Journal of Science, Innovation and Creativity
Vol. 5 No. 1 (2026): ISSN (Online): 2958-454X
DOI: https://doi.org/10.58721/yz2zk257
Deep Learning Approaches for Crime Detection in Surveillance Systems: A Comprehensive Review

Deep Learning Approaches for Crime Detection in Surveillance Systems: A Comprehensive Review

Hesbon Amwayi, Jairus Odawa & Cedric Okinda
Masinde Muliro University of Science and Technology, Kenya

Article History
Received: 2026-03-28
Revised: 2026-07-04
Accepted: 2026-07-07
Published: 2026-07-09

Keywords
Attention mechanisms
Crimes
Deep learning
Surveillance

How to cite:
Amwayi, H., Odawa, J., & Okinda, C. (2026). Deep Learning Approaches for Crime Detection in Surveillance Systems: A Comprehensive Review. *Journal Science, Innovation and Creativity, 5*(1), 222-241.

Abstract
The rapidly rising crime rates have necessitated advanced and automated security surveillance systems capable of robust, real-time detection, recognition, and prevention of crime. Although traditional surveillance systems have largely been deployed to enhance security and safety, they are inefficient, error-prone, and incapable of effectively processing and generating meaningful insights from the vast quantities of video data they produce. Further, they are adversely affected by extreme weather conditions and subject to human vandalism. The advent of deep learning has significantly transformed earlier automated crime detection, recognition, and prevention by enabling robust extraction and analysis of complex spatial and temporal features from surveillance videos. This research presents a comprehensive review and synthesis of various state-of-the-art deep learning architectures employed in surveillance video-based crime detection and recognition systems, including 3D Convolutional Neural Networks (3D-CNN), Residual Networks (ResNets), Recurrent Neural Networks (RNN), Bidirectional Long- and Short-Term Memory (BiLSTM), Gated Recurrent Units (GRUs), and the integration of attention mechanisms of Soft attention, hard attention, dual attention, and Multi-Head Self-Attention (MHSA). The study critically examines the architectures' contributions to enhancing detection accuracy, recognition, and the capability to prevent crime. The review further highlights challenges associated with existing systems, including data scarcity, privacy concerns, computational complexity, data class imbalance, and limited real-world adoptability. The paper finally outlines emerging research gaps and future directions in the development of intelligent deep learning-based surveillance crime detection systems.

Introduction

Rapid population growth, urbanisation, and socio-economic inequalities have increased criminal activity, prompting security practitioners to prioritise crime detection, recognition, and prevention systems to enhance public safety (Anser et al., 2020). Crime is a global phenomenon that adversely affects peace and security, negatively impacting emotional, social, and economic development (Jonathan et al., 2021). Globally, over 440,000 people die UNODC, (2019), and property of over $51 billion Heeks et al., 2018) is lost yearly due to crime-related activities.

Several security measures have been employed to improve human and property safety: door locks, passive sensor alarm systems, and closed-circuit television systems. Studies have reported that

constant human supervision is paramount to avert insecurity problems (Wolthers, 2013). However, physical security guards are ineffective, have high labour costs, are subject to human subjectivity, and lack real-time Surveillance. This led to slower response times and the failure to detect crucial security threats (Tsakanikas & Dagiuklas, 2018). Additionally, passive sensor alarm systems have a major limitation: frequent false-positive detections, undermining their applicability in real-world scenarios.

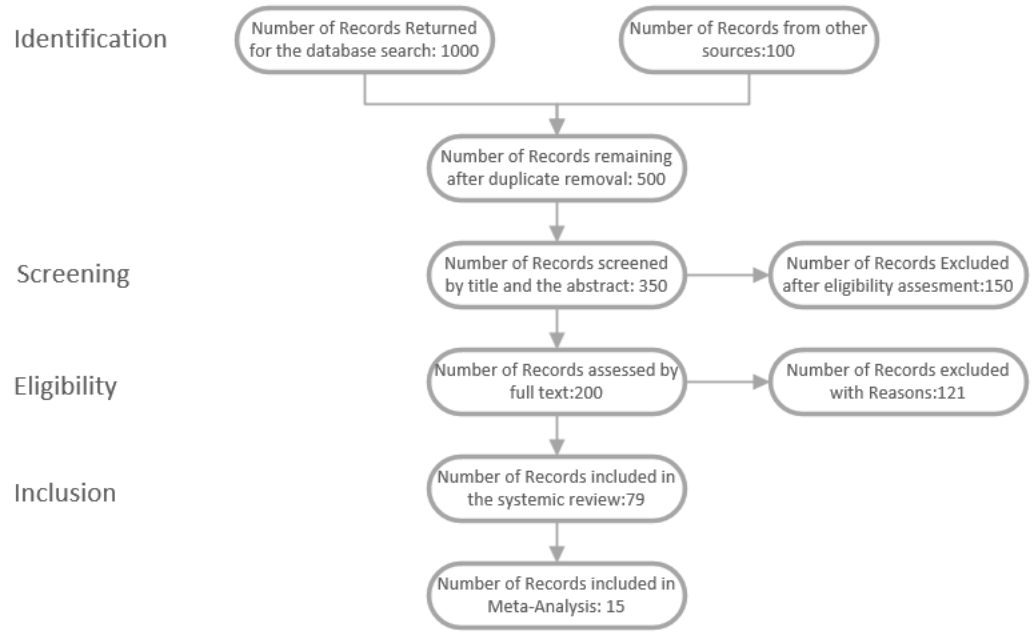
Video-based crime detection surveillance systems were designed to analyse surveillance videos for automatic detection and recognition of criminal activity and facilitate an informed and timely response. Machine learning systems are tailored to perform low-level tasks such as tracking, motion detection, and object identification. However, they have shown high false-positive rates in adverse environments (Ouardirhi et al., 2024). The complexity of criminal activities such as fighting, vandalism, arson, and robbery is unpredictable, highly associated with intra-class variability, and often remains visually subtle until escalation (Zhao et al., 2024). Furthermore, surveillance footage is affected by camera resolution, angle, crowd nature, background complexity, and weather conditions (Jeon et al., 2024). Consequently, this field has attracted significant research interests with substantial scholarly work and notable scientific publications

In light of this growing research interest, a systematic review of deep learning approaches for crime detection in surveillance systems is paramount. The review synthesised existing knowledge, evaluating the strengths and limitations of the current methodology, identifying prevailing trends, and providing future research directions. For transparency, reproducibility, and methodological rigour, this review adhered to Page et al. (2021). Further, the review analyses commonly used datasets and evaluation metrics, compares the performance of various architectures, and identifies emerging trends that will shape the future of intelligent security surveillance systems. The findings of the review are geared toward providing valuable insights for policymakers, researchers, security agencies, and practitioners who are targeting the development of robust, accurate, and efficient crime detection and recognition models that will enhance public safety.

Materials and Methods

We conducted the review in accordance with the PRISMA guidelines (Page et al., 2021), as shown in Figure 1.1. Literature search spans several scientific databases, including Scimago, Scopus, Web of Science, and Google Scholar and is limited to peer-reviewed journal articles, conference proceedings, and reports published between 2015 and 2025.

Figure 1: PRISMA Flow Diagram



The search employed the following combination of keywords and Boolean operators;

- “Crime Detection” AND “Crime action Recognition” AND “Video Surveillance systems”
- “Machine Learning” AND “Deep Learning”
- “2D-CNN” OR “3D-CNN” OR “3D-ResNet” OR “RNN” OR “LSTM” “OR, "BILSTM”, OR “GRU”
- “Spatial analysis “AND “Temporal analysis”, AND “Attention mechanism”

The included studies met the following criteria;

a. Focused discussion on crime detection and recognition in video surveillance.
b. Utilised various deep learning techniques.
c. Reported the model results and performance metrics.
d. Research published in peer-reviewed journals written in the English language.

Similarly, the review excluded studies that fell in the following category;

a. Addressed crime prediction rather than surveillance-based crime detection and recognition.
b. Lacked sufficient methodological details or experimental validation.
c. The full text of the publication was not available

The relevant information extracted focuses on the author and year of publication, the Crime category handled, the dataset utilised, the deep learning architecture employed, the performance metrics, key findings, and the architecture limitations, as indicated in Table 1.1.

Evolution of traditional feature engineering to deep learning

Initial video-based crime detection systems primarily relied on handcrafted feature engineering performed by domain experts. They designed descriptive visual and motion features to represent the underlying scene and human behaviour (Gibert et al., 2022). Various techniques, including Histogram of Oriented Gradients, Background Subtraction methods, Colour Histogram, Scale-Invariant Feature

Transform and Local Binary Pattern (Zhou et al., 2018), were used to extract low-level features such as edges, texture patterns, corners, optical flow, trajectory and silhouettes. The spatiotemporal descriptors used for motion analysis included Motion History Images, Space-Time Interest Points, and Optical Flow Histogram (Zhou et al., 2018). The handcrafted features were classified using renowned machine learning algorithms: Support Vector Machines, K-Nearest Neighbours, Hidden Markov Models, and Gaussian Mixture Models. However, the system exhibited high false-positive rates in adverse environments (Ouardirhi et al., 2024), rendering traditional feature engineering techniques ineffective. Furthermore, these systems were subject to delays and were reactive; the footage could only be reviewed after the incident had occurred, limiting their ability to prevent or reduce the impact of crime (Gibert et al., 2022). Moreover, the larger volume of video data collected from several high-definition cameras further strained its scalability and operational efficiency (Emmanuel Cadet et al., 2024).

The advent of computer vision marked a major technological shift in security surveillance systems. The enormous capabilities of CNNs enabled the system to automatically detect, recognise, and predict criminal activities within the viewshed in real-time (Apene et al., 2024), demonstrating robustness across varied environments.

Spatial-temporal modelling

Spatiotemporal modelling significantly enhances the performance of modern computer vision surveillance systems in action detection and recognition. It provides information on what happened at the scene, where it occurred, and the evolution of events over time. Effective detection and recognition of sophisticated criminal behaviours in frames depend on proper modelling of the spatial appearance of objects, people, or the scene layout, and Temporal dynamics of motion, interactions, and event progression (Natha et al., 2025).

Spatial modelling focuses on extracting meaningful visual features from each video frame, including objects, people, scene context and interactions. In video surveillance for crime detection, spatial cues such as human posture, object presence, clothing, sudden aggression, and individual proximity provide critical indicators for crime recognition. CNNs have become predominant in automatic, hierarchical feature extraction by combining both low-level features and high-level semantic concepts (Natha et al., 2025). To enhance accuracy, the system localises events within the scene, differentiates foreground and background activities, and clusters them across diverse environments of light intensity, viewpoint, and occlusion (Senussi et al., 2025).

Temporal modelling captures the motion patterns and dynamics of the activities across consecutive frames over time (Natha et al., 2025). It is the building block for recognising criminal actions within frames, since complex criminal activities cannot easily be recognised in a single frame. Additionally, it enables the system to learn short-term motion cues, such as sudden aggressive gestures, and long-term dependencies, like repeated suspicious behaviour.

Deep Learning in Video Analysis

Deep learning drastically changed the feature learning approach by providing automatic, hierarchical learning of high-level feature representations from surveillance videos using both 2D and 3D CNNs, such as 3D ResNet combined with LSTM and GRU (Ghauri et al., 2025). Further enhancement through spatial-temporal feature learning enables high recognition accuracy in complex scenarios such as crime and criminal activities. Various video datasets have been used in training these deep learning models, as indicated in Table 1.1.

Convolutional Neural Network (CNN)

CNNs are renowned for their ability to extract automatically relevant features from images without manual feature engineering. Its architecture has undergone technological enhancement and integration with other deep learning architectures to suit the world's dynamic needs (Noor Unnisa et al., 2025). The performance of the architecture can be enhanced through a combination of 2D and 3D CNNs. Further integration with deep residual architectures such as ResNet 50, ResNet 101, and their hybrid versions can greatly enhance performance (Noor Unnisa et al., 2025). This provides powerful architectures that perfectly learn and represent spatiotemporal features from crime video frames, as indicated in Figure 1.1

Figure 2: Convolutional Neural Network (CNN)

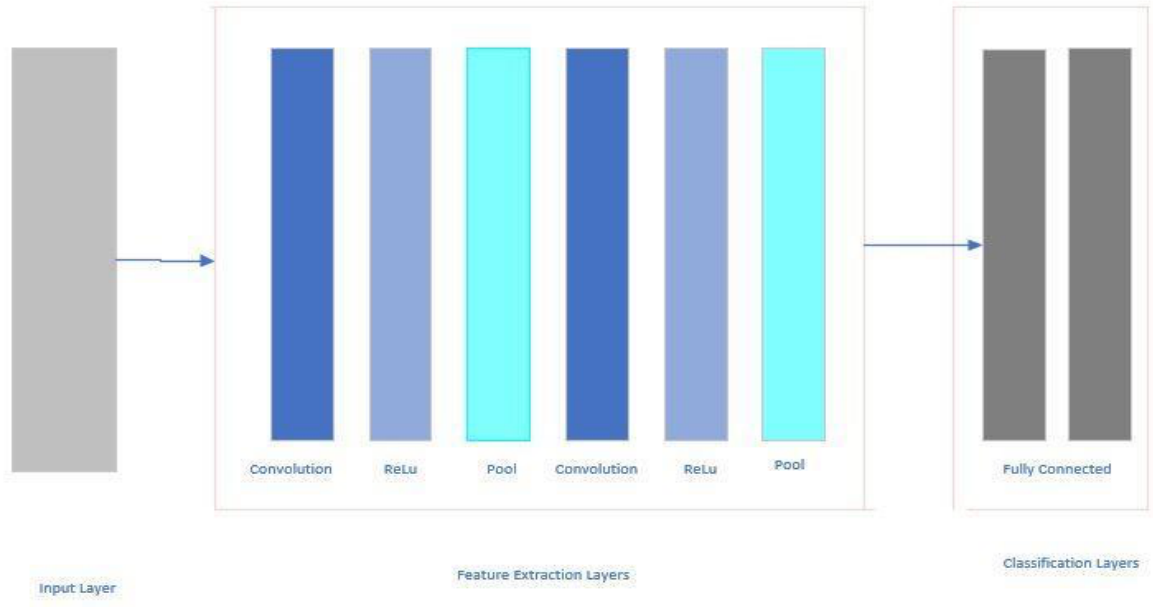


Table 1.1: Summary of Research Articles Reviewed

S/N	Authors, publication and year	Deep Learning architecture adopted	Dataset Utilised	Crime Category Addressed	Performance Metrics (Accuracy)	Key Findings and limitations
1	(Al-Madani et al., 2023)	3D-CNN+ BiLSTM	NLH Hocky, Movies Fights (MFS), Violent Flows (VFs), 1000 YouTube compilation of violent and non-violent videos.	Fighting	99.4	Performance of Models with self-Attention mechanisms outperformed those without. Limited crime coverage to fighting alone
2	(An et al., 2025)	ResNet18, ResNet50, EfficientNet, ViT, RetinaFace.	CRxK-6	Assault, intoxication, Robbery, Kidnapping, Burglary.	94.0	Higher accuracy compared by CNN over ViTs models. Vits,EfficientNet and ResNet18 performed well on Precision, Recall and Fi-Score. Lower precision (78%) and Fi-Score (86)for Robbery frames. Spatial-temporal Action localization.
3	(de Paula et al., 2022)	DNN-CNN and RNN	UFC-Crime, CamNuvem	Robbery	88.75%(AUC)	CamNuvem dataset are weakly labelled. innovative dataset required.
4	(Gandapur, 2022)	CNN, BiGRU, SRU	CAVIAR	Criminal activity	98.86%(accuracy)	BIGRU enhance detection and classification. No crime specificity.
5	(Ghauri et al., 2025)	DenseNet201+ LSTM	UCF-Crime	Robbery,fighting, shooting, Arrest, Arson, Assault	88.0%(AUC)	Dataset diversity enhancement, Incorpora attention mechanism and hybrid architectures. High false-positive.
6	(Jeon et al., 2024)	CLIP model.	ABODA and FireNet KISA, CCTV.	Intrusion, Loitering, Abandonment, Arson	99.08%(F1 - Score)	High false positives.

7	(A. Kumar et al., 2024)	MobileNetV, VGG16, BiLSTM	Violence situations	Violent and non-violent	93.25% (Accuracy)	High performance. Enhancement to complex situations handling.
8	(M. Kumar et al., 2024)	CNN, RNN, Attention mechanism	UCF50, UCF110, UCF	abnormal activity	Accuracy 96.94% (UCF50), 98.95%(UCF110) 62.04% (UCF)	Attention mechanisms Perfomance enhanced. Further research in aberrant human behaviour detection in complext environment.
9	(Londoño Lopera et al., 2025)	3D-CNN, LSTM	RWF2000, LAD2000,RVS, Ubi-Fights-urban Fights, UCF-Crime	robberies, assaults, and physical altercations	94%(Accuracy)	Dataset challenges of representativeness, class balance and precise labelling. Implementation of real-world tests to validate model effectiveness in complex conditions.
10	(Lu et al., 2025)	LCRNet, SAS MHSA	crime data in Los Angeles	Battery, Robbery, theft Vandalism, violent	97.76%(Accuracy)	Enhance the model interpretability and validating adaptability in resource-constrained environments.
11	(Mahmoodi & Nezamabadi-Pour, 2025)	Optical Flow and 2D-CNN	Hockey Fight, Violent Flows dataset, Action Movies , Surveillance Fight.	violent activity	Accuracy 98.54% (Hockey fights), 94.19% (Surveillance Fight), 100%(Action Movies)	High false-positive rate in distinction of violent and non-violent videos
12	(Mehmood, 2021)	CNN, optical flow stream	UFLV-DS ,URF-DS, Multicam-DS, Caviar-DS, (UMN-DS).	falling, loitering, violence	99%(falling), 97%(Loitering), 98%(Violence)	Spatiotemporal mechanisms enhaced performance. Further studies to include more abnormal behaviours.

13	(Nahar et al., 2022)	DenseNet, BiConvLSTM	hockey fight, movie fight, violent flows d, RWF-2000.	violent and nonviolent actions	99.50%, 97.50%	3D CNN integration enhanced performance. Modification of the network input (optical flow. pseudo-optical flow), integrate attention mechanism.
14	(Mukto et al., 2024)	YOLOV5 and MobileNetV2	bespoke	Weapon/violence detection, and face recognition	Accuracy 80% weapon 95% violence, 97% face	Better performance but struggled in areas with poor light illumination, hard to detection actions from low quality images, unable to detect holstered or completely concealed weapons.
15	(Natha et al., 2025)	DenseNet201 and BiLSTM	UCF, RAD	Road accidents, fighting, snatching, fires,explosions, stealing	86.2% accuracy	Enhanced detection accuracy, resource efficiency and minimised response time.

Two-Dimensional Convolutional Neural Networks (2D-CNNs)

2D-CNN architectures process single video frames as static images, representing each frame as a 3-Dimensional tensor (Kim & Lee, 2023). H denotes Height, W denotes width, and C denotes colour, as indicated in Eq. 1.1 below

$$H \times W \times C \tag{1.1}$$

The convolutional layers apply learnable features to capture local spatial patterns, followed by a non-linear activation function and a pooling layer operation. Feature extraction in 2D-CNNs automatically learns hierarchical spatial features from individual crime video frames, progressing from low-level visual cues to middle-level features to high-level semantic representations (Parmar & Morris, 2021).

The low-level cues detected by early layers include edges, corners, colour gradients, and texture. On the other hand, the middle-level feature detection by the middle-level layers includes objects, parts, silhouettes, and body posture. Finally, the high-level features represented by the deep layers include a complete semantic object representation and contextual cues, i.e., encoding what the object is, such as a person, fire, vehicle, or weapon (Parmar & Morris, 2021). The major limiting factor in the adoption of 2D-CNNs for video crime action detection and recognition is the inability to capture critical temporal relationships within a video frame. This is a crucial factor in the recognition and distinction of varied criminal activities. To complement 2D-CNN, the architecture is combined with other models, such as RNNs or Optical flow, to capture temporal dependencies within a video frame.

Three-Dimensional Convolutional Neural Networks (3D-CNNs)

The 3D-CNN architectures enhance the convolution operation to the temporal dimension by processing frames sequentially. Input data is represented as a 4D tensor in the form of Eq. 1.2 below, where T denotes temporal depth

$$T \times H \times W \times C \tag{1.2}$$

The 3D-CNN convolutional kernel operates across space and time, enabling the modelling of temporal dynamics. It extracts spatiotemporal features consisting of appearance and movement from video frames at three levels: low-level features (early layers) that capture local spatial features and short-term temporal dynamics, such as sudden limb movements or posture changes (Mehmood, 2021). The convolutional intermediate layers model interactions among multiple individuals (external temporal dynamics). Finally, the deep layers learn high-level crime video semantics, enabling recognition and distinction of different criminal activities (anomalies) from normal ones (Londoño Lopera et al., 2025). This makes the architecture suitable for crime action detection and recognition in surveillance videos by eliminating the use of handcrafted motion descriptors and enabling convolutional learning. However, the architecture struggles to learn long-range temporal dependencies beyond its fixed temporal receptive fields (Choukhairi et al., 2024). This is a critical drawback that technically impedes the adoption of single-module 3D-CNNs for crime action detection and recognition. Additionally, the architecture depicts high computational cost and requires a larger dataset to learn well.

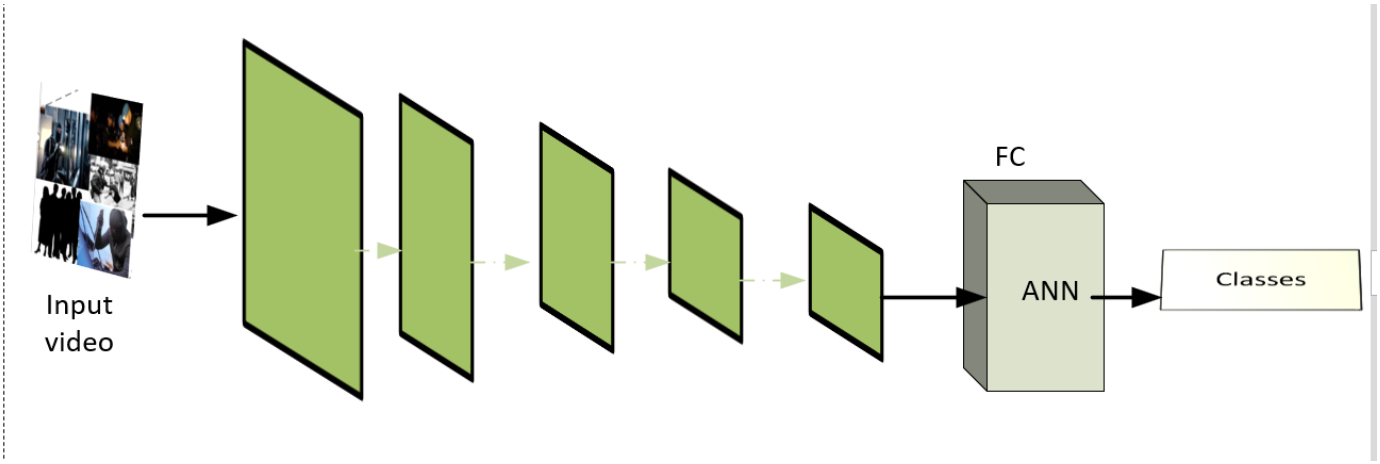
ResNet architectures

ResNets, designed for training very deep architectures through residual learning, helped address the vanishing gradient and degradation challenges in CNNs when the number of convolutions is increased (depth increase). Residual learning introduced a relatively unique way of learning features by introducing identity shortcut connections that bypass one or more convolutional layers. Rather than a direct learning map, the architecture learns a residual function that improves the architecture's gradient flow, speeds up convergence, and enhances training stability (Mao et al., 2024), as indicated in Eq. 1.3.

$$H(x) = F(x) + x \qquad (1.3)$$

Residual learning architectures play a significant role in crime action detection by enabling the training of deep spatiotemporal networks without degradation. This, therefore, enhances generalisation in complex surveillance contexts, supports transfer learning on large video datasets, and reuses features across layers. The 3D ResNet architecture is shown in Figure 1.2.

Figure 2: 3D ResNet Architecture of Crime Detection

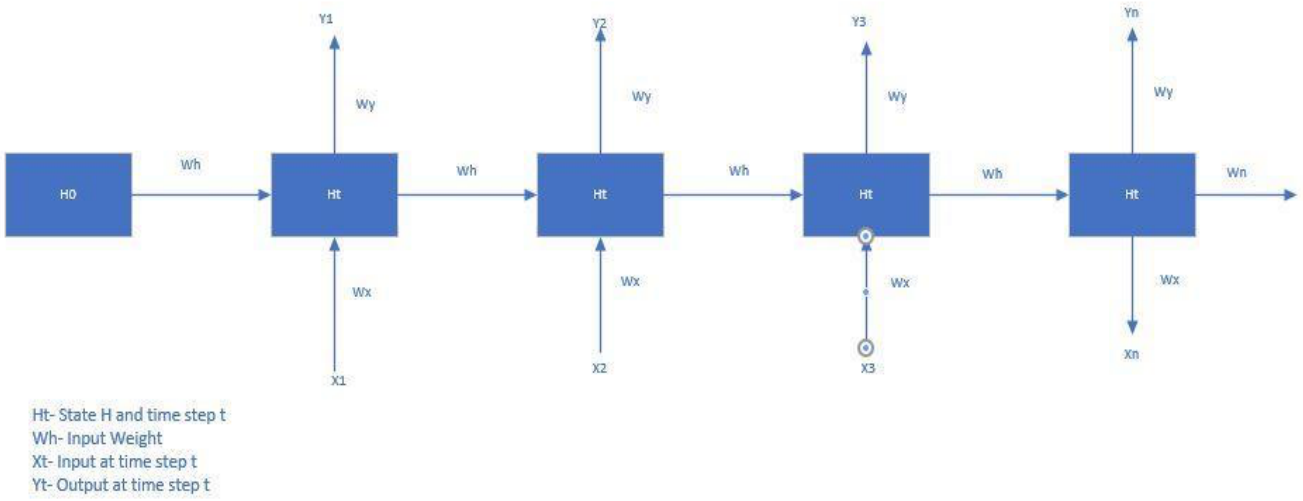


Two residual network architectures used in numerous crime detection applications in surveillance systems are ResNet-50 and ResNet-101. ResNet-50 architecture is made up of 50 layers that are structured in bottleneck residual blocks of $(1 \times 1, 3 \times 3, 1 \times 1)$ *convolutions*. ResNet-50 is used as the foundation for extracting spatial features due to its enhanced capability to balance the architecture's computational efficiency, depth, and accuracy (Udhaya et al., 2025). The ResNet-101 architecture is deeper, with 101 layers, enabling a more compressed hierarchical feature representation. It captures finer spatial details critical to accurate crime action recognition; however, additional layers increase the architecture's computational cost (Udhaya et al., 2025). The adoption of ResNet Architectures in the temporal domain enhances the extraction of robust spatiotemporal features from short video frames, necessary for accurate crime action recognition.

Recurrent Neural Networks (RNN)

Criminal activity detection and recognition require effective, robust modelling of temporal dynamics within frames. Distinctive action recognition can be achieved by integrating CNN architectures with an RNN proficient in modelling temporal relationships across frames over time. RNNs utilise time-series data for sequence learning and are renowned for modelling long-term temporal dependencies within frames (Ahmed et al., 2022). The edges provide input to the next time step for pattern recognition within a given input function, pass it through a hidden function, and produce output through an output function. The design structure of an RNN has an Input function Xt, a hidden-state function Ht, and an output function Yt. A recurrent Neural Network structure is designed to identify and recognise sequences within a given dataset over time, as shown below (Ahmed et al., 2022).

Figure 3: Recurrent Neural Network Architecture

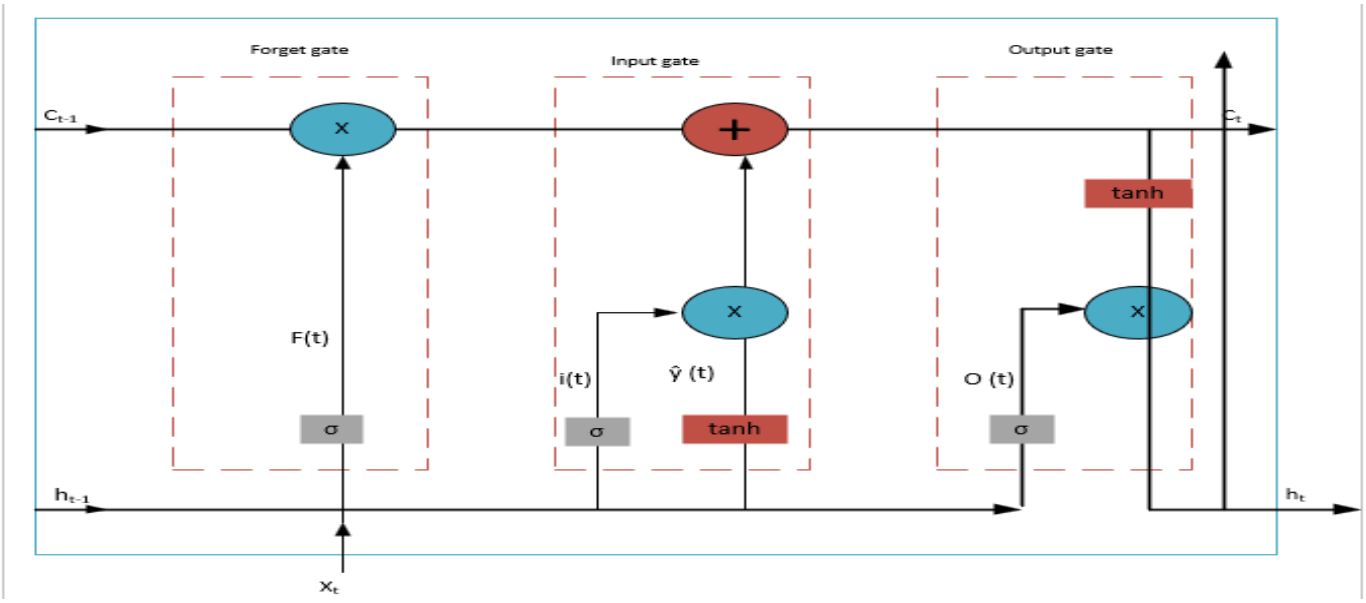


The dataset undergoes a sequence of computational processing in the intermediate hidden layers before being passed to the output layer. It has cycles within the network that serve as short-term memory. RNNs enable both sequential and parallel computation, resembling the human brain. An RNN architecture is best suited for function modelling, with the input and output constituting time-vector dependencies (Mohanty & Mohanty, 2022). These networks have loops in their connections that enable them to model temporal behaviour and achieve accuracy in domains such as language, time-series, text, and audio. The networks may establish an additional matrix that connects time steps and is trained to identify temporal relationships in the dataset. RNNs are classified into RNN cells and RNN network topologies. In the RNN cell taxonomy, there are simple RNN cells, LSTM cells, and GRU cells. Simple RNN cells constitute a simple recurrent input, with each cell representing an artificial neuron with inputs, biases, and outputs. LSTM and GRU resemble simple cells with additional gating functions that allow memory to persist and provide greater control (Krig, 2016).

Long-Short Term Memory (LSTM)

The LSTM network consists of three major aspects that distinguish it from the usual neuron in an RNN. These aspects include controls on when to allow input to enter the neuron, when to retain the previous computation at the previous time step, and when to allow the output to progress to the next time step, as indicated in Figure 1.5. LSTM was introduced to overcome the limitations of the Vanilla RNN architecture; it utilises gating mechanisms (Yadav & Thakkar, 2024).

Figure 4: LSTM architecture



A renowned aspect of the LSTM is that its decision is based solely on the current input. LSTM linearly integrates weights and bias over inputs and contains gates (Training capability) which determine when and how much the integrator listens to inputs to determine the output (Yadav & Thakkar, 2024). They regulate the flow of information, guided by the input, forget, and output gates indicated by the equations 1.4 to 1.7below.

$$f_t = \sigma\left(W_f[x_t, h_t^{-1}] + b_f\right) \quad (1.4)$$

$$i_t = \sigma(W_{\text{i}}[x_t, h_t^{-1}] + b_{\text{i}}) \quad (1.5)$$

$$o_t = \sigma(W_{\text{o}}[x_t, h_t^{-1}] + b_{\text{o}}) \quad (1.6)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot ReLU(W_c[x_t, h_t^{-1}] + b_c) \quad (1.7)$$

The LSTM has proven efficient in modelling long-term temporal dependencies, capturing contextual and gradual behaviour changes within video frames, and detecting complex crime sequences (Yadav & Thakkar, 2024). An advanced version of LSTM is the GRU (Gated Recurrent Unit) network architecture (Cho, Van Merriënboer, Gulcehre, et al., 2014), which simplifies backpropagation tuning by using two control gates, i.e, the input gate and the dynamic gate, as indicated by the equations below.

$$z_t = \sigma(W_z[x_t, h_t^{-1}]) \quad (1.8)$$

$$r_t = \sigma(W_r[x_t, h_t^{-1}]) \quad (1.9)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot h_t \quad (1.10)$$

GRU has some advantages over LSTM, including fewer parameters and faster training and inference. Finally, GRU is suitable for real-time and edge computing. The Bidirectional RNNs (BiLSTM and BiGRU) process sequences in both forward and backward directions, enabling the model to exploit past and future context for accurate action prediction simultaneously (Yadav & Thakkar, 2024). This architecture boosts the capability to capture long-range temporal patterns and forms the ideal model

for capturing sequences with complex temporal dynamics, while being robust to vanishing gradients due to gated memory (Yadav & Thakkar, 2024).

Attention mechanism in deep learning network architectures

Attention mechanisms in deep learning architectures have become a crucial component in enhancing models' effectiveness and robustness. They enable the architecture to focus on crucial informative parts of the input data while ignoring irrelevant information. The integration of attention mechanisms improves the learning efficiency, interpretability, and accuracy of computer vision tasks. These tasks primarily encompass image detection and recognition, natural language processing, and action detection and recognition in video frames (Shuvo et al., 2025). Attention mechanisms play a critical role in identifying the evolution of silent spatiotemporal features necessary for distinctive crime action recognition. Implementation of attention mechanisms in deep learning models can be done through Soft attention, Hard attention, Dual attention and Multi-Head Self-Attention (MHSA). The soft attention architecture operates on the principles of determinism and differentiability by assigning continuous attention weights to all input data elements. The weighted sum of features is computed, with the most relevant features given higher weights. The attention weights are computed using a compatibility function followed by a SoftMax function (Kim, 2025), as indicated by Eq. 1.11 below, where ei represents the relevance score of the i-th feature.

$$\alpha_j = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_i)} \tag{1.11}$$

The soft attention model architectures are fully differentiable and trainable via backpropagation, enabling simultaneous attention to all regions of a frame with varying emphasis. This makes the architecture computationally efficient and stable for action recognition in video frames, such as detecting aggressiveness and suspicious interactions. The hard attention architecture emphasises the stochastic mechanisms by selecting a discrete region of a frame to concentrate on in a non-differential process and training using reinforcement learning (Kim, 2025). Hard attention has limited the use of the video surveillance system, since the architecture has to make hard decisions on region selection, further complicating the training process, leading to high variance in gradient estimation, and slower convergence.

The dual attention model architectures simultaneously capture several complementary aspects of the dataset, including both spatial and temporal attention in video frames. The dual attention architecture comprises the spatial attention module, which concentrates on important regions in each frame, and the temporal attention module, which focuses on critical frame dynamics (Kim, 2025). These functionalities make the dual attention model an ideal architecture for crime action detection and recognition in surveillance systems. Finally, the Multi-Head Self-Attention (MHSA) model architecture employs an improved approach to computing attention within a single sequence by relating different elements to each other. This allows the architecture to capture long-range dependencies and contextual dynamics. For any given input, feature computation takes three dimensions: Query (Q), Key (K), and Value (V) (Lu et al., 2025) as indicated in Eq. 1.12.

$$Attension(QKV) = softmax\left(\frac{QK^t}{\sqrt{d_k}}\right)V \tag{1.12}$$

MHSA performs multiple attention operations in parallel, each with a different learnable projection, focusing on different aspects of the input dataset. MHSA model architecture outperforms other architectures by modelling long-range spatiotemporal dependencies, a major drawback of CNNs and

RNNs architectures. The architecture also demonstrates superiority in parallel feature representation, improved interoperability, high scalability, efficiency, and robustness to occlusion and noise.

Hybrid deep learning architectures

The integration of various hybrid deep learning architectures has performed significantly better than single-module architectures in crime action detection and recognition. The integration of 3D-CNNs with transformer-based self-attention mechanisms and sequential models provides an enhanced architecture for capturing multi-scale motion patterns and long-range dependencies (Zhang et al., 2025). 3D-ResNet has shown state-of-the-art performance in video classification due to its residual learning design, which enables deep hierarchical feature representation from consecutive video frames extraction and also solves the vanishing gradient challenge. However, it struggles to model long-range temporal relations in complex criminal activities (Gong & Luo, 2025). The integration of the MHSA mechanism addresses these limitations by enabling the model to focus on the most significant spatial and temporal features simultaneously (Lu et al., 2025) while ignoring background noise. Further, Criminal activities exhibit temporal evolution, which requires analysis in time-series. The architecture is complemented by Bidirectional Long-Short-Term Memory (BiLSTM), which learns both forward and backward dependencies within a video frame. This facilitates a finer understanding of the progression of events.

Therefore, the combination of 3D ResNet, MHSA, and BiLSTM provides an architecture that delivers state-of-the-art performance in crime detection and addresses multifaceted challenges in prior models. The architecture leverages the enhanced capability of 3D ResNet for the extraction of deep spatial and short-term temporal features (Udhaya et al., 2025), the MHSA architecture for global spatial-temporal attention and enhanced interoperability (Lu et al., 2025), and finally the BiLSTM architecture for robust and effective long-term temporal modelling (Yadav & Thakkar, 2024).

The integration of the three architectures into a single model enhances detection and recognition performance, reduces false-positive alarms, and improves generalisation to complex surveillance conditions (Panda et al., 2022; Udhaya et al., 2025)

Results

The review revealed that integrating deep learning architectures significantly improved the performance of security surveillance systems.

Enhance crime detection and recognition accuracy

Deep learning models achieved higher accuracy and precision in detecting and recognising crimes compared to traditional systems and machine learning models. The automated systems accurately identified crimes, including Robbery, Arson, fighting and vandalism.

Effective spatial feature extraction and temporal dynamic learning

ResNets were more effective at extracting spatial features from frames by effectively learning visual patterns associated with criminal activities. The temporal dynamics were effectively modelled using bidirectional long-short-term memory (BiLSTM), which processes information in both forward and backward directions.

Superiority of improved deep learning architecture (3D ResNet)

Residual Networks (ResNets), with enabled deeper neural networks, outperformed other deep learning architectures in feature representation and classification. Additionally, they were effective at addressing the vanishing gradient problem in CNN architectures.

Attention mechanism
The integration of the MHSA mechanism enabled the model to focus on the most relevant cues while ignoring background noise. This enhances the model's accuracy in crime detection and recognition.

Discussions
The integration of deep learning architectures into surveillance systems significantly enhanced crime detection, recognition, and prevention. The study demonstrated higher accuracy, efficiency, scalability, and real-time analysis in the nexus of deep learning models, machine learning models and legacy surveillance systems. The enhanced performance is attributed to the architecture's ability to automatically learn spatial and temporal features from frames.

Despite CNNs' limitations in capturing temporal dynamics and motion cues in frames, they are critical building blocks of modern surveillance systems, owing to their strong capability for hierarchical spatial feature extraction. Studies in CNN-based models often reported high accuracy in object detection, anomaly detection, and object and/or weapon recognition.

The study further showed that Residual networks (ResNet50 and ResNet101) exhibited superior feature extraction capabilities over other CNN-based models due to their deeper neural architectures, which addressed the vanishing-gradient challenge. Additionally, it achieved a higher feature representation and classification accuracy. However, CNNs struggle to model long-term temporal dependencies in complex criminal activities. This lowers the performance accuracy in crime detection and recognition. The improved models, such as SlowFast Networks, C3D, I3D and 3D-ResNet architecture, extend the CNNs' capability through their simultaneous extraction of spatial and temporal features from consecutive video frames (Mao et al., 2024). However, the real-world deployment of these architectures has faced several setbacks: require higher computational resources, memory capacity and longer training time. Furthermore, for optical performance, 3D-CNN models require large, high-quality annotated video datasets, which are constrained by privacy, ethical, and legal requirements.

The integration of recurrent architectures, including RNN, LSTM, BiLSTM and GRU, into 3D-CNNs improved the understanding of temporal dynamics (Ahmed et al., 2022). BiLSTM enables analysis of video data in both forward and backward directions. The integration of DL with various attention mechanisms, including soft attention, hard attention, dual attention, and MHSA, have greatly improved model performance. The MHSA focuses on the most informative cues in the surveillance video while ignoring the irrelevant background noises. This further enhances the model's robustness in accurately detecting criminal activities, especially in adverse environments.

Hybrid architectures outperform legacy architectures; however, the introduction of the attention mechanism increases computational overhead and thus requires careful optimisation to achieve optimal performance. The research indicated that the hybrid deep learning architectures demonstrated the best overall performance across all surveillance tasks. Integration of 3D ResNet, MHSA, and BiLSTM leveraged the complementary strengths of each module. This further strengthens the spatial feature extraction, temporal dynamic modelling and adaptive attention. A comparative analysis showed a higher accuracy, precision, Recall, F1-Score, and AUC for the hybrid model than the legacy models.

Generalisation limitation of the controlled environment nexus in real-world scenarios
Most of the crime detection models are trained and validated on benchmark datasets that were well-curated, well labelled and collected under relatively controlled environments, as indicated by the most used datasets of the UCF anomaly crime dataset, the Avenue dataset, ShanghaiTech, and UCSD

(Dhuri & Patil, 2025). These datasets do not reflect the real-world context of surveillance environments, which are associated with variation in illumination, adverse weather conditions, occlusion, crowd dynamics, camera motion, resolution, and varied angles. The existing models therefore have limited robustness and performance in real-world environments compared to the controlled environments in which they are trained and tested.

Limited availability of high-quality crime-specific Datasets

The scarcity of high-quality, realistic, and diverse crime-specific datasets poses a critical research gap. This limits the models' ability to extract rich spatiotemporal features for accurate and precise crime action recognition. Crimes in the real-world are difficult to predict, context-specific, and often constitute highly sensitive data, thereby hindering the systematic collection of large-scale datasets. Existing datasets, including UCF-crime, suffer from class imbalance because they fail to capture the true nature of crime and the complex activities of criminals in real-world environments (Londoño Lopera et al., 2025). There exists inconsistent annotation of the existing dataset, where the majority of the available dataset does not provide temporal boundaries or spatial localisation of criminal actions. Instead, it provides weak video labels such as 'crime' or 'non-crime'. This limits the model's ability to learn where and when the crime occurred within video frames (Londoño Lopera et al., 2025). The manual annotation of the crime dataset is time-consuming, costly, and ethically sensitive.

Ethical Privacy and Bias

Minimal consideration has been given to enhancing ethical standards, privacy, and avoiding bias. Most research has concentrated on model performance and computational efficiency while ignoring the importance of the social rubric of ethical and social implications. Therefore, their adoption could meet their functional requirements but cause social and legal predicaments. More conspicuous is the invasive nature of continuous visual monitoring, which collects large amounts of personal data without the data subject's consent. The footage that contains identifying features of a data subject, including facial features, body features, and behavioural patterns, constitutes a serious privacy breach. However, there is limited research on privacy-by-design approaches to minimise the exposure of private data while maintaining effective crime detection surveillance systems (Okunola & Ashun, 2024).

Computational constraints and scalability

Computer vision surveillance system adoption and deployment in the real-world struggle to balance high computational demand with limited scalability. While many state-of-the-art deep learning models for surveillance systems have shown satisfactory performance in controlled environments, they rely heavily on high computational resources. Deployment in the real-world may not be feasible due to the high associated cost of computational resources and the sophistication of the algorithms (Zhu et al., 2025).

Conclusion

The integration of the improved architectures enhances detection and recognition and improves generalisation in complex surveillance conditions. The complexity of criminal activities necessitates the adoption and implementation of Intelligent and autonomous security systems to enhance public and private safety. Additionally, this technology will not only improve detection and recognition but also facilitate post-event analysis, criminal behaviour monitoring and tracking, accurate evidence retrieval, and prediction of crime event occurrence. This is achieved through minimised involvement of human operators, enhanced situational awareness, and enabling data-driven decision-making. This, therefore, makes deep-learning-based surveillance systems essential components of modern security ecosystems. The deployment of these sophisticated architectures can, however, be enhanced

through technological advancements, including improved GPU computing, Edge computing, AI, and real-time streaming capabilities. Furthermore, well-curated, high-quality, and large datasets for training and testing the models are necessary. Finally, proper policy and legal challenges should be reviewed to provide a leeway for the development of a robust crime dataset that will simulate the real-world environment.

References

Ahmed, D. M., Hassan, M. M., & Mstafa, R. J. (2022). A review on deep sequential models for forecasting time series data. *Applied Computational Intelligence and Soft Computing*, *2022*(1), 6596397.

Al-Madani, A. M., Mahale, V., & Gaikwad, A. T. (2023). Real-Time Detection of Crime and Violence in Video Surveillance using Deep Learning. *First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022)*, 431–441.

Alshalawi, A., Abdul, W., & Muhammad, G. (2025). Fire Swin: Video anomaly detection using a hybrid model with convolutional layers, fire module, and Swin Transformer. *Journal of Engineering Research*. https://doi.org/10.1016/j.jer.2025.08.016

An, C., Lee, M., & Park, E. (2025). CRxK dataset: a multi-view surveillance video dataset for re-enacted crimes in Korea. *Scientific Reports*, *15*(1), 28946. https://doi.org/10.1038/s41598-025-15058-w

Anser, M. K., Yousaf, Z., Nassani, A. A., Alotaibi, S. M., Kabbani, A., & Zaman, K. (2020). Dynamic linkages between poverty, inequality, crime, and social expenditures in a panel of 16 countries: two-step GMM estimates. *Journal of Economic Structures*, *9*(1), 43. https://doi.org/10.1186/s40008-020-00220-6

Anser, M. K., Yousaf, Z., Nassani, A. A., Alotaibi, S. M., Kabbani, A., & Zaman, K. (2020). Dynamic linkages between poverty, inequality, crime, and social expenditures in a panel of 16 countries: two-step GMM estimates. *Journal of Economic Structures*, *9*(1), 43.

Apene, O. Z., Blamah, N. V., & Aimufua, G. I. O. (2024). Advancements in crime prevention and detection: From traditional approaches to artificial intelligence solutions. *European Journal of Applied Science, Engineering and Technology*, *2*(2), 285–297.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *ArXiv Preprint ArXiv:1409.1259*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *ArXiv Preprint ArXiv:1406.1078*.

Choukhairi, O., Choukhairi, M., Choukri, A., & Fakhri, Y. (2024). A Literature Review On Semantic Segmentation Of 3D Temporal Data Using Neural Networks. *2024 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, 1–8.

de Paula, D. D., Salvadeo, D. H. P., & de Araujo, D. M. N. (2022). CamNuvem: A Robbery Dataset for Video Anomaly Detection. *Sensors*, *22*(24), 10016. https://doi.org/10.3390/s222410016

Dhuri, K. G., & Patil, S. (2025). Crime prediction against women in surveillance videos using deep learning models: A review. *Computers and Electrical Engineering*, *128*, 110651.

Emmanuel Cadet, Olajide Soji Osundare, Harrison Oke Ekpobimi, Zein Samira, & Yodit Wondaferew Weldegeorgise. (2024). AI-powered threat detection in surveillance systems: A real-time data processing framework. *Open Access Research Journal of Engineering and Technology*, *7*(2), 031–045. https://doi.org/10.53022/oarjet.2024.7.2.0057

Francisco, G., & Brown, T. (2012). Fully integrated automated security surveillance system: managing a changing world through managed technology and product applications. *Unattended Ground, Sea, and Air Sensor Technologies and Applications XIV*, *8388*, 198–212.

Gandapur, M. Q. (2022). E2E-VSDL: End-to-end video surveillance-based deep learning model to

detect and prevent criminal activities. *Image and Vision Computing*, *123*, 104467. https://doi.org/10.1016/j.imavis.2022.104467

Ghauri, M. S., Bajwa, U. I., Saleem, G., Raza, R. H., & Anwar, M. W. (2025). SurveillanceNet: Spatio-temporal anomaly identification in surveillance videos using two-stream CNN and LSTM. *Multimedia Tools and Applications*, 1–25.

Gibert, D., Planes, J., Mateu, C., & Le, Q. (2022). Fusing feature engineering and deep learning: A case study for malware classification. *Expert Systems with Applications*, *207*, 117957. https://doi.org/10.1016/j.eswa.2022.117957

Gong, P., & Luo, X. (2025). A Survey of Video Action Recognition Based on Deep Learning. *Knowledge-Based Systems*, *320*, 113594. https://doi.org/10.1016/j.knosys.2025.113594

Heeks, M., Reed, S., Tafsiri, M., & Prince, S. (2018). *The economic and social costs of crime*. Home Office London.

Jeon, H., Kim, H., Kim, D., & Kim, J. (2024). PASS-CCTV: Proactive Anomaly surveillance system for CCTV footage analysis in adverse environmental conditions. *Expert Systems with Applications*, *254*, 124391.

Jonathan, O. E., Olusola, A. J., Bernadin, T. C. A., & Inoussa, T. M. (2021). Impacts of Crime on Socio-Economic Development. *Mediterranean Journal of Social Sciences*, *12*(5), 71. https://www.richtmann.org/journal/index.php/mjss/article/view/12607

Kim, H. (2025). *Attention Mechanisms in AI: A Comprehensive Overview*. https://medium.com/@donotgetgreed/attention-mechanisms-in-ai-a-comprehensive-overview-d4436c612562

Kim, S., & Lee, S.-P. (2023). A BiLSTM–transformer and 2D CNN architecture for emotion recognition from speech. *Electronics*, *12*(19), 4034.

Krig, S. (2016). Feature learning and deep learning architecture survey. In *Computer vision metrics* (pp. 375–514). Springer.

Leibovitch, A. (2021). A Wider Scope of Effects on Communities: Comment on Proactive Policing: Effects on Crime and Communities. *Jerusalem Review of Legal Studies*, *24*(1), 15–26. https://doi.org/10.1093/jrls/jlab003

Londoño Lopera, J. C., Bolaños Martinez, F., & Fletscher Bocanegra, L. A. (2025). Building a custom crime detection dataset and implementing a 3D convolutional neural network for video analysis. *Algorithms*, *18*(2), 103.

Lu, H., Chen, C., Ma, Y., & Ma, Y. (2025). Lightweight deep learning model for crime pattern recognition based on transformer with simulated annealing sparsity and CNN. *Scientific Reports*, *15*(1), 20311. https://doi.org/10.1038/s41598-025-07260-7

Lyon, D., & Wood, D. M. (2012). Security, surveillance, and sociological analysis. In *Canadian Review of Sociology/Revue canadienne de sociologie* (Vol. 49, Issue 4, pp. 317–327). Wiley Online Library.

Mahmoodi, J., & Nezamabadi-Pour, H. (2025). Violence Detection in Video Using Statistical Features of the Optical Flow and 2D Convolutional Neural Network. *Computational Intelligence*, *41*(2), e70034.

Mao, M., Lee, A., & Hong, M. (2024). Deep learning innovations in video classification: a survey on techniques and dataset evaluations. *Electronics*, *13*(14), 2732.

Mehmood, A. (2021). Abnormal behaviour detection in uncrowded videos with two-stream 3D convolutional neural networks. *Applied Sciences*, *11*(8), 3523.

Mohanty, M. D., & Mohanty, M. N. (2022). Verbal sentiment analysis and detection using recurrent neural network. In *Advanced Data Mining Tools and Methods for Social Computing* (pp. 85–106). Elsevier. https://doi.org/10.1016/B978-0-32-385708-6.00012-6

Ms. Noor Unnisa, Mohammed Ehtesham Ul baqui, Ms. Vibhavari N, Ms. Farheen Sultana, Shaik Irfan,

& Mohammed Mubashir Ul Baqui. (2025). Anomaly Detection Using Convolutional Neural Networks for Crime Prevention: A Deep Learning Approach. *International Journal of Latest Technology in Engineering Management & Applied Science*, *14*(6), 356–362.

Myagmar-Ochir, Y., & Kim, W. (2023). A survey of video surveillance systems in smart city. *Electronics*, *12*(17), 3567.

Nahar, J., Promi, Z. T., Ferdous, J., Ishrak, F., & Khurshid, R. (2022). *Anomalous behavior detection using spatio temporal feature and 3D CNN model for surveillance*. Brac University.

Natha, S., Ahmed, F., Siraj, M., Lagari, M., Altamimi, M., & Chandio, A. A. (2025). Deep BiLSTM Attention Model for Spatial and Temporal Anomaly Detection in Video Surveillance. *Sensors*, *25*(1), 251. https://doi.org/10.3390/s25010251

Okunola, A., & Ashun, A. (2024). *Ethical Considerations in AI-Powered Surveillance Systems: Balancing Security and Privacy in the Digital Age*.

Ouardirhi, Z., Mahmoudi, S. A., & Zbakh, M. (2024). Enhancing object detection in smart video surveillance: A survey of occlusion-handling approaches. *Electronics*, *13*(3), 541.

Panda, M. K., Sharma, A., Bajpai, V., Subudhi, B. N., Thangaraj, V., & Jakhetiya, V. (2022). Encoder and decoder network with ResNet-50 and global average feature pooling for local change detection. *Computer Vision and Image Understanding*, *222*, 103501.

Parmar, P., & Morris, B. (2021). HalluciNet-ing spatiotemporal representations using a 2D-CNN. *Signals*, *2*(3), 604–618.

Pouyan, S., Charmi, M., Azarpeyvand, A., & Hassanpoor, H. (2023). Propounding first artificial intelligence approach for predicting Robbery behavior potential in an indoor security camera. *IEEE Access*, *11*, 60471–60489.

Qian, Y., Ye, S., Wang, C., Cai, X., Qian, J., & Wu, J. (2025). *UCF-Crime-DVS: A Novel Event-Based Dataset for Video Anomaly Detection with Spiking Neural Networks*. http://arxiv.org/abs/2503.12905

Sahay, K. B., Balachander, B., Jagadeesh, B., Anand Kumar, G., Kumar, R., & Rama Parvathy, L. (2022). A real time crime scene intelligent video surveillance systems in violence detection framework using deep learning techniques. *Computers and Electrical Engineering*, *103*, 108319. https://doi.org/10.1016/j.compeleceng.2022.108319

Savitha N J and Lata B T. (2023). Deep Learning Approaches for Criminal Culprits Identification in Surveillance Systems. *Computer Research and Development*, *23*(12), 8.

Senussi, M. F., Abdalla, M., Kasem, M. S., Mahmoud, M., Yagoub, B., & Kang, H. S. (2025). A Comprehensive Review on Light Field Occlusion Removal: Trends, Challenges, and Future Directions. *IEEE Access*, *13*, 42472–42493.

Shafiq, M., & Gu, Z. (2022). Deep residual learning for image recognition: A survey. *Applied Sciences*, *12*(18), 8972.

Shah, N., Bhagat, N., & Shah, M. (2021). Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. *Visual Computing for Industry, Biomedicine, and Art*, *4*(1), 9.

Shuvo, M. R., Mekala, M. S., & Elyan, E. (2025). Deep learning and attention-based methods for human activity recognition and anticipation: a comprehensive review. *Cognitive Computation*, *17*(6), 1–28.

Simon Hall. (2025). *UCF Crime Images: Surveillance images extracted from UCF Crime Dataset*. https://www.kaggle.com/datasets/simonphall/ucf-crime-images

Stickle, B., & Felson, M. (2020). Crime rates in a pandemic: The largest criminological experiment in history. *American Journal of Criminal Justice*, *45*(4), 525–536.

Tsakanikas, V., & Dagiuklas, T. (2018). Video surveillance systems-current status and future trends. *Computers & Electrical Engineering*, *70*, 736–753.

Udhaya, T., Dhanush, R., & Jithender, S. (2025). Enhancing Home Security with Deep Learning-Based Smart Monitoring Using CNN and ResNet-50 for Danger Identification. *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, 1008–1013.

UNODC. (2019). *GLOBAL STUDY ON HOMICIDE*. https://www.unodc.org/documents/data-and-analysis/gsh/Booklet1.pdf

Wiliem, A., Madasu, V., Boles, W., & Yarlagadda, P. (2012). A suspicious behaviour detection using a context space model for smart surveillance systems. *Computer Vision and Image Understanding*, *116*(2), 194–209.

Wolthers, L. (2013). Self-surveillance and virtual safety. *Photographies*, *6*(1), 169–176. https://doi.org/https://doi.org/10.1080/17540763.2013.788852

Yadav, H., & Thakkar, A. (2024). NOA-LSTM: An efficient LSTM cell architecture for time series forecasting. *Expert Systems with Applications*, *238*, 122333.

Zhang, X., Nyamasvisva, T. E., & Liu, C. (2025). *A Framework Combining 3D CNN and Transformer for Video-Based Behaviour Recognition*. https://doi.org/https://doi.org/10.48550/arXiv.2508.06528

Zhao, R., Hua, F., Wei, B., Li, C., Ma, Y., Wong, E. S. W., & Liu, F. (2024). A review of abnormal crowd behaviour recognition technology based on computer vision. *Applied Sciences*, *14*(21), 9758.

Zhou, P., Ding, Q., Luo, H., & Hou, X. (2018). Violence detection in surveillance video using low-level features. *PLOS ONE*, *13*(10), e0203668. https://doi.org/10.1371/journal.pone.0203668

Zhu, J., Hua, L. S., Chen, B., Kou, W., Rong, J., Zhou, Q., & Sun, Y. (2025). YOLO-DCS: A high-precision edge detection method for latex bowl state and latex trail state. *Industrial Crops and Products*, *237*, 122310.

Zungu, N., Olukanmi, P., & Bokoro, P. (2025). SynthSecureNet: An Improved Deep Learning Architecture with Application to Intelligent Violence Detection. *Algorithms*, *18*(1), 39. https://doi.org/10.3390/a18010039