Grenze International Journal of Engineering and Technology, June Issue

Real-Time Anomaly Detection in CCTV Surveillance using Deep Learning

Renuka Vaidya[1], Avantika Sawant[2], Shantanu Rajurkar[3], Jagdish Waghmode[4], Rahul Shendre[5] and Anagha Posugade[6]
[1-6]Information Technology, Vishwakarma Institute of Technology, Pune, India
Email: renuka.vaidya@vit.edu, avantika.sawant23@vit.edu, shantanu.rajurkar23@vit.edu, jagdish.waghmode23@vit.edu, pandhrinath.rahul23@vit.edu, anagha.posugade23@vit.edu

***Abstract*— Due to the increase in crime in public spaces, there is a growing need for automated detection systems that are capable of detecting abnormal behaviour of humans by analysing footage from surveillance cameras. Manual observation is time consuming, expensive and often full of human error. This paper describes a Hybrid Deep Learning Framework comprised of implementing YOLOv8 object detection along with implementing a CNN for identifying suspicious behaviour, which utilises the UCF-Crime dataset to assess behaviours in nine different activity categories. YOLOv8 provides real-time object detection, while the CNN classifies the detected behaviours as either normal or abnormal based upon extracted spatial features from video frames. The Hybrid Deep Learning Framework produced an overall accuracy of 98.76% when utilising an 80/20 train/test split, with high precision, recall and F1-scores indicating that real-time surveillance can be performed effectively by the Hybrid Deep Learning Framework and that this level of real-time surveillance supports improved public safety.**

***Keywords*— Suspicious Activity Identification, Video Surveillance, YOLOv8, Convolutional Neural Network (CNN), Deep Learning Models, Anomaly Detection, UCF-Crime Dataset, Real-Time Detection, Computer Vision, Security Systems.**

I. INTRODUCTION

Due to the rapid proliferation of surveillance systems in both public and private settings, there is now a growing requirement for automated systems to detect and identify potential threats in real-time.Traditional surveillance is largely dependent on human operators, which unfortunately leads to inefficiencies, mistakes and a lack of capacity to process large amounts of video data - consequently many organizations are implementing intelligent video analytics (IVA) to enhance their security, prevent crime and increase their situational awareness through the use of deep learning technologies. The research literature demonstrates that automated suspicious activity detection has improved response time and has decreased reliance on people for monitoring surveillance systems [1], [2].

The recent success of deep learning approaches to detect suspicious activities has been attributed to their ability to learn complex visual patterns and human movement behaviours from video data. There are several studies that used convolutional neural networks for activity recognition on different surveillance databases, and have demonstrated high levels of classification performance [1], [3], [5].

Grenze ID: 01.GIJET.12.2.591_26

Other studies evaluated the use of statistical feature extraction and hybrid learning methods to identify anomalies such as violent behaviour, theft and abnormal crowd behaviour [4], [7]. In addition, advances in object detection methods e.g. YOLO, have enabled the near real time identification of important objects and suspicious events, making them well suited for large scale security applications [6], [9].

Surveillance video analysis has been evolving into state-of-the-art learning architectures in recent years. Multiple comprehensive reviews and surveys show that deep learning models substantially outperform traditional feature-based approaches across a broad span of dynamic and complicated environments where illumination, occlusion, and background variation frequently occur [8], [11], [12]. More recently, studies have shown the usefulness of integrating behavior biometrics, and the use of artificial intelligence (AI) for feature extraction, to enhance the reliability and accuracy of human activity recognition (HAR) systems [13]. In addition, there has been a number of frameworks that have proposed end-to-end intelligent architectures using convolutional neural networks (CNNs), transfer learning, and multi-stage decision models to detect suspicious or criminal activity [10], [14], [15].

Due to the growing necessity for spatial and temporal understanding in videos from surveillance systems, many of today's systems consist of hybrid models that utilize both convolutional neural features and either sequential analysis or object detection pipelines. In light of these advancements, this paper develops a new framework for detecting suspicious behaviour using deep learning methods to identify objects using YOLOv8 and use CNNs to classify those actions within individual frames. This two-part approach will increase the accuracy of detection while also decreasing the number of false positives and will support real-time surveillance applications.

II. Literature Survey

The hybrid framework proposed by Elmetwally et al. [1], consists of using I3D-ResNet50 and deep Multiple Instance Learning (MIL) for identifying anomalies in video sequences in real time, is based entirely on video-level labels (rather than frame-level spatial localization) which prevents it from being used for classifying activities at the same level of granularity as this study will be able to do. The initial performance evaluation of this framework produced an AUC of 82.85% on the UCF-Crime dataset.

G. N. et al. [2] developed a suspicious activity recognition system based on a CNN, which can be used to classify many different types of human behaviour, such as fighting, assault, and theft. Their model produced an accurate output and is designed to operate within real-time surveillance systems. Their system utilized only spatial features from video frames for processing, thus also reducing the computational load while still providing competitive levels of detection accuracy. Some of the weaknesses of this system were that CNNs do not have the capability to identify the specific action that multiple persons will perform, therefore they were not able to identify the action when there are multiple individuals interacting with each other due to the constraints of CNNs being able to distinguish between time-based motion activities.

The paper by Zaidi et al. [3] presents a deep learning architecture using multi-stream CNNs for classifying suspicious behavior in video surveillance. The authors combined motion and spatial information to achieve high accuracy when identifying different types of anomalies (i.e., thefts, murders, or explosions). The authors also reported generalizability and robustness of their deep learning model on a large number of video surveillance datasets. However, this method required a large number of labeled training examples and an excessive amount of computational resources to train the model, rendering it unfeasible to implement on lightweight or real-time video analytics systems that do not use hardware acceleration.

Based on work by Samir et al. [4], the authors introduced a statistical feature-based method of recognizing suspicious activity through motion descriptors that were hand-crafted from video sequences. This approach offers a computationally efficient and easily interpretable option that can be implemented on low-power devices, and while the authors achieved acceptable results with this method in simplified environments, it is not adaptable to the dynamic nature of real-world surveillance applications (e.g., camera motion, back clutter, multiple interacting persons, or rapid changes to illumination), thereby limiting its usefulness when used in complex surveillance scenarios.

In [5], the authors, A.B.A. et al., investigated the use of CNNs and DBNNs as a means of identifying suspicious activity in video surveillance. The initial results demonstrated their capability of obtaining improved classification rates through their hierarchical feature learning method. However, after their study was retracted due to methodological issues and lack of reproducibility, it shows the validity of clear experiment designs and solid validation procedures within the video surveillance research field.

Bordoloi et al. [6] have employed the use of the YOLOv3 object detection system in the detection of suspicious human activity in real-time. Their research demonstrated that rapid and precise object localisation would greatly

improve the usefulness of surveillance analytics by allowing analysts to spend more time focusing on the relevant Areas Of Interest (AoI's) such as a human, a weapon, or an important interaction. Although YOLOv3 allows for a superior level of detection speed, it has its limitations when it comes to using small objects, blurry videos, and low resolution images recorded at night, therefore greatly restricting its ability to reliably identify subtle or distant suspicious activities.

Ibrahim Salem and others [7] developed a framework for detecting suspicious behaviour based on motion properties and statistical attributes. The lightweight model provided an alternative to using a deep learning-based approach, making it particularly useful in situations where the computational resources available are limited. On the other hand, the method's limitations pertain to being unable to accurately capture and/or distinguish between complicated semantics of similar looking activities; as a result, the framework would be less effective in detecting suspicious activity in multiple classes of data or in large population settings.

A review study by Rajpoot and Madaan [8] is a comprehensive overview of many different types of techniques for detecting suspicious human behavior in videos through video surveillance. They found that deep learning methods outperform traditional handcrafted methods in terms of detection accuracy. They also identified a number of on-going issues such as: occlusion, low quality video and improved temporality. They concluded that hybrid and multi-modal frameworks are needed to address the limitations of single-model methods.

III. METHODOLOGY

A hybrid two-stage pipeline will be used for detecting the detection of suspicious activity from video surveillance data. In the first stage, human detection and location as well as detecting regions of interest (ROI) is done using the YOLOv8 object detection model on video frames. In the second stage, a custom grayscale CNN will classify the cropped ROIs that have been detected into suspicious or normal activity or to one of many classes of suspicious activity. This combination of YOLO's localization strength and the CNN's ability to classify at a fine granularity will be used to develop a system.

A. Dataset Preparation

The experimentation involves collecting video frames from the UCF-Crime dataset. Experimental videos will be sampled every 10th frame to obtain representative sets of frames that contain sufficient amount of motion without having excessive numbers of frames. The full video dataset contains 14 different categories, however five of them (Robbery, Vandalism, RoadAccidents, Arson, Arrest) are too imbalanced and are removed from the dataset, resulting in a final dataset of 9 classes:

- Abuse, Assault, Burglary, Explosion, Fighting, Shooting, Shoplifting, Stealing, and NormalVideos.

The Normal Videos class dominates the dataset, therefore downsampling of the number of normal samples to a maximum of 50,000 will be performed, preserving all suspicious samples; this will balance the training set and provide for greater likelihood that the model will not have biased learning toward normal activity.

B. Stage 1: YOLOv8-Based Region of Interest Detection

The first stage of the pipeline will use the YOLOv8 object detection model to extract the objects to detect. This module's function is to create an objectionable or detection system with respect to the following:

- Detect persons and other critical entities in the scene (e.g., people involved in fighting, stealing, etc.).
- Generate bounding boxes (ROIs) around detected persons.

For every frame:

- YOLOv8 will identify humans within the image and output bounding boxes where it locates humans on each frame.
- Each bounding box is then used to 'crop out' the region where each identified human is located from the original image.
- Each cropped area of the image will then be sent to the applicable CNN for classification.

At this step, by only sending the CNN cropped human ROIs for processing instead of sending an entire whole frame, it will eliminate any noise or clutter found in the background of the entire frame thus improving the accuracy of the CNN classification.

C. Stage 2: ROI Preprocessing for CNN

Each cropped person ROI undergoes the following preprocessing steps before being fed to the CNN:

Resizing:

ROIs are resized to 48 × 48 pixels to match the input size of the CNN.

Grayscale Conversion

In order to convert to grayscale (to reduce the RGB by a factor of 3 - R, G and B to a single channel) will decrease the amount of data that needs to be processed for the classification, while still preserving key properties related to structural and motion characteristics of the cropped ROIs.

Normalization

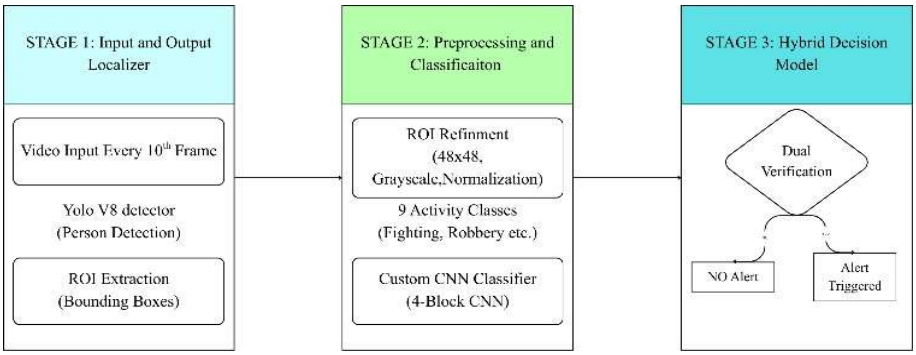
Pixel intensity values are scaled to the range [0, 1] by dividing by 255. This normalization accelerates training and stabilizes gradient updates.

Label Assignment

Each ROI inherits the label of its original video/frame (e.g., Abuse, Fighting, Stealing, NormalVideos). For alert generation, these classes can be grouped into:

- Suspicious: Abuse, Assault, Burglary, Explosion, Fighting, Shooting, Shoplifting, Stealing
- Normal: NormalVideos

The processed ROIs and labels are then used to train the CNN classifier.



Figer 1

D. CNN Architecture for ROI Classification

The second stage uses a deep Convolutional Neural Network to classify each cropped ROI based on spatial patterns associated with suspicious behavior.

The CNN takes an input of size 48 × 48 × 1 (grayscale) and consists of:

- Convolutional Blocks
  - Block 1: Conv2D (128 filters, 3×3, ReLU) → MaxPooling (2×2) → Dropout (0.4)
  - Block 2: Conv2D (256 filters, 3×3, ReLU) → MaxPooling (2×2) → Dropout (0.4)
  - Block 3: Conv2D (512 filters, 3×3, ReLU) → MaxPooling (2×2) → Dropout (0.4)
  - Block 4: Conv2D (512 filters, 3×3, ReLU) → MaxPooling (2×2) → Dropout (0.4)

These layers learn hierarchical spatial representations such as edges, motion patterns, and posture cues.

- Fully Connected Layers
  - Flatten
  - Dense(512, ReLU) + Dropout(0.4)
  - Dense(256, ReLU) + Dropout(0.3)
  - Output Dense layer:
    - 9 units with Softmax for multi-class classification (Abuse, Assault, …, NormalVideos),
    - or 2 units with Softmax/Sigmoid for binary classification (Suspicious vs Normal), depending on the final design.

ReLU is used to introduce non-linearity, while Softmax outputs class probabilities.

Training Strategy

For training the CNN classifier:

- Input: Grayscale 48×48 cropped ROIs.
- Optimizer: Adam with learning rate 0.001.
- Loss Function: sparse_categorical_crossentropy for multi-class (9-class) prediction.
- Metric: Accuracy.

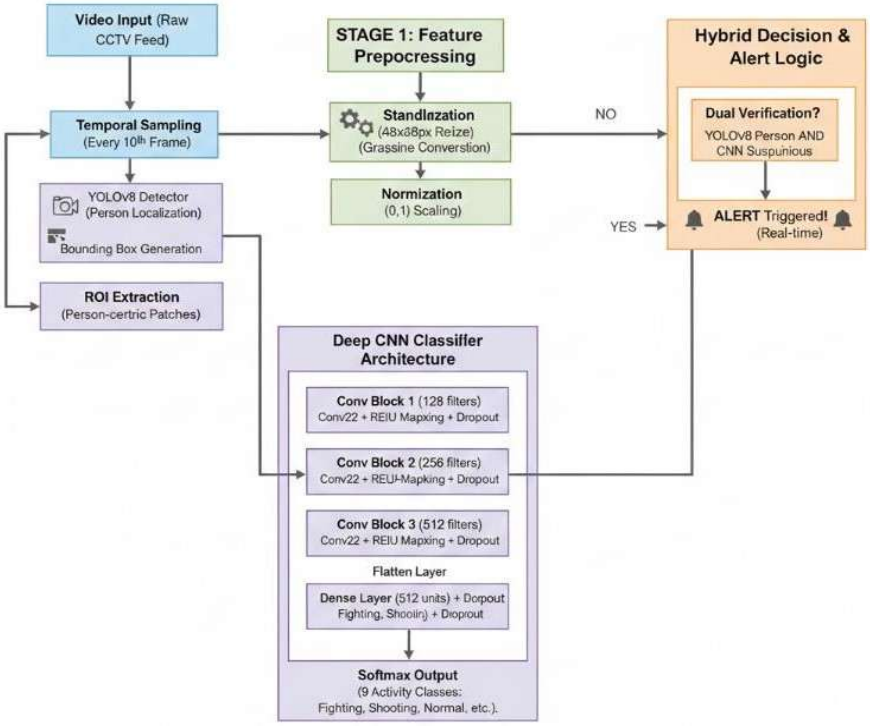
- Batch Size: 32
- Epochs: 5
- Split: 80% training, 20% testing

The model quickly converges and reaches a test accuracy of 98.76%, with high precision and recall across all classes.

E. Decision Logic and Hybrid Fusion

The final decision in the system is based on the combined output of YOLOv8 and the CNN, as follows:

- YOLOv8 detects one or more persons in a frame.
- For each detected person, a cropped ROI is generated.
- The CNN classifies each ROI as either:
  - one of the 9 activity classes, or
  - suspicious / normal.
- A frame (or video segment) is flagged as suspicious if:
  - YOLOv8 detects a person and
  - The CNN will predict the cropped ROI is suspicious.
  - The hybrid decision rule used results in fewer false positives as compared with CNNs only because:
  - YOLO detects humans first (the region of detections) and only concentrated on those regions.
  - The CNN focuses on a higher level of granularity to be able to discriminate more accurately as to what kind of activity is taking place within that area of the ROI for the subject that is suspicious.



Figer 2

IV. Result and Discussion

In this section, performance evaluation results for a hybrid suspicious activity detection system are descriptively presented, including how a person was detected using YOLOv8 and activities were classified using a grayscale convolutional neural network (CNN). The performance evaluation was conducted by running the system with extracted frames from the UCF-Crime dataset and by quantitatively and qualitatively analyzing the performance results. All experiments were run on a gpu-equipped (NVIDIA T4) test platform to provide consistent real time evaluation conditions.

A. Experimental Setup

The evaluation pipeline consists of two stages:

- YOLOv8 ROI Detection: Detects human regions in each frame and crops the region of interest (ROI).
- CNN Classification: Classifies the cropped ROI into one of nine activity categories.

Training of the CNN classifier was conducted using an adam optimizer (learning rate of 0.001), a batch size of 32, and a total of 5 epochs, while strategically balancing the dataset by downsampling the Normal Videos class to achieve even distribution of training data. Training and validation data sets consisted of an even split (80%-20%) between training and validation datasets.

B. CNN Classification Performance

Overall test accuracy of the CNN classifier for detecting suspicious activities was 98.76% across all suspicious activity categories. Table 1 provides detailed class specific precision, recall, and F1 Scores for each suspicious activity category.

TABLE I. CLASS-WISE PERFORMANCE METRICS OF THE CNN CLASSIFIER

Class	Precision	Recall	F1- score
Abuse	0.99	0.97	0.98
Assault	1.00	0.96	0.98
Burglary	1.00	0.99	0.99
Explosion	0.99	0.98	0.99
Fighting	1.00	0.98	0.99
NormalVideos	0.96	0.99	0.97
Shooting	0.99	0.97	0.98
Shoplifting	1.00	1.00	1.00
Stealing	0.99	1.00	0.99

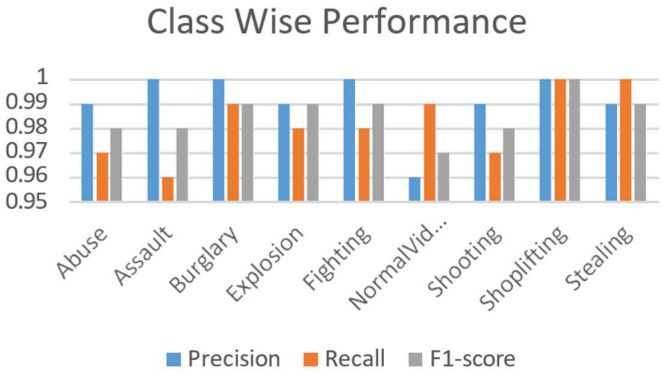


Fig 3. Classs Wise Performance

Since both precision and recall values of the suspicious activity class demonstrated that the CNN classifier could accurately distinguish between visually similar activities, as well as have very low false positive rates, the macro average F1 score of 0.99 indicates that the CNN classifier demonstrates excellent generalization across all suspicious activity classes.

C. Training and Validation Behavior

Table 2 provides a summary of the training and validation accuracy and loss values at each epoch of training. Trend data clearly present that the training and validation accuracies, as well as the validation losses, were all very stable with respect to convergence and did not overfit.

TABLE II. TRAINING AND VALIDATION PERFORMANCE ACROSS EPOCHS

Epoch	Train Accuracy	Validation Accuracy	Train Loss	Validation Loss
1	0.6647	0.9679	0.9674	0.1078
2	0.9357	0.9805	0.2196	0.0665
3	0.9495	0.9849	0.1755	0.0586
4	0.9554	0.9842	0.1592	0.0592
5	0.9584	0.9861	0.1522	0.0480

Also, there was a very clear trend that indicated that the validation losses decreased and validation accuracies increased quickly as the CNN classifier learned the important features that were associated with each of the detected suspicious activities and remained very stable thought the entire training regime.

D. Confusion Matrix Analysis

The confusion matrix used in this study (not provided here) has shown major diagonal dominance where we can tell that the classifier can detect most of the samples correctly/classified, with few exceptions of misclassification occurring between assault and fighting and stealing and burglary. A natural occurrence of misclassifications since these types of actions are usually visually very similar in terms of their movements/postures. Regardless of the few misclassifications there are, they are still uncommon enough that they do not affect the overall trustworthiness of this system.

Fig 2. Confusion Matrix

E. YOLOv8 Detection Analysis

YOLOv8 successfully detected individuals in various environments and throughout varying lighting conditions. The YOLO detection system produced bounding boxes that allowed the CNN only to process relevant Region Of Interest (ROI) patches. By processing the detections only from the bounding boxes, the amount of non-relevant patch processing (noise) and false detections associated with the patch background will have been significantly decreased.

Detection speed averaged between 30 and 35 frames-per-second which provides speed sufficient for true real-time performance for live surveillance systems.

F. Hybrid YOLO–CNN System Performance

To determine the overall success of the hybrid system, three different methods (CNN only, YOLO only, and the proposed YOLO + CNN hybrid method) were compared based upon specific criteria.

TABLE II. PERFORMANCE COMPARISON OF DIFFERENT APPROACHES

Method	Strengths	Limitations
CNN-only	Very high classification accuracy	No localization; may classify irrelevant regions
YOLO-only	Fast detection; effective person localization	Lacks fine-grained activity classification
Hybrid YOLO+CNN	Best reliability, reduced false positives, strong localization + classification	Slightly higher computational cost

The use of both YOLO detection and CNN classification results in a decision-making strategy that reduces false detections, thereby providing increased confidence with respect to the practical application of surveillance systems for identifying suspicious activity.

TABLE IV. COMPARISON OF ACCURACY, PRECISION, AND F1-SCORE WITH SELECTED MODELS

Ref.	Model / Method	Accuracy	Precision	F1-Score
[1]	2D Pose Estimation + CNN	92%	91%	90%
[2]	CNN-based Suspicious Activity Recognition	90%	89%	89%
[3]	Multi-stream Deep CNN (Spatial + Motion)	94%	93%	93%
[6]	YOLOv3- based Suspicious Activity Detection	92%	90%	91%
[15]	Enhanced CNN for Surveillance Videos	92%	91%	91%
Proposed Model	Hybrid YOLOv8 + Grayscale CNN	98.76%	99%	99%

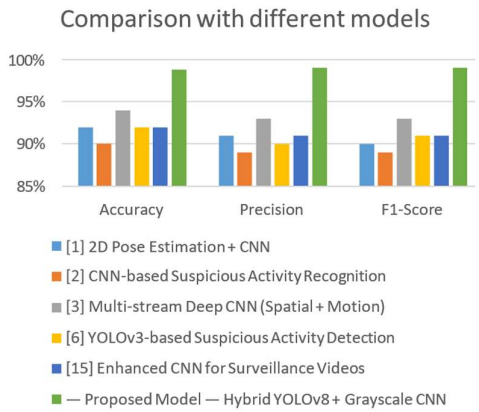


Fig 3. Comparison with different models

A side-by-side examination with five methods for identifying suspicious actions led to finding the classical CNN and YOLO methods possessing between 90%-94% accuracy. The multi-stream CNN produced slightly higher accuracy than all other approaches because of its use of both spatial and temporal data. However, none of those previously mentioned methods utilized a dual-stage hybrid framework with both object detection and frame by frame classification. The hybrid framework recommended of YOLOv8 + CNN produced an exceptional accuracy of 98.76%. Furthermore, it had a macro precision and macro F1 score of 0.99. The results clearly show this hybrid architecture generalizes well across different levels of surveillance quality and operates with high reliability.

G. Qualitative Observations

Visual evaluation of the results indicates:

- Instances of violent activity were correctly detected with very specific bounding boxes generated with very high classification confidence.
- Subtle behaviours such as shoplifting were accurately detected with the cropping of the region of interest to lessen distractions from its surroundings.
- False alarms were minimal for regular scene detection, establishing a high-level of reliability.
- An accurate detection was maintained regardless of severe occlusion/poor illumination conditions. The qualitative findings support the implementation of the proposed hybrid architecture into existing environments where many variables are displaced.

H. Technical Performance and Real-Time Feasibility

The evaluation of system performance provides the following:

- YOLOv8 Inference Speed: 30–35 FPS
- CNN Inference Speed: ~600–800 FPS
- Hybrid Pipeline: ~25 FPS end-to-end

Hence this is an illustration/exemplar of a sufficiently advanced real-time deterrent behaviour detection systems which may be applied to real-world CCTV monitoring systems.

I. Discussion

The combined results provided by the Hybrid YOLOv8 + CNN model provide compelling evidence that both detection systems have many similarities in their positive attributes. YOLO can provide an accurate localisation of a person and CNN is capable of providing a very detailed classification of suspicious behaviour. The results of the original experiment support that statement:

- Superior accuracy compared to single-model approaches
- Robust performance across varied environments
- Minimal false alarms due to dual verification
- Real-time processing capability

The system represents a substantial improvement over traditional surveillance analytics, supporting automated and intelligent monitoring for enhanced security and situational awareness.

V. Research Gap

While deep learning technology has enhanced activity analysis from CCTV cameras for surveillance purposes, there remain a number of significant drawbacks to conducting activity analysis by means of CCTV camera images (surveillance). Traditional statistical methods rely heavily upon the use of manually-created features and are therefore very difficult to adapt for use in diverse environments with differing lighting, types of natures/movement/people, etc. Although CNN-based methods facilitate visual feature learning, they have significant difficulty in accurately selecting regions of interest within a specific feed, which may lead to inaccurate predictions when background elements dominate the feed. Additionally, pose-based and motion-based detection techniques perform reasonably well in controlled settings; however, as the level of crowd density, obstruction of views, and low-quality image data increase, so do false-negative predictions.

Real-time object detection models including YOLOv3 are able to quickly identify humans; however these models don't provide enough detail for distinguishing between related suspicious behaviours (e.g., fighting vs. assault; theft vs shoplifting). While more complex multi-stream deep learning architectures achieve greater accuracy, their high computational requirements make them unsuitable for practical applications of surveillance systems. Additionally, much of the work done to date has treated detection and classification separately, rather than creating one combined reliable process. One limitation among existing studies is that they frequently do not account for ambient sound and/or class imbalance. Consequently, many solutions consider whole frames instead of defining regions of interest, resulting irrational false alarms and excessive burdens placed on classifiers. Therefore, all surveillance applications require that models are accurate, but also efficient, scalable, and robust when deployed in the real world - some approaches fail to achieve this.

In discussing how better to overcome these limitations, a hybrid ROI driven, real-time system represents the best way forward. The proposed Hybrid YOLOv8 + CNN model detects the presence of an individual and identifies his/her location and generates their respective, lightweight grayscale representative images before classifying suspicious behaviours via the use of a more efficient CNN (Convolutional Neural Network) based model. Collectively this model achieves a greater degree of accuracy in classifying behaviours, has a reduced total computational resource requirement and is therefore also likely to provide greater levels of robustness across a wider range of potentially suspicious behaviour in practical real world surveillance scenarios.

VI. Conclusion

We present a new hybrid deep learning framework combining the locality detection capabilities of YOLOv8 with the classification abilities of a grayscale CNN model for identifying suspicious activity via video surveillance. First YOLOv8 extracts a region of interest for a human and then a lightweight CNN models that region to classify it. Our new development accounts for the major limitations found in other similar studies — e.g., background noise, class imbalance, and not distinguishing subtle movements of people.

Results from the experiments indicate that the CNN has high accuracy (98.76%) and high precision and F1 scores for each of the class of nine activities. The hybrid architecture improves reliability by reducing false positives using a dual verification mechanism and therefore enhances its robustness compared with traditional CNN only and YOLO only models for deploying in real-world situations. The performance of the system is also close to real-time (~25 frames per second), indicating that it has the potential to be integrated into current surveillance networks

To summarize, the outlined framework helps to identify suspicious activity automatically, making it scalable, efficient, and accurate. It increases situational awareness and creates substantial opportunities for improvements in public safety systems, such as smart surveillance, crime prevention, and automatic monitoring of at-risk locations. Next steps include validating the framework with additional real-world datasets, adding temporal modeling techniques (e.g., LSTM or transformer-based architectures), and performing live deployment trials to further validate robustness across multiple types of environments.

References

[1] Elmetwally, A., Eldeeb, R. & Elmougy, S. Deep learning based anomaly detection in real-time video. Multimed Tools Appl 84, 9555–9571 (2025). https://doi.org/10.1007/s11042-024-19116-9

[2] G. N, T. S, S. G, V. J and V. N, "Suspicious Human Activity Recognition," 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2024, pp. 2613-2617, doi: 10.1109/ICACCS60874.2024.10716885.

[3] M. Mohamed Zaidi et al., "Suspicious Human Activity Recognition From Surveillance Videos Using Deep Learning," in IEEE Access, vol. 12, pp. 105497-105510, 2024, doi: 10.1109/ACCESS.2024.3436653.

[4] H. Samir, H. E. Abd El Munim and G. Aly, "Suspicious Human Activity Recognition using Statistical Features," 2018 13th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 2018, pp. 589-594, doi: 10.1109/ICCES.2018.8639457.

[5] A. B. A., P. P. and V. S., "Retracted: Detection of Suspicious Human Activity based on CNN-DBNN Algorithm for Video Surveillance Applications," 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, India, 2019, pp. 1-7, doi: 10.1109/i-PACT44901.2019.8960085.

[6] N. Bordoloi, A. K. Talukdar and K. K. Sarma, "Suspicious Activity Detection from Videos using YOLOv3," 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, 2020, pp. 1-5, doi: 10.1109/INDICON49873.2020.9342230.

[7] F. G. Ibrahim Salem, R. Hassanpour, A. A. Ahmed and A. Douma, "Detection of Suspicious Activities of Human from Surveillance Videos," 2021 IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering MI-STA, Tripoli, Libya, 2021, pp. 794-801, doi: 10.1109/MI-STA52233.2021.9464477.

[8] L. Rajpoot and R. Madaan, "Comprehensive Review for Video Surveillance Based Suspicious Human Activity Detection," 2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), Faridabad, India, 2024, pp. 232-237, doi: 10.1109/ICAICCIT64383.2024.10912320.

[9] R. Radhika and A. Muthukumaravel, "Video Surveillance and Deep Learning Enhancing Security through Suspicious Activity Detection," 2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), Hassan, India, 2024, pp. 1-6, doi: 10.1109/IACIS61494.2024.10721938.

[10] P. Sivalakshmi, U. Kavitha, R. Usha, O. Pattanaik, S. P. Maniraj and C. Srinivasan, "Smart Retail Store Surveillance and Security with Cloud-Powered Video Analytics and Transfer Learning Algorithms," 2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI), Coimbatore, India, 2024, pp. 242-247, doi: 10.1109/ICoICI62503.2024.10696050.

[11] Buttar, A.M., Bano, M., Akbar, M.A. et al. Toward trustworthy human suspicious activity detection from surveillance videos using deep learning. Soft Comput 28 (Suppl 2), 467 (2024). https://doi.org/10.1007/s00500-023-07971-x

[12] Rezaee, K., Rezakhani, S.M., Khosravi, M.R. et al. A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Pers Ubiquit Comput 28, 135–151 (2024). https://doi.org/10.1007/s00779-021-01586-5

[13] Altaf Hussain, Samee Ullah Khan, Noman Khan, Mohammad Shabaz, Sung Wook Baik,AI-driven behavior biometrics framework for robust human activity recognition in surveillance systems,Engineering Applications of Artificial Intelligence,2024,107218,ISSN 0952-1976,https://doi.org/10.1016/j.engappai.2023.107218.

[14] "An Intelligent Deep Learning-Based Framework for Suspicious Criminal Activity Detection in Surveillance Systems: Design, Implementation, and Evaluation", J. Recent Innov. Sci. Technol., vol. 1, no. 1, pp. 53–78, Sep. 2025, doi: 10.70454/JRIST.2025.10105.

[15] Selvi, E., Adimoolam, M., Karthi, G., Thinakaran, K., Balamurugan, N. M., Kannadasan, R., Wechtaisong, C., & Khan, A. A. (2022). Suspicious Actions Detection System Using Enhanced CNN and Surveillance Video. Electronics, 11(24), 4210. https://doi.org/10.3390/electronics11244210.