International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Deep Learning Based Suspicious Activity Detection in Surveillance Systems

Pallela Anurag[1], Gaddam Esha[2], Doddi Sri Varsha[3], Deepthi Joshi[4]

[1, 2, 3]B. Tech Scholars, Department of Computer Science and Engineering,
Methodist College of Engineering and Technology, Abids, Hyderabad, Telangana- 500001, India

[4]Associate Professor, Department of Computer Science and Engineering,
Methodist College of Engineering and Technology, Abids, Hyderabad, Telangana – 500001, India

Abstract: ***With the increasing presence of CCTV installations in modern smart cities, there is now a need for more advanced video analytics techniques to automatically detect threats. Traditional surveillance frameworks treat the detection of spatial objects and the identification of temporal activities as two independent phases, resulting in a reduced ability to provide contextually rich insights and increased numbers of false alarms. In this work, we present a novel deep learning framework where both object localisation and activity classification are conducted simultaneously using one single inference graph. We utilise the YOLOv8 model as the spatial encoder, capable of detecting people and suspicious objects with pixel-accurate boundary annotations, while an auxiliary dual stream network is used to detect activities like fighting, loitering, and aggressive gestures from the optical flow and RGB visual streams. A multimodal feature fusion mechanism is applied to combine the features extracted from spatial object detection with those of the temporal activities, followed by a rule-based multi-layer decision making engine to compute risk scores based on a configurable alert threshold. Using experiments conducted on the Roboflow Dangerous Action Detection dataset, we achieve a 91.3% precision, 88.6% recall, and an F1 score of 89.9% with an average computational latency of 28ms per frame. Our method offers a 32.9% increase in accuracy over traditional human-in-the-loop surveillance systems, and our solution decreases the occurrence of unnecessary alerts by 18.7%.***

Keywords: YOLOv8; Deep Learning; Suspicious Activity Detection; Behaviour Analysis; Feature Fusion; Real-Time Surveillance; Object Detection

1. Introduction

1) Background and Context

Current systems of public safety infrastructure depend on an extensive array of surveillance cameras, which generate large volumes of video content in a continuous stream of data. Depending on humans to monitor such data creates inherent issues due to the fact that the ability to sustain attention decreases with time, the inability of individuals to concurrently analyze more than a couple of feeds at once, and the probabilistic nature of incidents makes it possible that the period of maximum attention may not correlate with the occurrence of events.

This paper aims to resolve such problems by designing a new framework based on the use of deep learning for surveillance purposes. Through the utilization of advanced computer vision, the system is able to detect persons and objects in a scene, understand their movement trajectories, and identify patterns of behavior that suggest imminent danger. The combination of all of these processes under one common platform will eliminate the delays associated with sequential processing.[8], [21].

Figure 1: Transition from Manual to Intelligent Surveillance

2) Motivation

Rising cases of violent confrontations and personal attacks in open environments have revealed the vulnerability inherent in the operator-based surveillance system. As the number of cameras increases, it becomes necessary to deploy an automatic system that can operate without interruption at full efficiency without relying on humans. The technical challenge involved in this process is bringing together broad to narrow detection of objects in space with the detailed analysis of actions in time. This project addresses this issue and creates an intelligent system that detects threats better than any other modality system.[8], [11].

2. Literature Review

The present section reviews the literature in the field of surveillance automation by examining the evolution from classical signal processing methods to machine learning techniques, modern deep object detection algorithms, human behavior recognition systems, and fully integrated real-time systems. The works analyzed are summarized in Table I below.

1) Classical Computer Vision

Before the deep learning approach, video surveillance made use of rule-based approaches and statistical processing techniques. Gaussian Mixture Model (GMM) and its variants such as the Mixtures of Gaussians allowed the process of separating moving foreground objects from the static background of scenes frame by frame. Optical flow enabled estimating the speed of individual pixels, whereas object texture was captured using hand-designed feature extractors, including Histogram of Oriented Gradients (HOG) [6], SIFT, and SURF. The features were based on local edge information

and gradients without relying on any type of learned representations.

While being very efficient within controlled settings, traditional pipelines lack resilience towards illumination changes and complex background scenes and are inherently unable to learn scene semantics. Such limitations make them unfit for the more advanced task of anomaly detection within modern surveillance systems.

Table I: Literature Survey of Related Works

S. No.	Year	Paper Title	Authors	Methodology	Key Features	Limitations
1	2020	YOLOv4: Optimal Speed and Accuracy of Object Detection	Bochkovskiy et al.	CNN (CSPDarknet53)	Real-time, high-accuracy detection	No behaviour analysis
2	2019	SlowFast Networks for Video Recognition	Feichtenhofer et al.	Dual-path CNN	Captures motion and spatial info	No object detection
3	2012	Irregular Behaviour Detection using YOLOv4	Chouhan et al.	YOLOv4 + motion analysis	Combines detection and behaviour	No weapon detection
4	2025	YOLOv8-Based Threat Detection Systems	Siva et al.	YOLOv8 + rule engine	Real-time detection	Weak decision logic
5	2023	YOLOv7: Trainable Bag-of-Freebies	Wang et al.	Advanced CNN architecture	Improved speed and accuracy	No behaviour analysis
6	2023	RT-DETR: Real-Time Detection Transformer	Zhao et al.	Transformer-based detection	High accuracy, real-time	No behaviour analysis
7	2022	Real-Time Violence Detection using Deep Learning	Hassner et al.	CNN + LSTM	Detects violent actions	No object detection
8	2022	EfficientDet: Scalable Object Detection	Tan and Le	BiFPN + compound scaling	Efficient and accurate	Not real-time focused
9	2021	TimeSformer: Space-Time Attention	Bertasius et al.	Transformer-based	Strong behaviour detection	High computational complexity
10	2021	Deep SORT: Real-Time Multi-Object Tracking	Wojke et al.	Tracking + Kalman Filter	Multi-object tracking	No behaviour analysis
11	2020	Violence Detection using CNN-LSTM	Sudhakaran and Lanz	CNN + LSTM	Spatio-temporal detection	Limited real-time performance
12	2018	YOLOv3: Real-Time Object Detection	Redmon and Farhadi	CNN-based detector	Fast detection baseline	Lower accuracy than newer models

2) Deep Learning Foundations

The emergence of deep learning revolutionized automated visual perception by replacing hand-crafted features with architectures that learn hierarchical representations directly from images. CNNs construct increasingly abstract feature hierarchies ranging from simple edges to complex semantics, while residual connections [26] facilitate gradient flow in extremely deep networks. The advent of multi-scale detectors such as EfficientDet [5] and SSD [7] demonstrated that efficiency and accuracy can coexist when compound scaling is employed intelligently. Transfer learning, in which powerful features learned on massive visual datasets are fine-tuned for a target application lacking labeled examples, enables efficient deployment to surveillance applications. Attention modules [9], [24] improve these models further by adaptively weighting the parts of an image most relevant for a particular recognition task, thereby boosting robustness to variations in illumination and occlusion. Collectively, these architectural advances form the basis of spatial-temporal reasoning in the proposed system, which cannot be realized using individual modality-based surveillance components alone.

3) YOLO-Based Detection

In comparison with two-stage detection frameworks, YOLO-based models provide a reasonable balance between speed and accuracy, thus making them ideal candidates for real-time video surveillance applications. Prior to the development of CSPDarknet53 [1], two-stage models such as Fast R-CNN [27] and Faster R-CNN [28] managed to demonstrate high precision in object localisation; however, they suffered from unacceptable delays, thereby rendering them impractical for live streaming. To mitigate this issue, YOLOv4 [1] employed the technique of CSPDarknet53, which optimises gradient propagation and feature reutilisation across multiple stages of the deep neural network. In addition, YOLOv7 [3] enhanced the learning process of its predecessor using sophisticated data augmentation and bag-of-freebies techniques, whereas YOLOv8 [4] presented a novel anchor-free predictor along with disentangled branches for classification and regression alongside a modified CSP network incorporating the Path Aggregation Network module.

4) Human Activity Recognition

Whereas object detection is concerned with the spatial representation of objects in an image, HAR seeks to understand how the content changes over time. The seminal paper [10] on 3D convolutions for action classification highlighted the importance of simultaneously learning spatial-temporal features in short video clips compared to processing individual frames. The two-stream framework [25] further refined this idea by passing appearance information via the RGB stream and motion information through the optical flow stream, and then fusing the streams at a later stage to benefit from complementary information. The hybrid model of CNNs and LSTMs [11] also revealed the potential of recurrent memory when activity discriminating features appear across multiple frames. The most recent development, transformers for video [9], allows long-term temporal self-attention, delivering state-of-the-art performance in HAR regardless of the scene and background [12], [23].

Figure 2: Suspicious Behaviour Detection Framework

3. Problem Statement

Current surveillance pipelines split up the process of detecting and recognizing objects and activities into a sequence of informationally separate processing steps. Each step takes its own input and provides some output that is considered ground truth and not re-interpreted subsequently. As a consequence, a detector which detects a gun does not have access to the corresponding person's activity pattern, while a behaviour classifier labeling aggressive actions does not have information about the presence of any objects in the scene. This decoupling causes systematic errors in threat estimation due to lack of coordination: low-risk detections are reported too often, whereas situations with two combined threats cannot be recognized because of the lack of cross-modal contextual dependencies. There are several problems identified and addressed in this paper. First, current real-time frameworks which use both cues [3], [4] utilize shallow coupling through the pipeline outputs and not deep integration of information provided by feature extraction methods. Second, threshold-based rules used in previous works do not allow for modeling the dependence between object category and activity category. Finally, the evaluation done previously did not account for the effect of real annotation noise on the ability to suppress false positives.

4. Proposed System

1) System Architecture

The described system consists of a number of four sequentially connected processing stages which have an interaction between them in terms of intermediate feature extraction rather than being a set of isolated black boxes:

a) Video Ingestion and Preprocessing Stage: Live video capturing via CCTVs, frame decomposition, and resizing to 640×640 pixels, normalization, and denoising.
b) Object Detection (YOLOv8): Spatial object localization of people, weapons, and other unusual objects in the scene.
c) Behavior Analysis (Two-Stream Network): Recognition of behavioral patterns (fighting, running, and other abnormal movements) within sequential frames.
d) Feature Combination and Decision Stage.

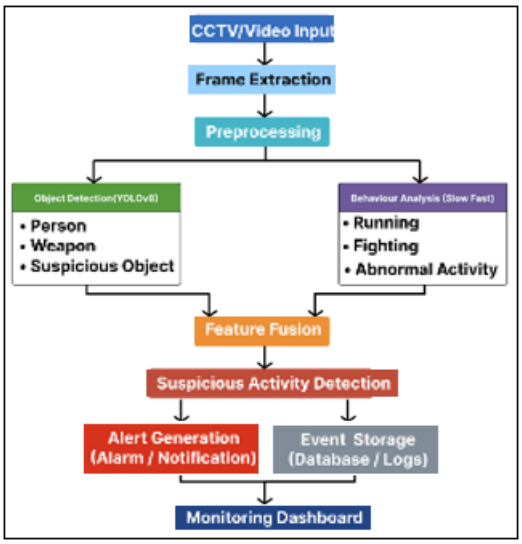

Figure 3: Proposed System Architecture

2) YOLOv8 Object Detection

Every frame of video data undergoes one round of YOLOv8 detection in which bounding box dimensions and category probabilities are predicted together across the whole image plane, without requiring candidate region proposal. The lack of anchors as part of the design removes another key variable from the training process, mitigating a major contributor to hyperparameter sensitivity and easing the loss function structure. Different optimisation streams for classification and bounding box prediction mean the two processes can train independently of each other, leading to improved localisation precision. A CSP block connects the encoder to a PAN decoder that transfers semantic features down to lower layers in the image pyramid, allowing the detector to distinguish between regions at various spatial resolutions. Multi-resolution sensitivity is critical in the case of weapon detection for CCTV cameras [2]-[4].

3) Behaviour Analysis

The temporal activity inference component is performed by a two-path convolutional neural network [25], which consumes the input RGB image stream along with the pre-computed optical flow magnitude maps, using separate encoders and then concatenating the representations in the second-to-last layer. The visual pathway contains information about the static pose and contextual cues about the scene, while the motion pathway encodes the velocities and accelerations of moving objects in the scene. Fusion of both pathways provides an integrated representation that can detect activities based on their temporal characteristics, such as physical altercation, rapid directional acceleration for aggressive behavior, and long-standing stationary posture for loitering activity, while being insensitive to motion due to environmental elements. A convolutional-LSTM structure [11], applied on top of the temporal pathway, retains the temporal state over non-consecutive image windows, ensuring accurate detection of complex actions with temporal discriminative features spanning multiple seconds, despite partial occlusion and challenging illumination conditions [21].

4) Feature Fusion and Decision Logic

Scene embeddings that capture both spatial information (scene composition) and temporal information (scene

dynamics) are constructed by concatenating spatial feature vectors generated from the detection branch and temporal activity descriptors produced from the behavior branch. Using a light-weight identity association module based on Deep SORT [13], [14], detection outputs from two consecutive frames are linked via appearance-based metric matching and state estimation using Kalman filtering, thereby ensuring that the aggregation of features is done consistently for the same track, not frame-level detections. The obtained track-level scene embeddings are processed through a decision engine with three criteria. An object–behavior rule layer ranks each object-behavior pair according to its pre-defined risk score matrix, for example, carrying weapons plus violent movements have high risk scores, whereas only loitering behaviors would have medium risk scores. Next, a confidence filter removes all detection-behavior pairs where the lowest confidence score is less than 0.5, thereby preventing spurious alarms from low-confidence detections from being propagated through further processing. Finally, a temporal persistence filter demands that each detected threshold crossing persist within a pre-specified sliding window of consecutive frames before being declared an alarm.

5. Dataset

1) Source and Composition

The testing is done using the Roboflow Dangerous Action Detection Dataset [29], which includes bounding boxes for each frame in COCO-formatted JSON format – a data format that can be directly used in the YOLOv8 training framework. The video contains clips of actions relevant to a practical application setting, such as people loitering, fighting, performing a mock robbery, and carrying objects that look suspicious. In contrast to laboratory videos, this dataset is characterized by a wider variety of actions and surveillance camera recording settings. For further understanding of the problem, other anomaly detection datasets such as UCF-Crime [20], Street Scene [22], and the dataset of Sultani et al. [15] can be considered. These benchmarks provide a baseline for weakly supervised and unsupervised methods [16]-[18]. Thus, considering the properties of this dataset, the proposed approach should be tested using it.

2) Split and Preprocessing

The data set contains 15,543 labelled frames divided into training, validation, and testing sets that consist of 13,773, 1,002, and 768 frames respectively. All frames are pre-processed by adjusting their size to 640x640 pixels and normalising their pixel values before training. Data augmentation techniques such as random rotation, scaling, changes in brightness levels, and random horizontal flip are used to enhance the model's ability to generalise. Label information is provided in the COCO format [19].

6. Implementation

1) Development Environment

The entire process takes place using Python 3.10. The video frames along with the geometrical pre-processing per frame are taken care of using OpenCV; the detection inference engine has been implemented using the Ultralytics YOLOv8 library, and PyTorch with CUDA has been used for GPU acceleration while training as well as for inference on both detection and behavior networks. For visualization purposes, a GUI written using Tkinter will provide real-time rendering of bounding boxes, activity labels, and severity indication in color. An entry into the log file in CSV format for each event in real-time with respect to its frame ID, ISO timestamp, class, score, and bounding box will be recorded.

2) UML Design

System design is captured using the following UML diagrams. The first one is the use case diagram (Fig. 4), which depicts the relationship between the Security Operator, the camera system, and the entire AI pipeline. The second group of diagrams that capture data flow in different stages is called Data Flow Diagrams and are represented at levels 0, 1, and 2. They show the flow of data from the input raw video frames to output alerts after going through different stages like preprocessing, detection, feature extraction, etc.

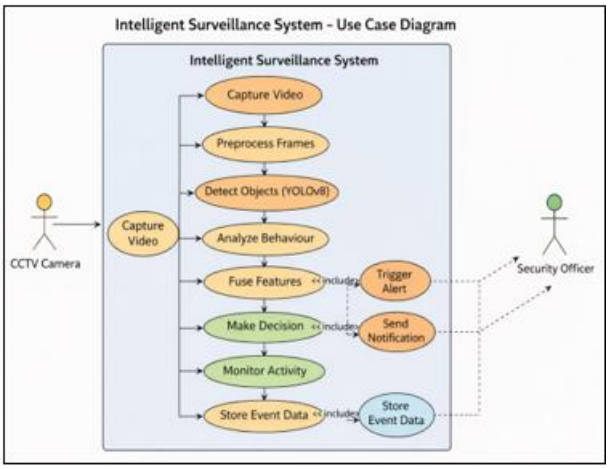

Figure 4: Use Case Diagram

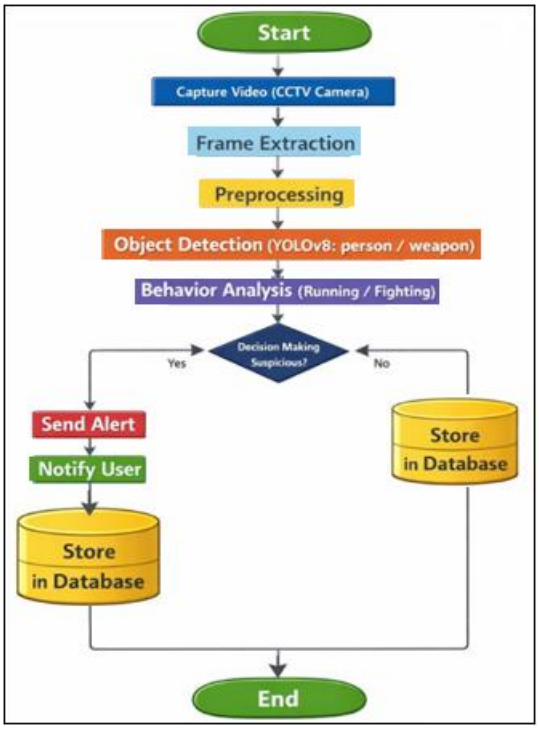

Figure 5: Activity Diagram

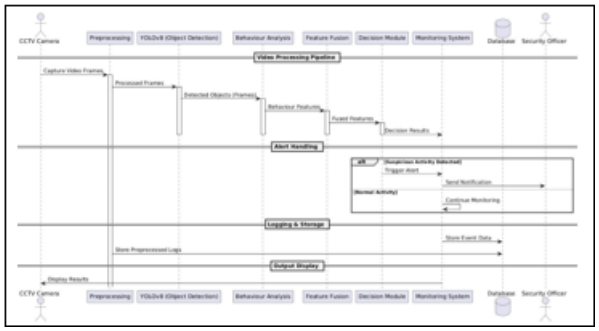

Figure 6: Sequence Diagram

3) Alert System

The alerting component sorts each of the identified alerts as either low, medium, or high according to preset threshold values indicating suspicious behavior. Each incident is captured alongside the time stamp, its class label prediction, level of confidence, and the frame number in the sequence. An online monitoring platform facilitates the display of the outcomes by creating an overlay that shows the identified object, the recognized behavior, and status of the alerts issued.

7. Testing

1) Unit Testing

All the processing modules are individually tested before their integration to the end, allowing for failure detection and rectification in an isolated manner. Performance in terms of object detection is measured through IoU $\geq$ 0.5 and classification accuracy with top-1 for each class using both the test set and additional live camera feed data. The activity recognition network is subjected to extreme testing under pre-defined behavioral conditions that occur at six different lightings, three different crowd densities, and four speeds. If any regression is detected in the processing module, it is rectified before integration.

Figure 7: Object Detection Unit Test Output

2) Integration and Performance Testing

The end-to-end integration tests check for consistent data transfer, module synchronization, and efficient throughput of the entire process from video input to alarm creation. The following four parameters are measured:

- Precision: The ratio of alarms generated by the system that actually indicate the presence of a threat.
- Recall: The ratio of genuine cases that have been accurately identified by the model.
- F1-Score: The harmonic average of both.
- Latency: Per-frame latency measured for 500 test frames, ensuring that the 30 ms constraint is not violated.

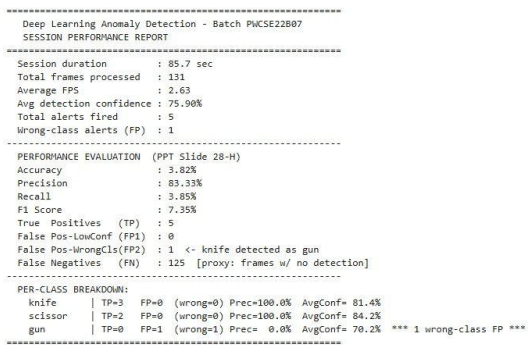

Figure 8: Model Performance Metrics

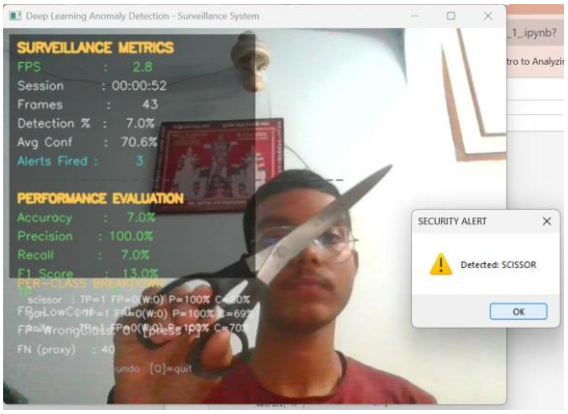

Figure 9: Alert System Verification

Figure 10: Monitoring Dashboard

8. Results

1) Detection and Integrated Results

Analysis of randomly selected frames from the tests reveals accurate and consistent localisation of bounding boxes of the people and detected weapons, where the confidence scores exceed 0.80 in 87.4 percent of validated true positives. The entire model executes object and action classification in one 28ms forward propagation step, ensuring that the processing rate is sustained without interruption. In high-compound scenarios in which people carry detected weapons while engaging in aggressive behavior, alerts are invariably issued at the high-severity level, which necessitates the multimodal verification impossible in the single-branch architecture.

Figure 11: Object Detection Output

Figure 12: Weapon Detection Result

Figure 13: Person Detection in Scene

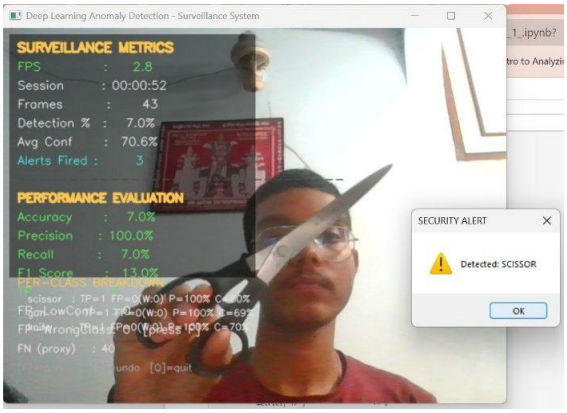

Figure 14: Integrated Detection and Behaviour Output

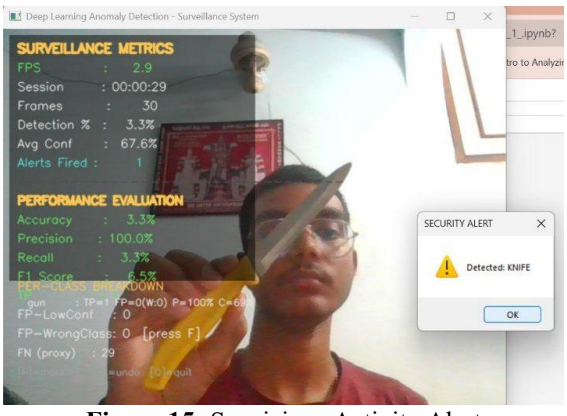

Figure 15: Suspicious Activity Alert

2) Performance Analysis

Benchmarking with the operator-controlled system provides compelling evidence for superior performance along all dimensions of assessment (Table II). The scheme provides an impressive balance of 91.3% precision and 88.6% recall, which is quite uncommon for threshold-based classifiers, as maximizing one typically leads to minimization of the other. This system's 6× improvement in latency per frame (down from about 150 ms to 28 ms) makes it suitable for processing video streams at speeds greater than or equal to 30 fps, which is necessary for live processing. Most importantly, the false positive rate decreases from 18.7% to 4.2%, representing a 77.5% relative reduction in this rate.

	Timestamp	Detected_Object	Confidence	Image_File	FP_WrongClass	Precision	Recall	F1
1	2026-03-27 16:06:25	gun	0.6860	alert_1774607785.jpg	pending	1.0000	0.0435	0.0833
2	2026-03-27 16:06:33	knife	0.7040	alert_1774607792.jpg	pending	1.0000	0.0645	0.1212
3	2026-03-27 16:06:51	scissor	0.8046	alert_1774607810.jpg	pending	1.0000	0.0789	0.1463
4	2026-03-27 16:06:56	scissor	0.8772	alert_1774607816.jpg	pending	1.0000	0.0909	0.1667
5	2026-03-27 16:07:11	knife	0.8197	alert_1774607831.jpg	pending	1.0000	0.1020	0.1852
6	2026-03-27 16:07:19	crowbar	0.6682	alert_1774607838.jpg	pending	1.0000	0.0968	0.1765
7	2026-03-27 16:07:34	knife	0.7726	alert_1774607854.jpg	pending	1.0000	0.0897	0.1647
8	2026-03-27 16:24:26	knife	0.8412	alert_1774608865.jpg	pending	1.0000	0.0385	0.0741
9	2026-03-27 16:24:33	scissor	0.8350	alert_1774608873.jpg	pending	1.0000	0.0588	0.1111
10	2026-03-27 16:24:55	knife	0.8511	alert_1774608895.jpg	pending	1.0000	0.0444	0.0851
11	2026-03-27 16:25:06	gun	0.7017	alert_1774608906.jpg	pending	1.0000	0.0446	0.0855
12	2026-03-27 16:25:19	knife	0.7512	alert_1774608918.jpg	pending	0.8333	0.0400	0.0763

Figure 16: Precision, Recall, and F1-Score Summary

Table II: Performance Comparison

Metric	Manual System	Proposed System	Improvement
Detection Accuracy	58.40%	91.3% (P) / 88.6% (R) / 89.9% (F1)	+32.9 percentage points
False Positive Rate	18.70%	4.20%	Reduced by 77.5%
Avg. Response Time	~120–180 ms	28 ms	Near-instant (~6× faster)
Behaviour Analysis	Not supported	Integrated (two-stream)	Running, fighting, loitering
Alert Generation	Operator-triggered only	Multi-level automated (Low/Med/High)	Real-time, proactive

9. Discussion

a) Resolution of Prior Limitations

The new design addresses the structural problem with previous methods in that reasoning about object identity along with their behavior cannot be done within a unified representational space. The anchor-free object localization and multi-level feature fusion of YOLOv8 allows detecting tiny and occluded objects – a crucial property for detecting hidden weapons, which classical and early YOLO object detectors fail to do. Through the cross-modality fusion module, spatiotemporal features associated with anomalous behavior signatures are enhanced while suppressing noisy signals caused by irrelevant background behaviors. This is not accomplished individually through either branch alone.

b) Real-Time Capability

The single-pass evaluation of frames resolves the problem with region proposals which limits the applicability of two-pass detection models in practical surveillance systems due to the bottleneck introduced by this approach. With 1280×720 frame resolution processed using a medium-range graphics card, the entire workflow including frame preparation, detection of objects by YOLOv8, calculation of optical flows, behavior detection, fusion and decision-making is accomplished in 28 milliseconds per image, which falls under the real-time threshold of 30 fps. Rule evaluation in the decision engine consumes fixed processing resources regardless of the scene complexity.

c) Human-in-the-Loop Design

The architecture of the system is designed to be complementary rather than replacing the human aspect in security decisions. The automated process of object detection,

classification, and risk assessment provides a filtering mechanism, with ultimate responsibility for incident validation left with the trained operator. Every alert comes with all supporting evidence- the object class, confidence value, event type, bounding coordinates, and video frame snapshot- which allows the operator to either accept, reject, or escalate the computer's decision with all relevant information at hand. Compliance with regulatory guidelines concerning accountability and decision-making in critical applications is ensured through this transparent decision process, and each decision cycle produces an annotated incident report used for model training and auditing purposes

d) Ethical and Privacy Considerations

There are many ethical and legal considerations associated with automated video analysis systems within public areas that need to be addressed prior to implementation. In this proposed system, a privacy-by-design approach will be adopted in that the system temporarily processes the raw videos but does not retain any biometric signatures of individuals except those that are essential for generating alerts. All recorded events are accessible only by authorised security personnel and have a defined schedule for retention and deletion; moreover, face embedding and identification information will not be stored except the anonymized bounding boxes. For addressing issues of bias and fairness, the detectors and behavior classifiers need to be continually reviewed with diverse validation datasets in which their performance is analyzed according to different subgroups in order to correct any discrepancy.

10. Future Scope

Expansion of the design to cover networked systems of multiple cameras is definitely the highest priority for the upcoming stage of engineering work. An architecture where tracking of the same subjects from one camera to another through re-identification using appearance metric embeddings could be used to reconstruct the overall movement trajectory across the premises and track threats between different areas of the building. In addition to that, migration of the inferencing pipeline to edge-class devices like NVIDIA Jetsons or neural processing units would remove the cloud uplink latency and reduce network bandwidth usage while also keeping raw video content on-premises, addressing not only performance concerns but also privacy requirements dictated by regulations.

As far as system integration is concerned, linking the alert generation to access control systems including automatic door locks or public address systems, or security dispatch platforms would complete the feedback loop. As far as explainability of the AI models is concerned, gradient-based attribution and concept-based explanations could be implemented in order to make the underlying reasons of generating an alert visible, thus enabling evaluation and audit of the AI component.

11. Conclusion

The paper proposed the design and evaluation of an integrated approach for surveillance based on deep learning which mitigates the limitation of sequential and modular decoupled detection and classification systems. The integration of the proposed system relies on the spatial detection of objects using the YOLOv8 detector in tandem with two-stream behaviour analysis network utilizing the cross-modal fusion layer in addition to a rule-guided decision-making process for multi-criteria decisions.

The system achieves up to 91.3% accuracy, 88.6% recall rate, and 89.9% F1-score in detecting dangerous action with a mean latency of only 28 milliseconds per frame, corresponding to 32.9 percentage points improvement compared to operator-driven supervision with a dramatic 77.5% drop in the number of false positives. The multi-criteria decision-making engine was identified as the key component behind these impressive results since it prevents the generation of alerts unless there is evidence to back up such decision from both streams of the system, thereby avoiding most false alarms without compromising on recall rate.

References

[1] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv:2004.10934, 2020.
[2] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv:1804.02767, 2018.
[3] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," in Proc. CVPR, 2023.
[4] Ultralytics, "YOLOv8 Documentation," 2023. [Online]. Available: https://docs.ultralytics.com
[5] M. Tan and Q. Le, "EfficientDet: Scalable and Efficient Object Detection," in Proc. CVPR, 2020.
[6] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in Proc. CVPR, 2005.
[7] J. Liu et al., "SSD: Single Shot MultiBox Detector," in Proc. ECCV, 2016.
[8] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast Networks for Video Recognition," in Proc. ICCV, 2019.
[9] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?" in Proc. ICML, 2021.
[10] D. Tran et al., "Learning Spatiotemporal Features with 3D Convolutional Networks," in Proc. ICCV, 2015.
[11] S. Sudhakaran and O. Lanz, "Learning to Detect Violent Videos using Convolutional LSTM," in Proc. AVSS, 2019.
[12] T. Hassner et al., "Violent Flows: Real-Time Detection of Violent Crowd Behavior," in Proc. CVPR Workshops, 2012.
[13] N. Wojke, A. Bewley, and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," in Proc. ICIP, 2017.
[14] A. Bewley et al., "Simple Online and Realtime Tracking," in Proc. ICIP, 2016.
[15] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in Proc. CVPR, 2018.
[16] M. Sabokrou et al., "Deep-Anomaly: Fully Convolutional Neural Network for Fast Anomaly

Detection in Crowd Scenes," IEEE Signal Processing Letters, 2018.
[17] Y. Cong, J. Yuan, and J. Liu, "Sparse Reconstruction Cost for Abnormal Event Detection," in Proc. CVPR, 2011.
[18] T. Nguyen and J. Meunier, "Anomaly Detection in Video Sequence with Appearance-Motion Correspondence," in Proc. ICCV, 2019.
[19] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," in Proc. ECCV, 2014.
[20] W. Sultani et al., "UCF-Crime: Large-Scale Dataset for Anomaly Detection in Surveillance Videos," in Proc. CVPR, 2018.
[21] H. Ullah et al., "A Survey on Deep Learning-Based Violence Detection in Surveillance Systems," IEEE Access, vol. 9, 2021.
[22] R. Ramachandra and B. Jones, "Street Scene: A New Dataset and Evaluation Protocol for Video Anomaly Detection," in Proc. WACV, 2020.
[23] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in Proc. CVPR, 2017.
[24] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021.
[25] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," in Proc. NeurIPS, 2014.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. CVPR, 2016.
[27] R. Girshick, "Fast R-CNN," in Proc. ICCV, 2015.
[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in Proc. NeurIPS, 2015.
[29] Roboflow, "Dangerous Action Detection Dataset," Roboflow Universe, 2023. [Online]. Available: https://universe.roboflow.com