed surveillance systems. The need persists for lightweight approaches that balance generalisation capability with CPU-grade real-time throughput.

2.3 Multimodal Fusion for Surveillance and Affective Understanding

Combining auditory and visual streams significantly enhances recognition robustness [23], particularly in degraded environmental conditions [11]. While video anomaly detection has advanced via weakly

supervised multiple instance Learning frameworks [12–14], these methods typically produce coarse clip-level probabilities that lack interpretability. Multimodal benchmarks like XD-Violence [15] confirm the value of audio cues, yet state-of-the-art architectures often rely on heavy 3D CNNs or transformers. Even recent self-training variants [16], despite improved accuracy, retain high computational complexity, which hinders practical real-time deployment [24].

2.4 Identified Gaps and Design Motivation

Positioning the present study within this landscape reveals several critical limitations in existing research. First, current systems frequently lack interpretable multimodal reasoning, as they typically output abstract anomaly scores without distinguishing whether an event stems from acoustic distress, visual aggression, or collective crowd tension. This opacity compels human operators to rely on subjective interpretation rather than actionable data. Second, a majority of frameworks perform assessments at the region or clip level, neglecting the per-person attribution that is necessary for forensic traceability and operator confidence. Finally, the prevalent dependence on high-capacity GPU infrastructures hinders deployment in resource-constrained public environments, highlighting an unmet need for efficient algorithms capable of real-time execution on standard CPUs.

Our proposed framework directly addresses these shortcomings by integrating fast visual detection/tracking [4,5], emotion-driven spatial clustering [7,8,17], and lightweight audio anomaly modeling [9,10,18,19]. Through a weighted late fusion mechanism informed by multimodal literature [11], the system yields explicit, human-interpretable alerts (e.g., "Overcrowding," "Emotion Cluster," and "Fused Alert") at real-time frame rates on commodity CPUs. By prioritizing interpretability, person-level anchoring, and deployability, our approach closes the gap between highly accurate research models and field-ready surveillance tools.

3 Methodology

As illustrated in Fig. 1, our proposed framework operates via two parallel processing streams derived from a single input source. The Visual Pipeline extracts semantic crowd cues, specifically identity, emotion, and spatial density, to compute a normalized visual anomaly score (S_{visual}). Simultaneously, the Audio Pipeline detects spectral irregularities using a lightweight CNN to generate an audio anomaly probability (S_{audio}). These complementary signals are integrated via a weighted late fusion mechanism to trigger interpretable, operator-facing alerts.

3.1 CNN-Based Audio Anomaly Classifier

The audio branch employs a lightweight one-dimensional convolutional neural network (1D CNN) to detect acoustic irregularities such as sharp acoustic events such as sirens, horns, and gunshots [25]. It is optimized for real-time inference on standard CPUs, balancing discrimination power with latency and interpretability.

3.1.1 Input and Feature Representation

The audio stream is segmented into overlapping windows, and forty-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) are extracted using `librosa` at 22.05 kHz, 25 ms frame length, and 10 ms hop size. These features form $L \times 40$ tensors normalized by a global scaler to ensure consistency across input samples.

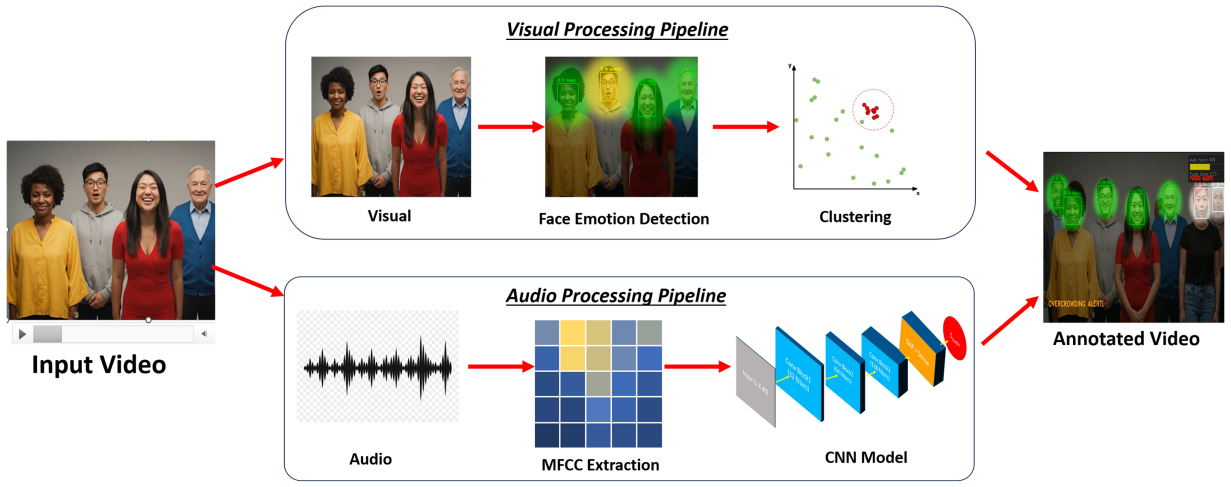


Figure 1: The proposed SAFE architecture. The visual pipeline processes frames through SSD-ResNet and DBSCAN clustering. The audio pipeline converts waveforms to MFCCs for 1D-CNN analysis. Both pipelines are integrated through a late fusion mechanism.

3.1.2 Network Architecture

The model structure is depicted in Fig. 2 and detailed specified in Table 1. It includes three convolutional blocks (`Conv1D-BN-ReLU-MaxPool`) with 32, 64, and 128 filters (kernel = 3) followed by global average pooling, a dropout layer (rate 0.5), and a sigmoid output neuron that yields the anomaly probability $S_{\text{audio}} \in [0, 1]$. The total parameter count remains under one million, meeting practical CPU constraints.

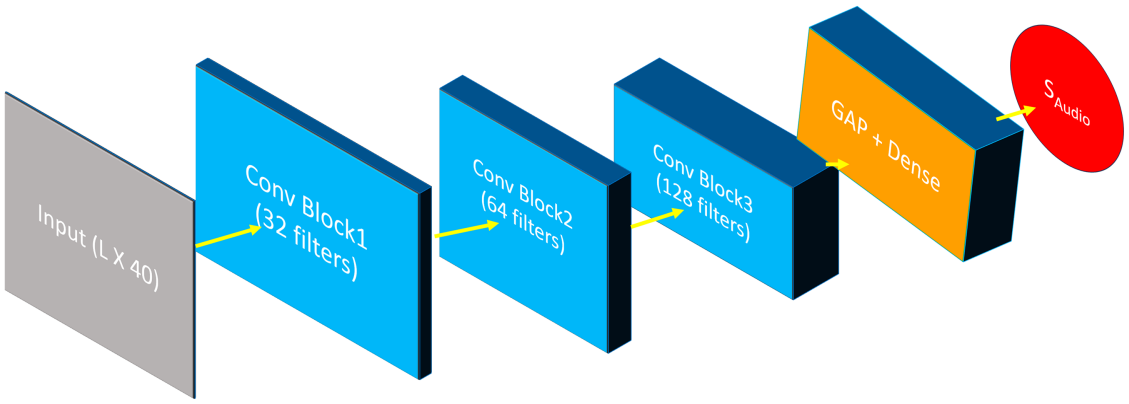


Figure 2: Block diagram of the 1D CNN for audio anomaly detection. Three convolutional blocks extract temporal–spectral features; global pooling and dropout regularization lead to a single sigmoid output S_{audio}.

3.1.3 Training and Deployment

The network is trained using binary cross-entropy loss optimized with Adam (10^{-4} learning rate) and early stopping on validation loss. Data augmentation (additive noise, mild pitch/tempo shifts) improves robustness to acoustic variability. For deployment, the model is exported in TensorFlow Lite format with 16-bit quantization, achieving inference latency below 2 ms per window. The resulting probability stream is exponentially smoothed before late fusion with the visual score.

Table 1: Detailed architectural specification of the 1D-CNN audio classifier.

Layer (Type)	Output Shape	Filters/Units	Kernel/Pool	Activation/Rate
Input Layer	$(T, 40)$	–	–	–
Conv1D (Block 1)	$(T, 32)$	32	$k = 3$	ReLU
Max Pooling 1D	$(T/2, 32)$	–	$s = 2$	–
Conv1D (Block 2)	$(T, 64)$	64	$k = 3$	ReLU
Max Pooling 1D	$(T/2, 64)$	–	$s = 2$	–
Conv1D (Block 3)	$(T/2, 128)$	128	$k = 3$	ReLU
Global Avg Pool	$(128,)$	–	–	–
Dropout	$(128,)$	–	–	0.5
Output Layer	$(1,)$	1	–	Sigmoid

Note: T denotes the temporal dimension (number of time steps) derived from the 4.0 s MFCC extraction window. With a sampling rate of 22,050 Hz and a hop size of 10 ms (220 samples), a 4.0 s window yields $T = \lfloor (88{,}200 - 551)/220 \rfloor + 1 \approx 400$ time steps in practice.

3.2 Visual Analysis Module

The visual analysis module converts raw video frames into person-centric, interpretable evidence for anomaly assessment. Faces are first localized with a fast SSD–ResNet detector, and detections are linked across time using a centroid tracker with Hungarian assignment to maintain stable IDs in crowded scenes. On these tracks, facial emotion recognition (FER) is executed asynchronously at fixed intervals (DeepFace) to estimate a per-person negative affect probability while controlling compute and reducing jitter. From these per-person signals, two complementary scene-level cues are derived: (i) overcrowding, computed from the number of active tracks, and (ii) localized distress, obtained by applying DBSCAN to the centroids of "alert" faces with an adaptive neighborhood to account for perspective. The cues are normalized and temporally smoothed into a visual anomaly score S_{visual}, which also drives the operator overlays (IDs, emotions, clusters) used in downstream fusion and alerting.

3.2.1 Face Detection and Tracking

For each frame, a pre-trained SSD-based face detector identifies faces with a confidence score above a threshold τ_{conf}. A centroid tracker enhanced with the Hungarian algorithm associates these detections with existing tracks by minimizing the Euclidean distance between centroids. Tracks are terminated if a face is unobserved for $F_{\text{max\_disappeared}}$ frames.

3.2.2 Emotion Recognition

To provide sentiment cues without impacting real-time performance, Facial Emotion Recognition (FER) is handled asynchronously. For each tracked individual, the DeepFace library periodically analyzes their face ROI. The dominant emotion is used for the operator-facing visual overlay, while a "negative affect" signal, aggregating 'angry,' 'fear,' and 'sad,' is computed for anomaly analysis. This signal is temporally smoothed to prevent alert flicker from transient expressions or minor misclassifications. Finally, the centroids of individuals exhibiting persistent negative affect are passed to the DBSCAN module for spatial clustering.

3.2.3 Visual Anomaly Score Calculation

Once the per-person states (location and emotion) are established, the framework aggregates this information to compute a high-level, interpretable visual anomaly score S_{visual}. The anomaly score relies on two complementary conditions:

- **Overcrowding:** Triggered when the number of tracked faces exceeds a defined threshold T_{crowd}.
- **Emotion Clustering:** Activated when DBSCAN detects spatial clusters of at least $N_{\text{min\_cluster}}$ individuals exhibiting "alert" negative affect.

Let N_t denote the number of active tracks in frame t and A_t the number of those tracks whose smoothed affect state is classified as "alert" (angry, fear, or sad). Each cue is normalized to $[0,1]$ and the score is taken as the more severe of the two, so that either condition alone is sufficient to raise it:

$$\tilde{S}_{\text{visual}}(t) = \max\left(\min\left(\frac{N_t}{T_{\text{crowd}}}, 1\right), \frac{A_t}{\max(N_t, 1)}\right) \tag{1}$$

No relative weighting is applied between the cues. Temporal smoothing of both modality scores is defined in Section 3.4.

3.3 Audio Analysis Module

The audio analysis module provides a critical complementary signal to the visual pipeline, capturing auditory distress cues—such as screams or panicked shouts—that may lack a clear visual signature. The acoustic stream is partitioned into windows of fixed duration T_{win}, from which 40-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) are extracted to form a feature tensor of shape $(T, 40)$, where T denotes the temporal sequence length. These features are standardized via a pre-fitted scaler and processed by a specialized 1D Convolutional Neural Network (CNN). As detailed in Table 1, the model utilizes a series of temporal convolutions and Global Average Pooling (GAP) to collapse the spectral-temporal features into a latent representation. The final layer applies a sigmoid activation to generate a probabilistic anomaly score:

$$\tilde{S}_{\text{audio}}(t) = \text{CNN}_\theta(\text{Scale}(\text{MFCC}(x_t))) \tag{2}$$

where x_t denotes the audio window at time t. Each instantaneous score is stabilised by a short-term exponential moving average (EMA), $S_m(t) = \lambda \tilde{S}_m(t) + (1-\lambda) S_m(t-1)$ for $m \in \{\text{audio}, \text{visual}\}$, where the smoothing factor is set to $\lambda = 0.3$.

3.4 Audio-Visual Fusion and Alerting

To produce a single, reliable metric, the scores are combined using a weighted linear sum in a late fusion scheme. Each instantaneous score is stabilised by a short-term exponential moving average (EMA).

$$S_{\text{fused}}(t) = w_{\text{audio}} S_{\text{audio}}(t) + w_{\text{visual}} S_{\text{visual}}(t) \tag{3}$$

where $w_{\text{audio}} + w_{\text{visual}} = 1$. A final **Fused Alert** is triggered if $S_{\text{fused}}(t)$ surpasses the threshold τ_{fused}.

3.5 System Output and Operator-Centric Visualization

The final output of our proposed framework is an annotated video stream designed to provide a human operator with immediate, actionable insights. For each frame, the system overlays bounding boxes on tracked individuals, along with their assigned ID and detected emotion. A dashboard panel displays the real-time audio, visual, and fused anomaly scores [26]. When an alert is triggered, a prominent notification is displayed

on screen. A baseline system state in a non-anomalous scenario is shown in Fig. 3, demonstrating stable operation with neutral affect and zero anomaly scores.

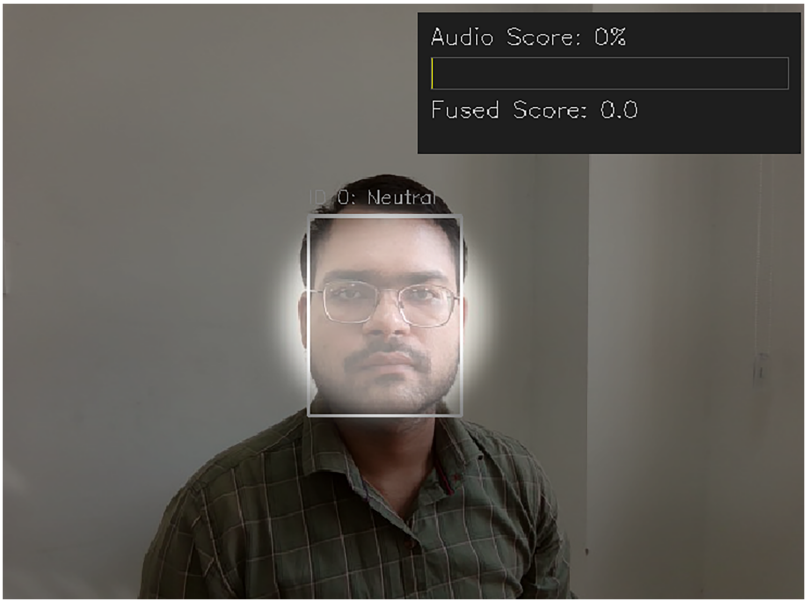


Figure 3: Example of the final annotated output in a non-anomalous, real-time scenario. The system correctly identifies a 'Neutral' face and reports baseline scores, illustrating operational stability.

4 Implementation Details

This section summarizes the software/hardware stack, pipeline configurations, training/inference setup, and deployment choices that enable interpretable, CPU real-time operation. Each module includes concise prose and bullet point checklists for clarity.

4.1 Software and Hardware Stack

The system is implemented in Python, using OpenCV for video I/O/rendering, NumPy/SciPy for numerics, scikit-learn for DBSCAN and scaling, DeepFace for FER, librosa for audio features, and TensorFlow/Keras for the audio CNN. All reported results are from CPU-only inference on a standard multi-core workstation; no discrete GPU is required.

- Core libraries: OpenCV (DNN for SSD), NumPy/SciPy, scikit-learn (DBSCAN, StandardScaler), librosa (MFCC), TensorFlow/Keras (1D CNN), DeepFace (FER).
- Environment: Python 3.14 with version-pinned dependencies; random seeds fixed; model/scaler artifacts stored with configs.
- Hardware: multi-core CPU workstation; results reflect CPU-only inference and rendering.
- Reproducibility: deterministic preprocessing, serialized weights/scalers, and structured logging of metrics and alerts.

4.2 Visual Pipeline

The visual engine processes frames via an SSD–ResNet detector, filtering detections by confidence τ_{conf} and maintaining identity stability through a Hungarian-based centroid tracker. To preserve CPU resources, FER is executed asynchronously at fixed intervals (K_{FER}), focusing on "alert" emotions (Angry, Fear, Sad).

Anomaly reasoning is derived from the convergence of physical overcrowding and localized psychological distress, the latter identified via a perspective-aware DBSCAN clustering over distressed individuals. These signals are normalized and temporally smoothed to generate the final visual anomaly score, S_{visual}. All core operational hyperparameters and heuristic thresholds governing this pipeline are consolidated in Table 2.

Table 2: SAFE visual pipeline configuration and hyperparameters.

Category	Parameter	Symbol/Variable	Value/Setting
Detection	Confidence Threshold	τ_{conf}	0.6
Tracking	Max Disappeared Frames	F_{max}	20 frames
	Max Distance Threshold	–	75 pixels
Emotion	Analysis Interval	K_{FER}	4 frames
	Alert Emotion Set	–	{angry, fear, sad}
Clustering	Min Cluster Size	N_{min}	3 individuals
	Distance Threshold	–	150 pixels
Crowd	Overcrowding Threshold	T_{crowd}	>6 individuals

4.3 Audio Pipeline and Multimodal Fusion

The acoustic engine processes segmented audio windows through a feature extraction layer to produce 40 Mel-Frequency Cepstral Coefficients (MFCCs). These features are standardized via a pre-fitted scaler and analyzed by a lightweight 1D Convolutional Neural Network (CNN) to generate an anomaly probability. Temporal stability is maintained using a Short-term Exponential Moving Average (EMA) to filter transient acoustic spikes, yielding the final score S_{audio}. The multimodal fusion layer then orchestrates the interaction between visual and acoustic signals through a weighted late fusion scheme, governed by a hysteresis band to ensure reliable operator-facing alerts. The complete architecture and fusion parameters are detailed in Table 3.

Table 3: SAFE audio pipeline and fusion configuration.

Category	Parameter	Symbol/Variable	Value/Detail
Features	Sampling Rate	f_s	22,050 Hz
	MFCC Dimensions	N_{MFCC}	40
	Window Duration	T_{win}	4.0 s
1D-CNN	Architecture	–	3 × Conv1D + GAP
	Activation	–	ReLU (Hidden)/Sigmoid (Output)
	Dropout Rate	–	0.5
Fusion	Audio Weight	w_{audio}	0.5
	Visual Weight	w_{visual}	0.5
	Fused Threshold	τ_{fused}	0.5
	EMA Smoothing Factor	λ	0.3

4.4 Experimental Environment and System Specifications

To validate the real-time performance of the SAFE framework, all experiments were conducted on a standardized mobile workstation representing a typical edge-deployment scenario. The system utilized a producer-consumer architecture with asynchronous multithreading to decouple the high-latency FER and CNN inference from the primary tracking loop. The hardware and software specifications used for all reported experiments are detailed in Table 4.

Table 4: Hardware and software experimental environment.

Component	Specification/Version
Processor (CPU)	Intel Core i7-12700H (14 Cores, up to 4.7 GHz)
Memory (RAM)	16 GB DDR4 3200 MHz
Operating System	Windows 11/Linux (Ubuntu 22.04 LTS)
Deep Learning Lib	TensorFlow 2.15/Keras 3.0
Inference Engine	TensorFlow Lite (16-bit Quantization)
Computer Vision	OpenCV 4.8.1/DeepFace 0.0.79
Audio Processing	Librosa 0.10.1/Sounddevice 0.4.6

5 Experimental Results and Discussion

We evaluate our proposed framework in three stages and maintain a consistent mapping between qualitative scenarios (Scenario A: Sparse crowd, Scenario B: Dense crowd, and Scenario C: High motion) and quantitative test cases. First, we assess the core audio/visual modules on benchmarks. Next, we present qualitative results aligned with the three test cases. Finally, we report real-time CPU performance on the same three cases to ensure comparability across analyses.

5.1 Datasets Used

Table 5 summarizes the datasets utilized in this study for training and evaluation of the framework's individual modules. We employ three public benchmarks: one visual and two acoustic, chosen to capture complementary modality characteristics relevant to crowd anomaly detection.

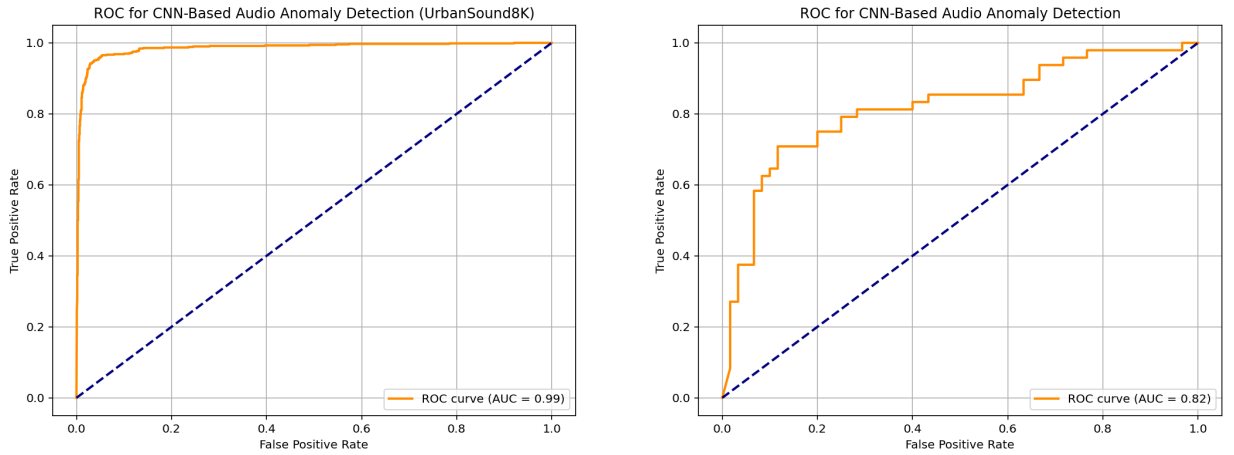
Table 5: Summary of used datasets.

Dataset	Domain	Classes	Samples/Clips	Usage
FER-2013 [17]	Visual (faces)	7 emotions	35,887 images	FER training and validation
UrbanSound8K [18]	Audio (urban)	10 types	8732 clips	Audio CNN training/validation
ESC-50 [19]	Audio (environmental)	50 types	2000 clips	Cross-domain benchmark test

FER-2013 provides gray-scale facial images labeled across seven emotion categories (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral). It trains and validates our Facial Emotion Recognition component to derive per-person affect probabilities used in the visual anomaly analysis.

UrbanSound8K is selected for in-domain training of the 1D CNN audio anomaly detector. It contains short urban event clips such as sirens, dog barks, and children playing, which are acoustically relevant to public space surveillance. For the binary anomaly task, the classes *gun_shot*, *siren*, and *car_horn* are labelled anomalous and the remaining seven classes ambient; for ESC-50, the corresponding alarm-type classes (siren, car horn, glass breaking, fireworks) are labelled anomalous and the remainder ambient. Eight of the

ten predefined folds are used for training, one for validation and decision threshold selection, and one as the held-out test set on which the ROC of Fig. 4a is computed; ESC-50 is used exclusively as an unseen out-of-domain test set.



(**a**) ROC on UrbanSound8K (AUC = 0.99)

(**b**) ROC on ESC-50 (AUC = 0.82)

Figure 4: ROC curves demonstrating the audio model's classification performance.

ESC-50 acts as an out-of-domain benchmark to assess the generalization capability of the audio CNN. The dataset encompasses a wide variety of natural and artificial environmental sounds across five major categories (animals, natural, human, domestic, and exterior). Performance consistency between UrbanSound8K and ESC-50 demonstrates the modality's robustness to domain shift.

5.2 Evaluation Metrics

To provide a quantitative and reproducible assessment, both component-level and system-level metrics are defined formally.

5.2.1 Classification Metrics

The accuracy and F1 score are computed for each module, reflecting recognition quality on respective benchmarks. For a set of N test instances with ground truth labels y_i and predictions $\hat{y}_i$,

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(y_i - \hat{y}_i) \tag{4}$$

where $\mathbf{1}(\cdot)$ is the indicator function.

Precision (P), Recall (R), and F1 score are expressed as

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN}, \tag{5}$$

$$F1 = 2 \times \frac{P \times R}{P + R}, \tag{6}$$

where TP, FP, and FN denote true positive, false positive, and false negative counts. For multi-class problems (e.g., FER-2013), these metrics are macro averaged across classes. Confusion matrices are additionally reported to analyze per-class misclassification trends.

5.2.2 System-Level Latency and Throughput

Real-time feasibility is evaluated through frame-level latency and end-to-end throughput measured during live inference on a CPU-only workstation:

$$\text{Latency (ms)} = \frac{1}{M}\sum_{j=1}^{M} t_{proc}^{(j)}, \tag{7}$$

$$\text{Throughput (FPS)} = \frac{1}{\text{Average Processing Time per Frame}}. \tag{8}$$

Latency represents the average processing time per frame, encompassing detection, tracking, FER, audio inference, fusion, and visualization. Throughput (frames per second) quantifies operational speed; values at or above 20 FPS are considered full real-time. The values in the high teens (≥18 FPS) are considered near real-time and remain operationally viable for human-in-the-loop monitoring.

5.2.3 Alert Level Evaluation

For system-level alerts, framewise precision, recall, and F1 over temporal windows against operator-annotated ground truth are the intended evaluation protocol. In this study, we report qualitative interpretability through visual examination of alert overlays across three representative scenarios, with full quantitative alert-level metrics and inter-annotator agreement left to future work.

These metrics collectively provide a balanced quantitative and qualitative understanding of the proposed system's performance, spanning both recognition accuracy and operational responsiveness.

5.3 Component-Wise Performance Analysis

We first evaluate the visual and audio modules independently to quantify their standalone accuracy and robustness, and to calibrate hyperparameters for the integrated system. The visual pipeline is assessed on FER-2013 (confusion matrix and negative affect reliability), while the audio CNN is validated on UrbanSound8K and ESC-50 (ROC/AUC) under a consistent, CPU-only inference protocol.

5.3.1 Audio Anomaly Detection

The 1D-CNN was trained using a supervised regimen on the UrbanSound8K dataset, employing a binary cross-entropy objective and Adam optimization; the specific feature engineering constraints and hyperparameter settings are detailed in Table 6.

We assess the MFCC + 1D CNN audio module in-domain on UrbanSound8K and out-of-domain on ESC-50 to gauge both accuracy and generalization. The ROC curves in Fig. 4a,b illustrate clear separability between anomalous (e.g., gun shots, sirens) and non-anomalous ambient sounds: on UrbanSound8K the curve hugs the top-left corner with an **AUC of 0.99**, indicating near-perfect discrimination and high true-positive rates even at very low false-positive rates; on ESC-50 the **AUC of 0.82** confirms robust cross-dataset performance despite domain shift. Practically, we select the decision threshold on a validation split (e.g., maximizing Youden's J or F1), then apply light temporal smoothing to reduce spurious spikes. These results

validate the audio stream as a strong, complementary signal—especially useful when visual evidence is weak (low light, occlusion) and well suited for late fusion with the visual score.

Table 6: Training and feature engineering specifications.

Category	Parameter	Value/Setting
Preprocessing	Sampling Rate (f_s)	22,050 Hz
	Window Duration	4.0 s
	Feature Mapping	40-dimensional MFCC
	Temporal Alignment	Constant Zero-padding
Optimization	Optimizer	Adam
	Loss Function	Binary Cross-Entropy
	Learning Rate (α)	10^{-4}
	Batch Size	32
	Training Epochs	30

5.3.2 Facial Emotion Recognition (FER)

The FER component, evaluated on the FER-2013 dataset, demonstrates strong overall performance with an accuracy of **78.25%**, a weighted F1 score of **0.78**, and a precision of **0.79**. Class-specific results indicate that the network performs best on **Happy** (F1 = 0.90), **Surprise** (F1 = 0.84), and **Neutral** (F1 = 0.74) expressions, while slightly lower recall values are observed for emotion classes such as **Fear** (F1 = 0.71) and **Sad** (F1 = 0.72). The corresponding confusion matrix in Fig. 5 confirms that most misclassifications occur between semantically similar negative categories (e.g., Fear ↔ Sad, Angry ↔ Neutral), which is acceptable for our framework, as these emotions jointly convey distress; this overlap is consistent with the category structure reported in meta-analytic reviews of facial expression recognition [27].

To address these overlaps, our system incorporates several stabilizing strategies: **(i)** it aggregates all negative emotions (Angry, Fear, Sad) into a unified "alert" signal, reducing the effect of intra-class confusion; **(ii)** it applies temporal smoothing to suppress short-term misclassifications; and **(iii)** it employs DBSCAN clustering to detect only spatially persistent regions of collective negative affect. Through these design choices, the FER module supplies an interpretable and statistically reliable cue for the visual anomaly scoring process. It is worth noting that severe lighting degradation (e.g., low-illumination night scenes or strong backlighting) can reduce facial contrast, making the subtle textural differences between Fear, Sad, and Angry harder to resolve and potentially amplifying the intra-class confusion already observed between these categories. In such conditions, the temporal smoothing and negative-affect aggregation strategies described above become especially important, as they tolerate individual frame-level misclassifications; furthermore, the complementary audio stream provides an independent anomaly signal that partially compensates when visual evidence is degraded. We note that this explanation is theoretical; a dedicated quantitative evaluation of FER accuracy under controlled low-light and backlit conditions was not performed in this study and is identified as a priority for future work.

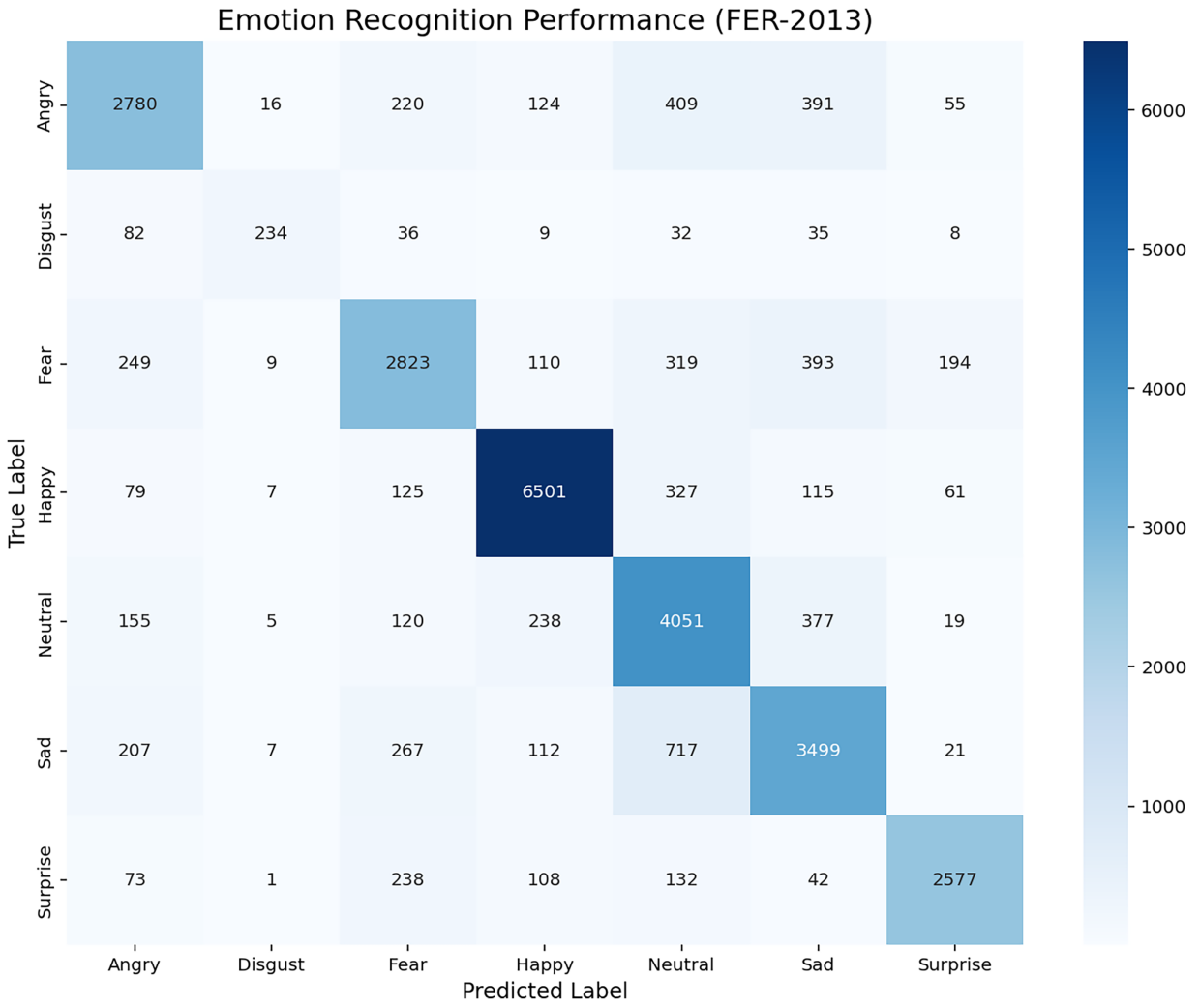


Figure 5: Confusion matrix for the FER module on the FER-2013 dataset.

5.4 Qualitative Analysis Aligned with Test Cases

To bridge the gap between quantitative benchmarks and real-world performance, we present a qualitative analysis of three representative sequences. These scenarios are directly aligned with our quantitative test cases—a **Sparse** crowd (Scenario A), a **Dense** crowd (Scenario B), and a scene with **High Motion** (Scenario C). This analysis demonstrates how the framework's synergy of audio-visual cues and semantic reasoning translates into clear, interpretable, and context-aware alerts for a human operator. The visual outputs for each case are shown in Fig. 6.

Scenario A (Sparse Crowd)

In this low-density scene, as shown in Fig. 6a, a single individual is detected moving quickly through a crowd. While a visual-only system might dismiss this as an outlier, our framework identifies the person's expression as 'Sad', triggering a high visual anomaly score. More importantly, it correlates this visual cue with a moderately elevated audio score of 0.47. As seen in the alert banner, the two modality scores combine to produce a fused score of 0.73, which exceeds the configured fusion threshold and consequently triggers a **"FUSED ALERT!"**. This example illustrates how complementary audio and visual evidence can strengthen

the alert score when neither modality alone provides sufficient confidence. However, this qualitative case does not establish superiority over single modality systems across all conditions.

(**a**) Scenario A: Sparse crowd (**b**) Scenario B: Dense crowd (**c**) Scenario C: High motion

Figure 6: Qualitative results aligned with test cases, demonstrating system performance across varying crowd densities and motion levels.

Scenario B (Dense Crowd)

This scenario depicts a dense crowd where the dominant emotions are positive, with multiple individuals correctly identified as 'Happy' (Fig. 6b). A naive system might incorrectly assume the scene is safe. However, our framework demonstrates its modality-specific reliability. As clearly indicated by the on-screen alert, it correctly identifies that the high number of tracked faces has surpassed the configured threshold, triggering a specific **"OVERCROWDING ALERT!"**. This showcases the system's ability to provide distinct, interpretable alerts based on different risk factors (density vs. sentiment) and to avoid being misled by positive emotional cues when a separate physical risk is present.

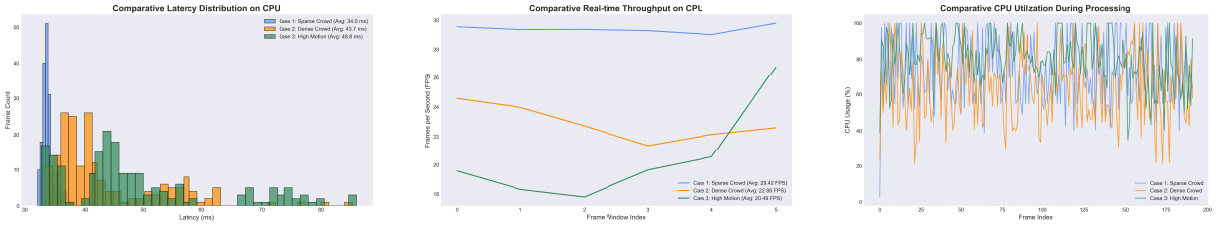
Scenario C (High Motion)

This challenging scenario features a dense, chaotic, and dynamic crowd where tracking and recognition are most stressed. As the output in Fig. 6c demonstrates, the full capabilities of the system are evident. Concurrently, it detects and flags multiple distinct anomalies. The high density triggers an **"OVERCROWDING ALERT!"**, while the DBSCAN module identifies a spatially coherent cluster of individuals exhibiting negative emotions ('Sad' and 'Fear'), highlighted by the colored overlays. The system explicitly reports both the **"Emotion Cluster"** and **"Overcrowding"** alerts, demonstrating the profound benefit of its multifaceted analysis in complex situations and providing operators with a high-confidence, context-rich understanding of a potentially escalating event.

Taken together, these cases demonstrate that our framework generates semantically distinct, operator-aligned alerts for different conditions. In sparse scenes, audio augments vision to detect isolated distress; in dense crowds, specific density alerts are triggered without spurious emotional alarms; and in complex scenes, converging evidence from multiple cues prompts a high-priority response. This behavior, enabled by our late fusion and spatial emotion clustering approach, enhances interpretability and reduces errors, validating the system's practical design.

5.5 Real-Time Performance on CPU

To assess the framework's deployability on commodity hardware [28], we profiled the end-to-end pipeline on a standard multicore CPU across our three representative test cases: **Sparse**, **Dense**, and **High motion**. The results, visualized in Fig. 7, confirm the system's efficiency and predictable scaling, validating its feasibility for practical, real-world application.



(**a**) Latency distributions (**b**) Throughput (FPS) (**c**) CPU usage

Figure 7: Performance of the proposed framework across the three test cases: Sparse crowd, dense crowd, and high motion.

The real-time throughput curves, shown in Fig. 7b, illustrate that the system maintains stable and distinct performance levels for each scenario. The sparse case consistently operates near **29 FPS**, while the more demanding dense and high motion cases stabilize around **22 FPS** and **20 FPS**, respectively, confirming sustained real-time operation. A more granular analysis is provided by the latency distribution histograms in Fig. 7a, which show how the frame processing time shifts to the right with increasing scene complexity. The broader tail in the high motion case, for instance, is consistent with the increased computational overhead from its complex tracking logic. Finally, the CPU utilization traces in Fig. 7c reveal the computational demand of each scenario. While the system is resource intensive, particularly in dense and high motion scenes where usage frequently approaches 100%, it operates effectively without crashing, delivering a stable, operator-facing output without perceptible jitter.

The results, summarized in Table 7, demonstrate the system's efficiency and predictable scaling. In sparse scenes, the pipeline achieves a comfortable real-time throughput of **28.40 FPS** with an average latency of 32.93 ms. As scene complexity increases, performance scales gracefully: throughput decreases moderately to 21.76 FPS in the dense scenario and **19.37 FPS** in the most demanding high-motion case, with latency rising accordingly. Crucially, the system maintains near real-time performance (**≥19 FPS**) across all conditions, validating its feasibility for efficient, **CPU-only deployment** without requiring specialised GPU hardware. Regarding temporal resolution, the asynchronous FER interval of $K_{\text{FER}} = 4$ frames (Table 2) was designed for the nominal frame rate of approximately 28 FPS, yielding an emotion update every ~143 ms. At the lower throughput of 19.37 FPS observed during high-motion scenes, the same 4-frame interval corresponds to an update every ~207 ms, which remains within the typical duration of a sustained facial expression (generally >500 ms [29]) and is therefore still sufficient to capture genuine affective state changes without introducing false negatives. This performance profile confirms that the framework is not only accurate but also practical for real-world application.

Table 7: Performance summary of our proposed framework.

Test Case	Density	Latency (ms)	Throughput (FPS)	Peak CPU (%)
Sparse crowd	Low	32.93	28.40	78
Dense crowd	High	43.40	21.76	92
High motion	Medium	48.78	19.37	85

5.6 Comparison with Existing Methods

To contextualize our contributions, we compare the proposed framework with representative baselines spanning video-only weakly/semi-supervised anomaly detection and audio–visual approaches. Specifically, we include influential video only methods [12–14,16] and the multimodal XD-Violence baseline [15].

Table 8 contrasts representative video-only and audiovisual baselines along deployment-oriented axes: modality, supervision, decision granularity/interpretability, fusion strategy, and CPU real-time feasibility (CPU RT). Video-only methods [12–14,16] produce clip/frame scores without audio cues and are generally not CPU real-time, offering limited operator interpretability.

Table 8: Comparison with existing methods.

Method	Modality	Superv.	Granularity	Fusion	CPU RT
Sultani et al. [12]	V	Weak	Clip/frame score	–	No
RTFM [13]	V	Weak	Clip/frame score	–	No
Georgescu et al. [14]	V	Semi	Frame score	–	No
XD-Violence [15]	A + V	Weak	Clip/segment score	Late/MIL	No
MIST (WSVAD) [16]	V	Weak	Clip/frame score	–	No
Proposed SAFE	A + V	Sup.	**Person-level overlays + fused score**	**Late (weighted)**	**Yes**

Key observations: (i) Video-only SOTA (e.g., [12–14,16]) achieves strong accuracy but lacks audio cues and per-person interpretability; most require a GPU for real-time. (ii) Audio-visual baselines (e.g., XD-Violence [15]) incorporate audio but remain clip-level and rely on heavy backbones. (iii) Our framework provides interpretable, operator-facing overlays (IDs, emotions, DBSCAN clusters), leverages audio for robustness in low light/high motion scenes, and sustains CPU real-time throughput (Table 7), making it practical for on-premise deployment.

6 Conclusion and Future Work

We presented a real-time multimodal framework for crowd anomaly detection that fuses complementary audio and visual cues into interpretable, operator-facing alerts. The visual pipeline combines an SSD face detector, Hungarian tracking, FER-based sentiment analysis, DBSCAN clustering of negative emotions, and an overcrowding signal. The audio pipeline extracts MFCCs and uses a lightweight 1D CNN, while a weighted late fusion scheme produces unified alerts including overcrowding, emotion cluster, and fused signals. Experiments demonstrate accurate component performance (audio AUC 0.99 on UrbanSound8K, 0.82 on ESC-50), consistent behavior across sparse, dense, and high motion scenarios. CPU near real-time throughput of 19.37–28.40 FPS across scenarios with predictable scaling. Table 8 positions SAFE against representative baselines along deployment-oriented axes (modality, supervision, granularity, fusion, CPU real-time feasibility). This is a conceptual comparison rather than a benchmark under unified datasets, metrics, or hardware, and a controlled head-to-head accuracy comparison remains future work. The system nonetheless offers materially greater interpretability and CPU-only deployability than the compared baselines.

Despite its effectiveness, our proposed framework currently depends on visible facial regions for reliable FER, which can reduce accuracy under severe occlusion or poor lighting. The audio pipeline may experience domain shifts in highly reverberant or acoustically cluttered scenes, affecting anomaly discrimination [30]. Moreover, the fusion process uses fixed thresholds that may require scene-specific calibration when operating across diverse environmental or cultural contexts. Similarly, the equal fusion

weights ($w_{\text{audio}} = w_{\text{visual}} = 0.5$) represent a balanced baseline that may be sub-optimal in environments where one modality is severely degraded—for instance, heavy acoustic noise would inflate S_{audio} spuriously, while dense occlusion would suppress S_{visual} below its true level. These limitations highlight practical considerations for future deployments and opportunities for refinement.

Upcoming research will focus on adaptive fusion mechanisms (e.g., attention or confidence-gating) that dynamically reweight modalities based on estimated per-frame signal quality—assigning lower weight to the audio stream when acoustic SNR is poor and reducing visual weight under heavy occlusion or low light—as well as end-to-end multimodal training with self-supervised or cross-attentive learning for enhanced robustness. Broader person cues, such as body pose, gesture, and re-identification, will be integrated to mitigate FER dependence and improve spatial continuity. We also plan to scale multi-camera and edge-cloud coordination using model compression techniques (quantisation, pruning, distillation) to maintain low latency at scale. Finally, efforts will focus on continuous active learning to handle domain shifts and strengthen reliability, along with ethical safeguards such as uncertainty calibration, human-in-the-loop verification, and privacy-by-design analytics to support responsible deployment.

Acknowledgement: The authors thank the Central University of Rajasthan for providing the necessary facilities.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization, Akrisht Singh; methodology, Akrisht Singh; software, Akrisht Singh; validation, Ravi Saharan; writing—original draft preparation, Akrisht Singh; writing—review and editing, Prakash Choudhary; supervision, Ravi Saharan. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used (FER-2013, UrbanSound8K, ESC-50) are publicly available.

Ethics Approval: Not applicable for this study as it utilizes publicly available datasets.

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

References

1. Illiyas FT, Mani SK, Pradeepkumar AP, Mohan K. Human stampedes during religious festivals: a comparative review of mass gathering emergencies in India. Int J Disaster Risk Reduct. 2013;5:10–8. doi:10.1016/j.ijdrr.2013.09.003.
2. Mahadevan V, Li W, Bhalodia V, Vasconcelos N. Anomaly detection in crowded scenes. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2010 Jun 13–18; San Francisco, CA, USA. p. 1975–81. doi:10.1109/CVPR.2010.5539872.
3. Li W, Mahadevan V, Vasconcelos N. Anomaly detection and localization in crowded scenes. IEEE Trans Pattern Anal Mach Intell. 2014;36(1):18–32. doi:10.1109/TPAMI.2013.111.
4. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot multibox detector. In: Computer Vision–ECCV 2016. Cham, Switzerland: Springer; 2016. p. 21–37. doi:10.1007/978-3-319-46448-0_2.
5. Bewley A, Ge Z, Ott L, Ramos F, Upcroft B. Simple online and realtime tracking. In: Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP); 2016 Sep 25–28; Phoenix, AZ, USA. p. 3464–8. doi:10.1109/ICIP.2016.7533003.
6. Li S, Deng W. Deep facial expression recognition: a survey. IEEE Trans Affect Comput. 2022;13(3):1195–215. doi:10.1109/TAFFC.2020.2981446.
7. Serengil SI, Ozpinar A. HyperExtended LightFace: a facial attribute analysis framework. In: Proceedings of the 2021 International Conference on Engineering and Emerging Technologies (ICEET); 2021 Oct 27–28; Istanbul, Turkey. p. 1–4. doi:10.1109/ICEET53442.2021.9659697.

8. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD); 1996 Aug 2–4; Portland, OR, USA. p. 226–31.
9. Chachada S, Kuo CCJ. Environmental sound recognition: a survey. APSIPA Trans Signal Inf Process. 2014;3(1):e14. doi:10.1017/ATSIP.2014.12.
10. Hershey S, Chaudhuri S, Ellis DPW, Gemmeke JF, Jansen A, Moore RC, et al. CNN architectures for large-scale audio classification. In: Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2017 Mar 5–9; New Orleans, LA, USA. p. 131–5. doi:10.1109/ICASSP.2017.7952132.
11. Gajšek R, Štruc V, Dobrišek S, žibert J, Mihelič F, Pavešić N. Combining audio and video for detection of spontaneous emotions. In: Biometric ID management and multimodal communication. Berlin/Heidelberg, Germany: Springer; 2009. p. 114–21. doi:10.1007/978-3-642-04391-8_15.
12. Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 6479–88. doi:10.1109/CVPR.2018.00678.
13. Tian Y, Pang G, Chen Y, Singh R, Verjans JW, Carneiro G. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 4955–66. doi:10.1109/ICCV48922.2021.00493.
14. Georgescu MI, Bărbălău A, Ionescu RT, Khan FS, Popescu M, Shah M. Anomaly detection in video via self-supervised and multi-task learning. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 12737–47. doi:10.1109/CVPR46437.2021.01255.
15. Wu P, Liu J, Shi Y, Sun Y, Shao F, Wu Z, et al. Not only look, but also listen: learning multimodal violence detection under weak supervision. In: Computer Vision–ECCV 2020. Vol. 12375. Cham, Switzerland: Springer; 2020. p. 322–39. doi:10.1007/978-3-030-58577-8_20.
16. Feng JC, Hong FT, Zheng WS. MIST: multiple instance self-training framework for video anomaly detection. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 14004–13. doi:10.1109/CVPR46437.2021.01379.
17. Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, et al. Challenges in representation learning: a report on three machine learning contests. Neural Netw. 2015;64(1):59–63. doi:10.1016/j.neunet.2014.09.005.
18. Salamon J, Jacoby C, Bello JP. A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM International Conference on Multimedia; 2014 Nov 3–7; Orlando, FL, USA. New York, NY, USA: Association for Computing Machinery; 2014. p. 1041–4. doi:10.1145/2647868.2655045.
19. Piczak KJ. ESC: dataset for environmental sound classification. In: Proceedings of the 23rd ACM International Conference on Multimedia; 2015 Oct 26–30; Brisbane, Australia. New York, NY, USA: Association for Computing Machinery; 2015. p. 1015–8. doi:10.1145/2733373.2806390.
20. McFee B, Raffel C, Liang D, Ellis DPW, McVicar M, Battenberg E, et al. Librosa: audio and music signal analysis in Python. In: Proceedings of the 14th Python in Science Conference; 2015 Jul 6–12; Austin, TX, USA. p. 18–25. doi:10.25080/Majora-7b98e3ed-003.
21. Joshi K, Patel N. Supervised deep learning approaches for anomaly detection and recognition in crowd scenes. ELCVIA Electron Lett Comput Vis Image Anal. 2025;24(1):31–50. doi:10.5565/rev/elcvia.1631.
22. Sindagi VA, Patel VM. A survey of recent advances in CNN-based single image crowd counting and density estimation. Pattern Recognit Lett. 2018;107(3):3–16. doi:10.1016/j.patrec.2017.07.007.
23. Rehman AU, Ullah HS, Farooq H, Khan MS, Mahmood T, Khan HOA. Multi-modal anomaly detection by using audio and visual cues. IEEE Access. 2021;9:30587–603. doi:10.1109/ACCESS.2021.3059519.
24. Wang Y, Zhao Y, Huo Y, Lu Y. Multimodal anomaly detection in complex environments using video and audio fusion. Sci Rep. 2025;15(1):16291. doi:10.1038/s41598-025-01146-4.
25. Liang C, Chen Q, Li Q, Wang Q, Zhao K, Tu J, et al. HADNet: a novel lightweight approach for abnormal sound detection on highway based on 1D convolutional neural network and multi-head self-attention mechanism. Electronics. 2024;13(21):4229. doi:10.3390/electronics13214229.

26. Kalyta O, Barmak O, Radiuk P, Krak I. Facial emotion recognition for photo and video surveillance based on machine learning and visual analytics. Appl Sci. 2023;13(17):9890. doi:10.3390/app13179890.
27. Xu P, Peng S, Luo Y, Gong G. Facial expression recognition: a meta-analytic review of theoretical models and neuroimaging evidence. Neurosci Biobehav Rev. 2021;127(3):820–36. doi:10.1016/j.neubiorev.2021.05.023.
28. Zhao Z, Wen Y, Yang S, Ning L, Liu Y, Gao J. Real-time crowd counting for embedded systems with lightweight architecture. Vicinagearth. 2025;2(1):13. doi:10.1007/s44336-025-00025-w.
29. Yan WJ, Wu Q, Liang J, Chen YH, Fu X. How fast are the leaked facial expressions: the duration of micro-expressions. J Nonverbal Behav. 2013;37(4):217–30. doi:10.1007/s10919-013-0159-8.
30. Rezaee K, Rezakhani SM, Khosravi MR, Moghimi MK. A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Pers Ubiquitous Comput. 2024;28(1):135–51. doi:10.1007/s00779-021-01586-5.