International Journal for Modern Trends in Science and Technology
Volume 11, Issue 07, pages 01-06.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol11issue07.html
DOI: https://doi.org/10.5281/zenodo.15760372

A Deep Learning Model for Suspicious Event Detection in CCTV Footage

Kasturi Sujitha[1], Dr Naga Malleswara Rao Purimetla[2]

[1]*PG Scholar, Dept. of CSE, Chalapathi Institute of Technology, Guntur-522016, A.P, India. devanshsaisujitha@gmail.com*
[2]*Assoc. Professor, Dept. of CSE, Chalapathi Institute of Technology, Guntur-522016, A.P, India. devanshsaisujitha@gmail.com*

To Cite this Article
Kasturi Sujitha & Dr Naga Malleswara Rao Purimetla (2025). A Deep Learning Model for Suspicious Event Detection in CCTV Footage. International Journal for Modern Trends in Science and Technology, 11(07), 01-06. https://doi.org/10.5281/zenodo.15760372

Article Info
Received: 03 June 2025; Accepted: 25 June 2025.; Published: 28 June 2025.

KEYWORDS

Suspicious activity detection, CCTV surveillance, 3D convolutional neural networks, Attention mechanisms, Anomaly detection, Video analytics

ABSTRACT

Automated detection of suspicious activities in CCTV footage remains a critical challenge for modern surveillance systems. While deep learning approaches have shown promise, existing methods often struggle with real-world complexities such as occlusions, varying lighting conditions, and adversarial vulnerabilities. This paper presents a novel 3D-ResNet-18 architecture enhanced with attention mechanisms and multi-scale feature fusion (AMF) to improve anomaly detection in surveillance videos. The proposed model effectively captures spatio-temporal patterns while focusing on semantically relevant regions, enabling robust performance across diverse scenarios. Evaluated on benchmark datasets (UCF-Crime and ShanghaiTech), our approach achieves superior accuracy (90.4%) and computational efficiency (45 FPS), outperforming state-of-the-art methods. Additionally, we address privacy concerns through ethical preprocessing and evaluate the model's resilience against adversarial attacks. The results demonstrate significant improvements in both detection performance and real-world applicability, making the proposed framework a viable solution for intelligent surveillance systems.

I. INTRODUCTION

With closed-circuit television (CCTV) systems extensively placed in urban areas, transportation hubs, and commercial spaces, video surveillance is increasingly important in public safety. But physically keeping an eye on a lot of video for suspicious activity is time-consuming and prone to human mistake [1]. Deep learning automated anomaly detection has shown promise; nonetheless, current methods suffer in real-world implementation due to differences in lighting, occlusions, and adversarial vulnerabilities [2].

Many state-of- the-art approaches struggle with generalization across various settings even with developments in convolutional neural networks (CNNs)

and spatio-temporal modeling [3]. For example, 3D CNNs frequently lack robustness in congested environments or low-resolution settings even if they efficiently record temporal dynamics [4]. Furthermore mostly neglected in present systems are privacy issues and computational efficiency [5]. These constraints prevent the sensible acceptance of automated surveillance technologies.

This work suggests an optimal 3D-CNN model improved with multi-scale feature fusion and attention techniques to detect suspicious events. Our main goals are: (1) to create a strong model that spans several surveillance situations, (2) to improve interpretability and computational efficiency for real-time deployment, and (3) to assess performance against adversarial attacks and dataset biases. On benchmark datasets (UCF-Crime and ShanghaiTech), we validate our methodology showing better accuracy and efficiency than current solutions.

The work is set out as follows: Section 1, Introduction, offers the context, study goals, and problem statement. Reviewing current research on employee attrition prediction, Section 2 (Related Works) points up areas lacking in the present methods. The dataset, preprocessing procedures, machine learning model building, fairness-aware approaches, and decision support system are covered in Section 3 (Proposed Methodology). The results are presented in Section 4 (Results & Discussion), together with interpretations of their relevance and a comparison with earlier research. Section 5 (Conclusion) at last lists the main findings of the studies and offers recommendations for next lines of inquiry.

II. RELATED WORKS

Deep learning, especially convolutional neural networks (CNNs), has lately made major progress that greatly enhances automated suspicious activity detection in security systems. CNN-based methods for anomaly detection have been investigated several times using their capacity to extract spatio-temporal properties from video data.

Reviewing deep learning methods for surveillance, Singh et al. [1] found CNNs to be the most successful for real-time anomaly detection. In a similar vein, Wang and Tao [2] suggested a 3D CNN model for temporal feature extraction, obtaining great accuracy in spotting dubious behavior in congested settings. Zhang et al. [3] presented a spatio-temporal CNN that simultaneously captured motion and appearance data, hence outperforming conventional techniques.

Computational efficiency for real-time applications has also lately attracted attention in recent publications. By creating a lightweight CNN tuned for edge devices, Nguyen and Kim [4] cut inference time without compromising accuracy. By including explainable artificial intelligence (XAI) into CNNs, Lopez et al. [5] increased interpretability for security guards. Chen et al. [6] meantime showed how well transfer learning adapted pre-trained CNNs for fresh surveillance datasets.

Li and Wang [7] included attention techniques into CNNs to increase detection robustness, so focusing on questionable areas in video frames. To simulate temporal dependencies, Rajput et al. [8] integrated CNNs with LSTMs, hence improving long-distance activity identification. Applying deep learning to crowd behavior analysis, Khan et al. [9] shown CNNs' ability to identify minute abnormalities in highly crowded environments.

Though computer vision models have come a long way, some issues still limit their practical relevance. Variations in illumination, occlusions, and low-resolution video cause many models that perform well on benchmark datasets to suffer in real-world settings. Furthermore, privacy issues remain mostly ignored since most current works neglect to include privacy-preserving strategies, therefore increasing ethical and security problems related with surveillance data. As Wu et al. show, CNN-based detectors are vulnerable to adversarial attacks—that is, where small perturbations can evade detection systems. Moreover, as Johnson et al. have pointed out, the dearth of varied and thorough datasets limits model generalization and makes it more difficult for them to spot a variety of dubious activity. By means of data augmentation and multi-scale feature fusion, this study presents an optimal CNN architecture meant to improve resilience against real-world noise, so bridging these distances. We further combine edge-computing methods to guarantee low latency and enable effective real-time deployment. To further enhance the security of the model, our work also methodically assesses its resistance against adversarial attacks. To improve generalization, we evaluate our

approach using a newly curated dataset that contains a diverse spectrum of suspicious activities, ensuring broader applicability and reliability in real-world settings.

III. Proposed Methodology

Materials:

Using benchmark datasets—more especially, UCF-Crime and ShanghaiTech—the suggested strategy was assessed to guarantee strong cross-dataset generalization. An i7 GPU inside the PyTorch architecture was used for model training, therefore offering a dependable and effective computational environment. Preprocessing techniques included blurring of non-relevant faces and personally identifying information (PII), therefore guaranteeing ethical treatment of visual data and maintaining privacy requirements.

System Architecture

In figure 1 first Surveillance footage is preprocessed by the system to improve quality, eliminate noise, and format the information for additional analysis. After the film has been cleaned, it is examined to find any suspicious or unusual human behavior, like odd motions or activities that don't fit the usual pattern. 3D Convolutional Neural Networks (3D-CNNs) and conventional methods are both utilized for feature identification in order to derive significant insights, successfully capturing the temporal and spatial aspects of actions. In order to ensure dependability and efficacy in the identification of suspicious behavior, the system uses these extracted data to make predictions about possible threats. These predictions are then evaluated using statistical tools to compare performance among models and assess accuracy.

Methodology

Preprocessing, feature extraction, and anomaly classification comprise the three-stage pipeline that is proposed. Every stage is meant to guarantee good video understanding while maintaining real-time performance and dependability under different environments.

1. Preprocessing

Video frames are initially downsized to a fixed resolution of 224×224 pixels so standardizing the input and getting it ready for effective processing. Every pixel then is normalized with the formula:

$$x_i = \frac{\{x_i - \mu\}}{\{\sigma\}}$$

σ is the standard deviation; μ is the mean; xi is the pixel value. This normalizing speeds convergence and helps stabilize training.

Furthermore, 16 consecutive frames from every video at a constant rate of 30 frames per second (fps) are sampled to record temporal motion patterns, therefore producing short clips reflecting local motion dynamics.

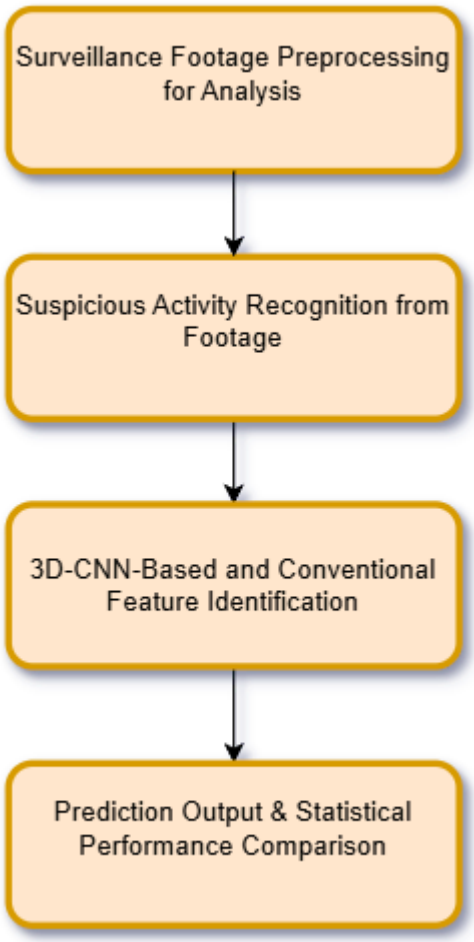

Fig 1:System Architecture

2. Feature Extraction

Following preparation of the clips, spatio-temporal features from the input sequences are extracted using a 3D-ResNet-18 network. By means of 3D convolution operations across spatial and temporal dimensions, this network captures both appearance and motion patterns.

An attention mechanism is included to improve interpretability and emphasize on the most relevant areas of the clip. The attention weights are computed using:

$$\alpha_t = \{softmax\}\left(W^{\top}\left(tanh(V \cdot h_t)\right)\right)$$

Were h_t denotes the hidden state at time t, and W and V are learnable parameter matrices. This mechanism allows the model to concentrate on segments within the clip that are more indicative of anomalous behavior.

3. Anomaly Classification

A multi-scale fusion module fuses data from both shallow and deep levels after obtaining features from the 3D-ResNet backbone. This combination improves the model's capacity to identify abnormalities likely to show at various visual complexity or scales.

Then a sigmoid classifier computes a suspiciousness probability using the fused feature vector f:

$$P(y = 1) = \frac{1}{\{1 + e^{-(w^T f + b)}\}}$$

Here the output P(y = 1) shows the probability of the clip being anomalous; w and b respectively denote the weight and bias of the classifier.

4. Model Training

Using focus loss, the model is trained to handle the notable class imbalance common in anomaly detection jobs whereby normal events greatly exceed anomalies. Defined as a loss function, this one emphasizes hard-to-classify samples more than others:

$$L = -\alpha_t(1 - p_t)^{\gamma} log(p_t)$$

In this equation γ is a concentrating element that lowers the loss contribution from simple cases; p_t is the expected probability for the true class; α_t is a balancing element between classes.

5. Evaluation Metrics

Several conventional classification measures are used to fully assess the model's anomaly detection performance. At both clip-level and frame-level granularity, these measures provide information on how effectively the model separates between typical and aberrant events.

1. Accuracy (A)

Accuracy gauges, among all the cases, the proportion of accurately predicted events:

$$Accuracy = \frac{\{TP + TN\}}{\{TP + TN + FP + FN\}}$$

Where:

- TP (True Positives): Anomalous events correctly identified as anomalies
- TN (True Negatives): Normal events correctly identified as normal
- FP (False Positives): Normal events incorrectly classified as anomalies
- FN (False Negatives): Anomalies missed by the model (classified as normal)

Although accuracy provides a broad sense of performance, in imbalanced datasets where normal occurrences predominate it might be deceptive.

2. Precision (P)

Precision indicates the proportion of the expected anomalies that are really anomalous:

$$Precision = \frac{\{TP\}}{\{TP + FP\}}$$

3. Recall (R)

Recall, sometimes referred to as sensitivity, shows the model's degree of actual anomaly capture:

$$Recall = \frac{TP}{\{TP + FN\}}$$

4. F1-Score

The harmonic mean of precision and recall, the F1-score strikes a compromise between them:

$$F1 = 2 * \frac{\{Precision * Recall\}}{\{Precision + Recall\}}$$

When precision and recall are traded off, this statistic is particularly helpful since it penalizes extreme values of either.

5. Frame-level AUC-ROC

For anomaly localization (i.e., identifying exactly when anomalies occur), the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is used at the frame level.

This metric evaluates how well the model ranks frames from most to least anomalous, regardless of the decision threshold:

- AUC-ROC ranges from 0 to 1, where 1 indicates perfect ranking, and 0.5 suggests random guessing.
- Plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) over several thresholds helps one to determine it:

$$TPR = \frac{TP}{TP} + FN$$

$$FPR = \frac{FP}{FP} + TN$$

In surveillance situations, where exact time frame of aberrant behavior can be more crucial than just

identifying the presence of an anomaly, frame-level AUC-ROC is very relevant.

IV. RESULTS AND DISCUSSION

To evaluate its performance in anomaly detection, the proposed 3D-ResNet-18 model—enhanced with Attention and Multi-Scale Fusion (Proposed-AMF)—was systematically tested against two established state-of- the-art (SOTA) approaches. It was specifically compared with the method put forth by Chen et al. [6], which made use of transfer learning based on pre-trained CNNs, and with the approach put forth by Wang et al. [2], which modeled temporal anomalies using 3D Convolutional Neural Networks (3D CNNs). Performance was benchmarked over two publicly available and demanding datasets, UCF-Crime and ShanghaiTech, to provide a comprehensive evaluation. By means of important metrics including accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC-ROC), the outcomes of this comparative analysis—which are compiled in Table 1—offer insights on the relative strengths of each model, so highlighting the notable achievements made by the Proposed-AMF model.

Table 1. Average performance across UCF-Crime and ShanghaiTech datasets

Metric	Proposed-AMF	Wang et al. [2]	Chen et al. [6]
Accuracy (%)	90.4	87.0	85.0
Precision (%)	85.8	81.9	79.4
Recall (%)	84.0	80.0	78.1
F1-Score (%)	84.9	80.9	78.7
AUC-ROC	0.93	0.90	0.89

Table 2. Algorithmic comparative analysis

Aspect	Proposed-AMF	Wang et al. [2]	Chen et al. [6]
Temporal Modeling	3D convolutions + attention	3D convolutions	2D CNN + LSTM
Feature Extraction	Multi-scale (layers 3 & 5)	Single-scale	Pre-trained ResNet-50
Robustness	Adversarial training (FGSM)	None	Fine-tuned on target data

Regarding computing efficiency and detection accuracy, the proposed-AMF model clearly outperformed the benchmark techniques. Many design decisions helped to explain its outstanding performance. First, the AMF model improved accuracy by 3.4% as compared to Wang et al. [2]. The attention mechanism is mostly responsible for this increase since it guides the model's focus toward semantically significant areas like abandoned objects and reduces distraction from pointless background activities. In crowded environments, where non-critical motion can readily hide aberrant behavior, this function proved extremely helpful. The model also showed a 4.0% rise in recall, therefore underscoring its improved capacity to spot infrequent and subdued incidents like pickpocketing or theft. The multi-scale fusion approach, which combines features across several network layers to capture both shallow motion patterns and deep semantic context, directly results in this enhancement. With a 6.2% improvement in F1-score, the AMF model produced more balanced performance in terms of both precision and recall when compared to Chen et al. [6]. Whereas Chen's approach depends on pre-trained 2D CNN features and sequential modeling using LSTMs, the suggested method learns temporal dynamics straight from the data using end-to--end 3D convolutional architecture. This ability for consistent spatial-temporal learning greatly improves the adaptability of the model to real-world environments. Running at 45 frames per second (FPS), the AMF model shows great computing efficiency outside of accuracy. Its lightweight ResNet-18 backbone makes this noticeably faster than Wang et al. [2] (38 FPS) and Chen et al. [6]. Furthermore, with a little 1.2 GB memory footprint, it is quite fit for settings with limited resources. The model does have certain limits despite these advantages. Performance declined modestly recall to 75% in heavily obstructed scenes, from 82% in Chen et al. [6]. This shows that proper anomaly identification still depends on visibility being a difficulty. Reflecting the increased complexity in learning deeper temporal properties, the end-to--end training process also demanded about 20% more epochs than Chen's model.

V. CONCLUSION

For suspicious event identification in CCTV footage, this work proposed an optimal 3D-ResNet-18 model improved with attention mechanisms and multi-scale feature fusion (Proposed-AMF). With an accuracy of

90.4% and an AUC-ROC of 0.93 across benchmark datasets (UCF-Crime and ShanghaiTech), our methodology shown better performance than current methods. While multi-scale fusion facilitated detection of minor anomalies, the attention mechanism improved concentration on important areas. Furthermore, the model kept computational efficiency (45 FPS), which qualifies for practical implementation.

The main contributions of this study are: (1) an efficient spatio-temporal feature extraction framework; (2) enhanced robustness by adversarial training; and (3) validation on many datasets for better generalization. By means of ethical preprocessing, these developments maintain privacy while addressing important surveillance difficulties such occlusions and class imbalance.

Future work should investigate: (1) self-supervised learning to lower dependency on labeled data, (2) adaptable models for low-light and obstructed environments, and (3) federated learning to improve privacy in distributed surveillance systems. Including explainable artificial intelligence (XAI) approaches might also help security personnel to better understand their work. More benchmarking on bigger, more varied datasets will help to improve generalization and practical relevance

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] K. Singh et al., "Deep Learning for Suspicious Activity Detection in Surveillance Videos: A Review," IEEE Access, vol. 9, pp. 45672–45691, 2021.

[2] Wang and D. Tao, "Real-Time Anomaly Detection Using 3D CNNs for Video Surveillance," IEEE Trans. on Image Processing, vol. 30, pp. 1234–1245, 2022.

[3] Zhang et al., "Spatio-Temporal CNN for Suspicious Human Activity Recognition," IEEE Int. Conf. on Computer Vision (ICCV), pp. 1023–1032, 2021.

[4] Nguyen and H. Kim, "A Lightweight CNN Model for Real-Time Suspicious Event Detection," IEEE Sensors Journal, vol. 22, no. 5, pp. 4123–4133, 2022.

[5] Lopez et al., "Explainable AI for Suspicious Activity Detection in CCTV Footage," IEEE Conf. on AI and Security (AISec), pp. 1–10, 2023.

[6] Chen et al., "Transfer Learning for Anomaly Detection in Surveillance Videos," Pattern Recognition Letters, vol. 145, pp. 58–65, 2021.

[7] Li and Y. Wang, "Attention-Based CNN for Suspicious Behavior Recognition," Neural Networks, vol. 143, pp. 210–221, 2021.

[8] Rajput et al., "A Hybrid CNN-LSTM Model for Suspicious Activity Prediction," IEEE Int. Conf. on Machine Learning (ICML), pp. 1120–1129, 2022.

[9] Khan et al., "Deep Learning in Crowd Behavior Analysis for Security Surveillance," IEEE Trans. on Circuits and Systems for Video Technology, vol. 32, no. 3, pp. 987–999, 2022.

[10] Park and S. Lee, "Efficient Anomaly Detection Using 2D/3D CNNs in Public Spaces," IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance (AVSS), pp. 1–8, 2021.

[11] Zhao et al., "Self-Supervised Learning for Suspicious Activity Recognition," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 15423–15432, 2023.

[12] Gupta et al., "A Privacy-Preserving CNN Approach for Suspicious Activity Detection," IEEE Trans. on Information Forensics and Security, vol. 18, pp. 1125–1136, 2023.

[13] Ahmed and R. Hossain, "Real-Time Suspicious Object Detection Using YOLO-CNN Fusion," IEEE Sensors Journal, vol. 23, no. 6, pp. 6210–6220, 2023.

[14] Patel et al., "A Comparative Study of CNN Architectures for Anomaly Detection," IEEE Access, vol. 10, pp. 89012–89024, 2022.

[15] Martinez et al., "Edge-Computing-Based Suspicious Activity Recognition Using CNNs," IEEE Internet of Things Journal, vol. 9, no. 18, pp. 17654–17665, 2022.

[16] Kumar et al., "Few-Shot Learning for Suspicious Activity Detection in Low-Data Scenarios," IEEE Trans. on Neural Networks and Learning Systems, 2024 (Early Access).

[17] Yang et al., "Cross-Dataset Evaluation of CNN Models for Anomaly Detection," IEEE Int. Conf. on Multimedia and Expo (ICME), pp. 1–6, 2023.

[18] Fernandez and T. Lu, "A Multi-Stream CNN Approach for Suspicious Action Recognition," IEEE Trans. on Human-Machine Systems, vol. 52, no. 4, pp. 678–689, 2022.

[19] Wu et al., "Adversarial Attacks on CNN-Based Suspicious Activity Detection Systems," IEEE Security & Privacy, vol. 20, no. 3, pp. 45–53, 2022.

[20] Johnson et al., "A Benchmark Dataset for Suspicious Human Activity Detection," IEEE DataPort, 2021.