Open Access ISSN: 3007-1437 (Online), ISSN: 3007-1429 (Print)

IRAQI Academic Scientific Journals

Alkadhim Journal for Computer Science (KJCS)

Journal Homepage: https://alkadhum-col.edu.iq/JKCEAS

KJCS ALKADHIM JOURNAL FOR COMPUTER SCIENCE

Multi-Dimensional Deep Learning for Unusual Detection in Security Footage: A 3D CNN, LSTM, and Attention-Based Approach

[1]Nadia Mahmood Ali*

[1]Middle Technical University, Institute of Medical Technology Al-Mansur, Baghdad, Iraq

Article information

Article history:

Received: June,14, 2025

Accepted: August,23, 2025

Available online: september,25, 2025

Keywords:

Video Anomaly Detection,
3D Convolutional Neural Networks,
Fight Detection,
Surveillance Videos,
Violence Detection.

Corresponding Author:

Nadia Mahmood Ali

nadiam@mtu.edu.iq

DOI:
https://doi.org/10.61710/kjcs.v3i3.114

Abstract

Automated detection of unusual activities in surveillance videos remains a significant challenge due to the massive amount of recorded footage and the low occurrence of anomalous events. Here we report a novel deep learning framework designed to address this problem by integrating three core components: three-dimensional Convolutional Neural Networks (3D-CNNs) for extracting spatiotemporal features, Long Short-Term Memory (LSTM) networks for capturing sequential dependencies, and an attention mechanism for emphasizing the most salient regions of video data. The study aimed to design a robust model capable of classifying surveillance video clips into "usual" and "unusual" categories with high accuracy while handling class imbalance and environmental variations. The proposed model was trained and evaluated on three large-scale benchmark datasets: UCF-Crime, XD-Violence, and CCTV-Fights, which represent real-world anomalies under diverse conditions. Experimental results demonstrated that the framework achieved an overall accuracy of 97.41% on UCF-Crime, 98.11% on XD-Violence, and 98.50% on CCTV-Fights, alongside consistently high values of precision, recall, and F1-score. These findings indicate that combining spatiotemporal modelling with attention-driven context aggregation substantially improves anomaly detection performance compared to existing baselines. The significance of this research lies in showing that integrating temporal modelling and attention can advance current surveillance systems, providing a more scalable and effective approach for anomaly detection in surveillance videos.

1. Introduction

Modern video surveillance technology has become an essential element for maintaining public safety across multiple spaces, including cities, transportation terminals, schools, and business enterprises [1]. These systems create video data that requires efficient analysis solutions and accurate monitoring techniques for detecting security threats and public disturbances [2].

Identifying abnormal conduct within video monitoring falls under anomalies and suspicious conduct [3,4]. Identifying abnormal patterns must happen quickly and precisely because it enables immediate actions to reduce their impact [5]. The manual review of surveillance footage is labour-intensive, prone to errors, and too demanding to manage the enormous data output [6]. Current conditions demand automated anomaly detection systems that function in real-time operations throughout various operational environments [7].

The traditional methods that detect video anomalous events depend on human-designed features and statistical pattern recognition systems [8]. Automated surveillance benefited from these initial methods, yet they tend to fail when dealing with the real-world environment complexity and variation [9,10]. These standard methods face major hurdles when implementing them in complex environments because of changing backgrounds along with different lighting conditions and the specific situational nature of anomalies [11].

Deep learning transforms computer vision through advanced tools for extracting features while performing pattern recognition operations [12]. Both convolutional neural networks (CNNs) and recurrent neural networks (RNNs) demonstrate outstanding success through deep learning models when tackling object detection, together with action recognition and scene understanding tasks [13,14]. The models excel at arranging hierarchical structures from untreated data, which positions them as optimal candidates for video surveillance anomaly identification [15,16]. Multiple hurdles continue to hinder video anomaly detection systems, even with the advantages brought through deep learning technology. Heterogeneous video quality (e.g., varying resolutions, frame rates, and lighting conditions) often reduces model robustness and accuracy [17]. The lack of available labelled anomalous data represents a substantial problem since anomalies naturally occur infrequently between multiple diverse patterns. For practical deployment it is essential to achieve high accuracy with real-time performance operations [18].

The proposed research focuses on creating a reliable method to find anomalies in surveillance videos which exploits deep learning architecture capabilities. The technique works to overcome traditional process weaknesses through combination of spatial-temporal modelling and attention strategies to understand video information patterns across time and space. The model operates as a binary classifier whose main objective is to detect normal versus abnormal incidents while generating useful response outputs intended for surveillance functions.

The contributions of this study are as follows:

- Integration of Spatiotemporal Features: The model employs 3D convolutional layers to extract spatial and temporal features simultaneously, capturing the intricate dynamics of video sequences.
- Temporal Modelling with LSTM: This approach effectively models temporal dependencies by incorporating long-short-term memory networks, enhancing the detection of anomalies that unfold over time.
- Attention Mechanism: Including an attention layer allows the model to focus on critical segments of the video, improving interpretability and performance.
- Comprehensive Evaluation: The method is evaluated on multiple benchmark datasets, including UCF-Crime, XD-Violence, and CCTV-Fights, demonstrating its generalizability across various scenarios.

This paper endeavors to advance the field of video anomaly detection by presenting a deep learning-based approach that addresses current challenges and offers practical solutions for surveillance systems. Figure 1 shows usual and unusual activity inside an ATM in a surveillance camera frame [19].

Figure 1: Usual and unusual activity inside an ATM in a surveillance camera frame. The usual activity is on the right side, while the unusual one is on the left [19].

The rest of this paper is organized as follows: Section 2 reviews related work in anomaly detection in video footage. Section 3 explains the details and steps of the proposed method. Section 4 discusses the obtained results. Finally, Section 5 concludes.

2. RELATED WORKS

Unusual activity detection within security footage has significantly progressed because of recent developments in deep learning and computer vision technology. Several cutting-edge methods are designed to address detection problems in video surveillance information through innovative techniques. The research team of Feroze et al. [20] presented a group anomaly detection system that analyses collective actions instead of single actions for identifying abnormal group dynamics. The approach identifies relationships between several moving entities, so it detects anomalies better in dense environments, yet shows decreased performance when abnormalities show up faintly in isolated person movements. Mahdi et al. [21] developed a whole system for reporting unusual activities through deep learning techniques that used convolutional neural networks for normal/abnormal behaviour separation. The research work provided an effective base for surveillance applications through its high precision performance but its data requirements are significant together with limited response time in actively changing environments. The integration of CNNs with LSTM networks for violence detection in crowded environments achieved by Sharma et al. [22] enabled real-time analysis of temporal violent incident patterns. The dual-architecture system developed by their team provides superior detection performance yet applies mostly to active physical contact between subjects without broad ability to detect diverse anomalies.

The research by Ramzan et al. [23] took novelty steps through video pattern analysis and temporal event detection for anomaly identification processes. This study demonstrated how combining multiple domains of data helps detect more anomalies especially in security and healthcare applications. The additional complexity of data fusion systems together with their implementation difficulties creates obstacles for simple deployment and expansion.

Kalshetty and Parveen [24] proposed an anomaly detection model based on an improved ResNet101 architecture with non-linear analysis for real-time surveillance. Their method achieves high accuracy and provides actionable insights, but it demands high computational resources and may need optimization for resource-constrained environments. Shukla [25] contributed an intrusion detection system that combines teaching-learning-based optimization with support vector machines, thus extending anomaly detection into the cybersecurity domain. While this robust multi-domain framework bridges visual and cybersecurity applications, the integration of optimization algorithms with SVMs adds complexity and tuning challenges.

Jaramillo-Alcázar et al. [26] explored anomaly detection in innovative industrial environments by integrating IoT and machine learning techniques, demonstrating the method's versatility and applicability to industrial settings. Nevertheless, adapting this approach to other domains might require significant modifications. Mane [27] proposed a sustainable, automated method for real-time anomaly detection using CNNs combined with deep learning strategies. The study focused on efficiently managing large video datasets while maintaining high detection accuracy; however, its reliance on deep architectures increases training time and computational demands. Wani et al. [28] developed the Efficient and Accurate Suspicious Activity Detection (EASAD), which integrates an enhanced Squeeze-Net architecture and an improved U-Net segmentation. This fully automated approach significantly improves detection performance, yet it may face limitations when applied to heterogeneous datasets with varying quality and frame rates. Gawande et al. [29] presented a Novel person detection and suspicious activity recognition using enhanced YOLOv5 and motion feature map, specifically addressing complex criminal behaviours such as arson and vandalism in urban settings. Although it introduces a domain-specific focus that effectively discriminates between various crime types, the model's complexity and dependency on domain-specific tuning can limit its broader applicability without further adjustments.

In this paper, we propose a novel approach that builds on the strengths of these earlier methods by integrating spatiotemporal feature extraction via 3D CNNs, temporal dynamics modelling with LSTM networks, and an attention mechanism to focus on critical segments of the video selectively. The proposed method addresses the limitations identified in previous works by improving the detection accuracy in heterogeneous and dynamic environments, optimizing computational efficiency for real-time performance, and ensuring adaptability across diverse application domains. Table 1 shows a summary table of the related works discussed, with a brief presentation of the strengths and weaknesses of each paper.

Table 1: Summary of related works with a brief presentation of the strengths and weaknesses.

Ref.	Approach (Method Used)	Strength Points	Weak Points or Restrictions
[20]	Group anomaly detection focusing on collective behaviours	Captures contextual interactions in crowded scenarios	Less effective with subtle individual anomalies
[21]	CNN-based framework for unusual activity detection	High accuracy; strong foundation for surveillance applications	Requires extensive annotated data; may struggle in real-time dynamic settings
[22]	Integration of CNN and LSTM for violence detection	Enhanced temporal analysis; real-time performance	Limited generalizability beyond overt physical interactions
[23]	Multi-modal integration of video and temporal signals	Captures a broader range of anomalies through data fusion	Increased complexity and fusion challenges
[24]	Improved ResNet101-based anomaly detection with non-linear analysis	High accuracy; actionable real-time insights	High computational resource demands; requires optimization
[25]	Intrusion detection using teaching-learning optimization with SVMs	Bridges the visual and cybersecurity anomaly detection	Added complexity; tuning challenges
[26]	IoT and machine learning-based anomaly detection in industrial settings	Versatile, effective IoT integration	May require significant modifications for broader domain adaptability
[27]	Sustainable automated anomaly detection using CNNs and deep learning	Efficient management of large-scale data; balanced accuracy	Increased training time and computational demands
[28]	Efficient and Accurate Suspicious Activity Detection (EASAD) using enhanced Squeeze-Net	Fully automated; significant detection performance improvements	Limited by heterogeneous data quality and varying frame rates
[29]	Detection and suspicious activity recognition using enhanced YOLOv5	Effectively discriminates between various crime types	High model complexity; requires domain-specific tuning for broader application

3. PROPOSED METHOD

This paper proposes a new deep learning framework for Robust Unusual Activity Detection in Surveillance Videos, introducing innovative elements to improve recognition accuracy and robustness. Figure 2 shows the graphical representation of the proposed method.

The method is structured into three main stages: data collection and preparation, the model architecture, and the model training strategy.

3.1 Dataset Collection and Preparation

For robust evaluation and training of the proposed model, we employ three complementary datasets that capture a wide range of abnormal events within surveillance footage:

- **UCF-Crime Anomaly-Detection-Dataset-UCF:**
  This large-scale dataset [30] consists of 1,900 long surveillance videos spanning more than 128 hours, annotated into abnormal activities such as robbery, assault, burglary, and arson, along with normal activities. Its diversity in environmental conditions (day/night, indoor/outdoor) and heterogeneous video

quality makes it a challenging and comprehensive benchmark for anomaly detection. For this study, the dataset was split into three subsets: 70% for training (1,330 videos), 15% for validation (285 videos), and 15% for testing (285 videos). This ensures a balanced representation of both anomalous and normal activities across all sets. All anomaly classes available in the dataset were included in the experiments to ensure comprehensive evaluation and to capture the wide spectrum of real-world abnormal activities.

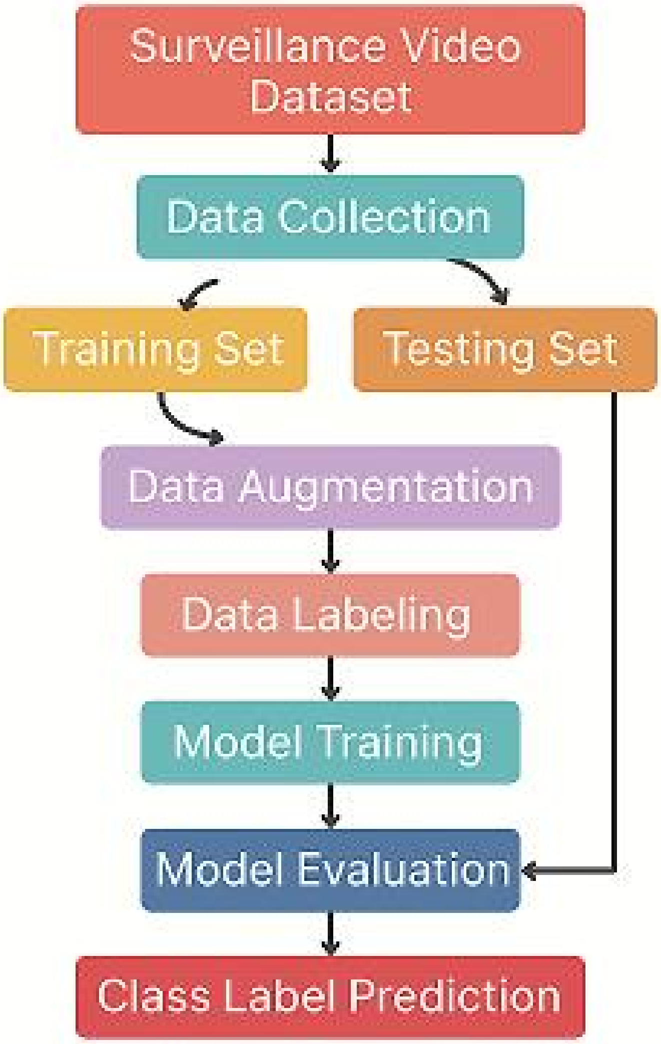

Figure 2: Schematic diagram of the proposed method. The workflow starts with collecting surveillance video data and splitting it into training and testing sets. After applying augmentation and labelling, the data is used to train and evaluate the anomaly detection model, which ultimately predicts class labels for usual and unusual activities.

- **XD-Violence:**
  The XD-Violence dataset [31] contains approximately 4,754 untrimmed videos with a total duration of 217 hours. It includes both violent and non-violent scenarios captured in diverse environments such as streets, stadiums, and transportation hubs. This variability in video quality, camera viewpoints, and scene complexity enhances the model's capability to generalize to real-world conditions. To enable robust evaluation, the dataset was divided into training (3,328 videos, 70%), validation (713 videos, 15%), and testing (713 videos, 15%). This division maintains consistency while preserving class diversity in each subset.

- **CCTV-Fights:**
  This dataset [32] comprises over 1,000 annotated video clips focusing on physical altercations and

aggressive behaviour in CCTV recordings. The videos represent short surveillance clips from different crowd densities, camera angles, and lighting conditions, further diversifying the abnormal event scenarios and complementing the larger datasets. The dataset was divided into 700 clips (70%) for training, 150 clips (15%) for validation, and 150 clips (15%) for testing, with equal proportions of fight and non-fight clips in each split.

Together, these datasets ensure comprehensive coverage of diverse anomaly types, video qualities, and environmental contexts, providing a solid foundation for training and evaluating the proposed framework. Figure 3 show the sample frames from surveillance video datasets.

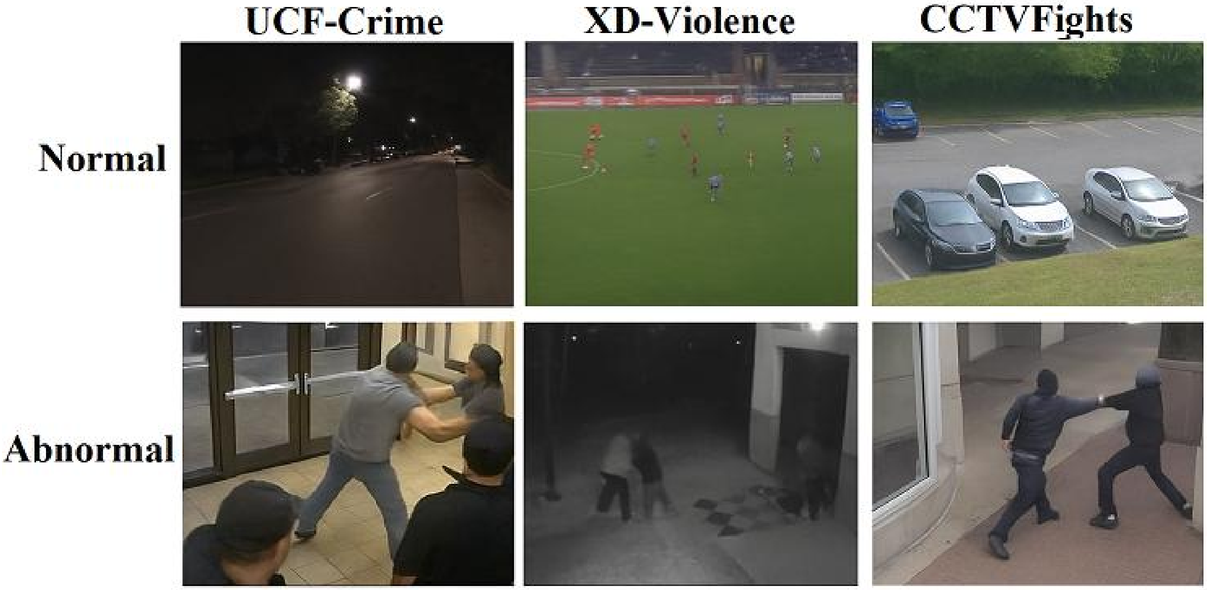

Figure 3. Representative sample frames from surveillance video datasets. Each column corresponds to a different dataset. Normal frames are shown on the top, while anomalous frames are on the bottom.

The preparation procedures in the proposed method include the following:

- **Frame Standardization:** All video frames are resized to a consistent resolution to standardize spatial dimensions (224 × 224).
- **Normalization:** Pixel intensities are normalized to ensure the input data are on comparable scales, thus promoting stable convergence during training.
- **Augmentation:** To mitigate class imbalance, a comprehensive augmentation pipeline, including spatial transformations (random horizontal flips), photometric adjustments (brightness and contrast variations), and temporal modifications (frame shuffling and cropping), is applied. By exposing the model to a wide range of conditions and activity variations during training, this stage helps improve its ability to generalize and recognize abnormal events.
- **Dataset Partitioning:** The collected videos are subsequently partitioned into training, validation, and test sets, adhering to a 70:15:15 ratio, ensuring each set contains a balanced representation of usual and unusual events.
- **Annotation:** Each video sample is rigorously labelled to indicate the presence of either usual or unusual activities, serving as ground truth for supervised learning.

3.2 The Proposed Model Architecture

The proposed method introduces a hybrid deep learning framework that combines 3D CNNs for spatiotemporal feature extraction, LSTMs for temporal dynamics modelling, and an attention mechanism for prioritizing important features, culminating in binary classification. Figure 4 illustrates the Architecture of the proposed model.

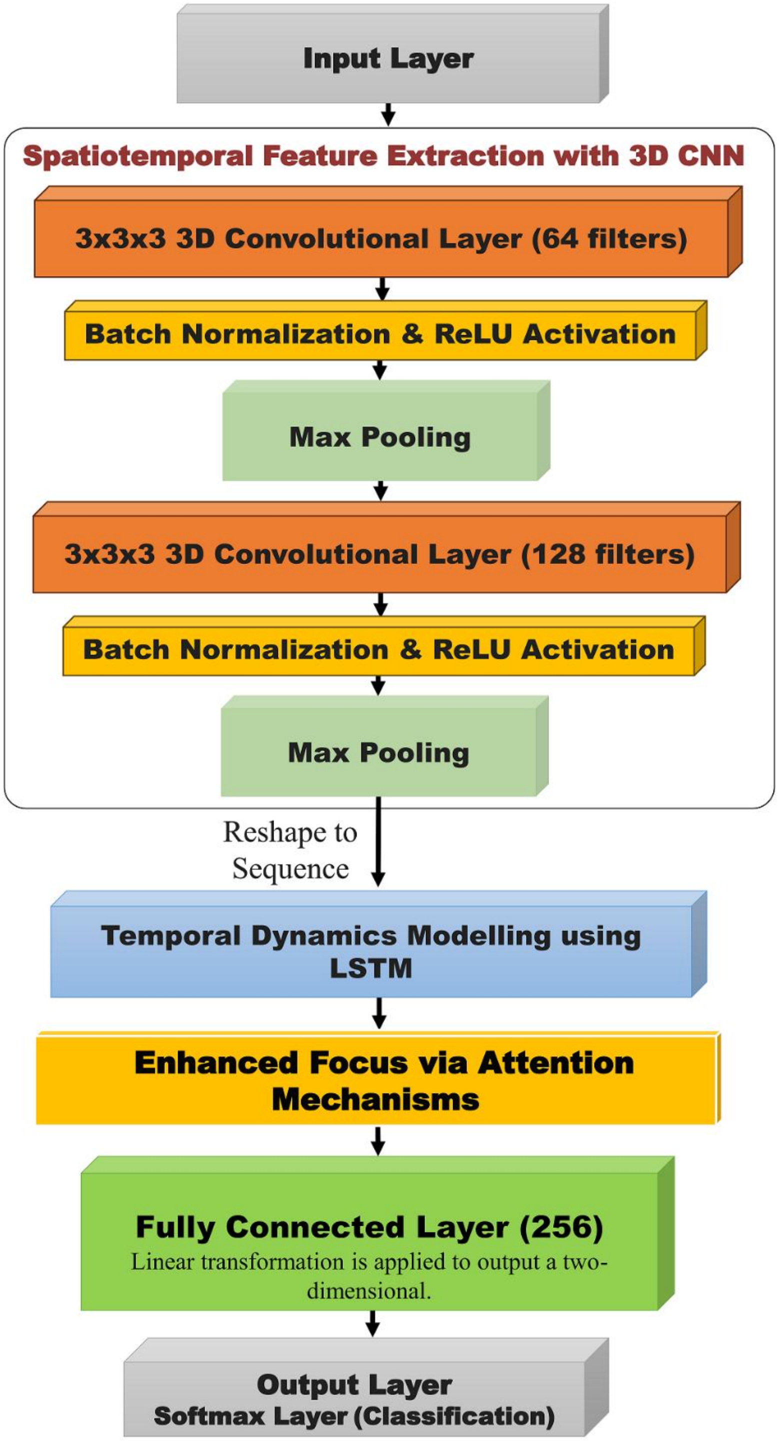
Input Layer
Spatiotemporal Feature Extraction with 3D CNN
3x3x3 3D Convolutional Layer (64 filters)
Batch Normalization & ReLU Activation
Max Pooling
3x3x3 3D Convolutional Layer (128 filters)
Batch Normalization & ReLU Activation
Max Pooling
Reshape to Sequence
Temporal Dynamics Modelling using LSTM
Enhanced Focus via Attention Mechanisms
Fully Connected Layer (256)
Linear transformation is applied to output a two-dimensional.
Output Layer
Softmax Layer (Classification)

Figure 4: The Architecture of the proposed model. The proposed architecture extracts spatiotemporal features using 3D CNN layers, models temporal dependencies with LSTM, and applies an attention mechanism to emphasize key segments before producing the final binary classification.

3.2.1 Spatiotemporal Feature Extraction with 3D CNN:

The proposed method leverages a tailored 3D Convolutional Neural Network (3D CNN) to capture both spatial details and short-term temporal dynamics directly from video inputs. The architecture is composed of two primary convolutional blocks. The first block begins with a 3D convolutional layer containing 64 filters of size

3×3×3, applied with stride 1 and padding 1 to preserve the temporal and spatial dimensions. This is followed by Batch Normalization to stabilize training and a ReLU activation function for non-linearity. A Max Pooling layer with a kernel size of (1,2,2) is then applied, reducing the spatial dimensions (height and width) while retaining the temporal resolution. The second block extends this process with a 3D convolutional layer of 128 filters, also sized 3×3×3 with stride 1 and padding 1, again followed by Batch Normalization and ReLU activation. A second Max Pooling layer with a kernel size of (1,2,2) further reduces the spatial dimensionality. The output feature maps are subsequently rearranged so that the temporal dimension serves as the sequential axis for the LSTM module, enabling the model to learn long-term temporal dependencies across video frames.

3.2.2 Temporal Dynamics Modelling using LSTM:

To effectively model long-range temporal dependencies across video frames, the 3D CNN feature maps are reshaped into a sequence of per-frame feature vectors, where each vector is formed by flattening the spatial dimensions and channel information into a single representation. These vectors are then processed by a two-layer Long Short-Term Memory (LSTM) network configured with a hidden state dimension of 256. At each time step, the recurrent gates (input, forget, and output) play a crucial role in selectively retaining relevant temporal context and filtering out noise, allowing the model to capture the progression of activities throughout the video sequence. The LSTM ultimately generates a sequence of hidden states, which serves as the input for the subsequent attention-based processing stage, ensuring that meaningful temporal patterns are preserved for accurate anomaly detection.

3.2.3 Enhanced Focus via Attention Mechanisms

The attention mechanism is integrated to highlight critical temporal segments that signify abnormal events, ensuring that less informative frames do not weaken the model's focus. Specifically, a learnable linear layer is applied to each LSTM hidden state, and the resulting values are normalized into scalar scores using a SoftMax function across the temporal dimension. These normalized weights are then used to compute a weighted sum of the LSTM outputs, generating a context vector that aggregates the most salient temporal features. The final context vector effectively summarizes the video sequence by concentrating on the frames most relevant to identifying unusual activity.

3.2.4 Classification Layer:

The final stage employs a fully connected layer that maps the 256-dimensional context vector to binary output classes, determining whether the activity is usual or unusual. A linear transformation produces a two-dimensional logits vector corresponding to the two classes, followed by a SoftMax activation during training and inference to generate probabilistic class scores. To address class imbalance, the training process incorporates a weighted cross-entropy loss function, while regularization techniques such as dropout and weight decay are applied to prevent overfitting and ensure robust end-to-end training of the proposed architecture.

Table 2 shows the hyperparameter values used in the experiments. Table 3 provides a summary of the network architecture parameters.

Table 2: Hyperparameter values used in the proposed model.

Hyperparameter	Value	Description
Learning Rate	1e-4	Initial learning rate for the Adam optimizer
Batch Size	16	Number of video samples per batch
Number of Epochs	50	Total training iterations over the entire training set
LSTM Hidden Dimension	256	Dimensionality of the LSTM hidden layer
LSTM Layers	2	Number of LSTM layers stacked for temporal modelling
Dropout Rate	0.5	Dropout probability applied after fully connected layers
Weight Decay	1e-5	L2 regularization parameter
Pooling Kernel Size	(1, 2, 2)	Kernel size for spatial pooling in 3D CNN

Hyperparameter	Value	Description
Attention Mechanism Input	256	Dimensionality corresponding to the LSTM hidden states used in attention

Table 3: Summary of network structural parameters

Layer Type	**Parameters**	**Description**
3D Convolutional Layer 1	64 filters; kernel size = 3×3×3; stride = 1; padding = 1	Extracts initial spatiotemporal features while preserving spatial and temporal dimensions.
Batch Normalization (Layer 1)	N/A	Normalizes feature maps from Convolutional Layer 1 to promote stable and efficient training.
ReLU Activation	N/A	Introduces non-linearity into the network, enabling the learning of complex feature hierarchies.
Max Pooling Layer 1	Kernel size = 1×2×2	Reduces the spatial dimensions while retaining the whole temporal dimension in Block 1.
3D Convolutional Layer 2	128 filters; kernel size = 3×3×3; stride = 1; padding = 1	Captures deeper spatiotemporal features, enhancing the feature representation for subsequent layers.
Batch Normalization (Layer 2)	N/A	Stabilizes and accelerates training by standardizing the outputs from Convolutional Layer 2.
ReLU Activation	N/A	Introduces non-linearity into the network, enabling the learning of complex feature hierarchies.
Max Pooling Layer 2	Kernel size = 1×2×2	Further reduces the spatial resolution, creating compact feature maps while preserving temporal consistency.
LSTM Layer	2 layers; hidden state size = 256	Models long-range temporal dynamics by processing flattened feature vectors sequentially.
Attention Layer	Linear transformation, SoftMax normalization	Computes attention weights over the LSTM outputs to highlight salient temporal features and generate a context vector for classification.
Fully Connected Layer	Input size = 256; output size = 2 (binary classification)	Transforms the context vector into class logits, with a SoftMax activation subsequently applied to provide probabilities for the binary outcomes (Usual vs. Unusual).

3.3 Training Model

The training strategy is meticulously devised to optimize model performance and generalization:

- **Loss Function:**
    - A weighted cross-entropy loss function addresses the class imbalance, ensuring that the minority class (unusual) receives appropriate emphasis during optimization.
- **Optimization Strategy:**
    - The Adam optimizer is employed with an initial learning rate of 1×10^{-4}.
    - A learning rate scheduler dynamically adjusts the rate based on validation performance, promoting steady convergence.
- **Regularization Techniques:**
    - Dropout (with a rate of 0.5) is applied after the fully connected layer to prevent overfitting by randomly deactivating units during training.
    - Batch Normalization is inherent within the convolutional layers to stabilize training by maintaining consistent feature distributions.
- **Training Procedure:**
    - The model is trained over a specified epoch using mini-batch stochastic gradient descent with a batch size 16.

- Online data augmentation is performed during training to enhance model robustness further and simulate various realistic conditions.
- Model evaluation is periodically conducted, with performance metrics such as accuracy, precision, recall, and F1-score guiding iterative improvements and implementing early stopping mechanisms.

Upon completion of training, the trained network efficiently classifies incoming video sequences into 'usual' or 'unusual' categories based on the learned spatiotemporal features and attention-weighted context.

Algorithm 1 summarizes the complete anomaly detection process in the proposed security scenes, outlining the key stages of the proposed method from input to binary classification.

Algorithm 1: High-Level Steps for the Proposed Unusual Activity Detection Method

```
Input: Pre-processed surveillance video
Output: Binary classification label (Usual or Unusual)
    1. Data Collection and Preprocessing
        o Acquire surveillance video data from specified datasets.
        o Standardize frame dimensions via resizing and normalize pixel intensities.
        o Apply data augmentation (spatial flips, brightness/contrast adjustments, temporal cropping/shuffling)
          and partition data into training and testing sets.
    2. Spatiotemporal Feature Extraction
        o Process the input video using a 3D CNN:
            ▪ Block 1: 3D Convolution with 64 filters (3×3×3, stride 1, padding 1), Batch Normalization,
              ReLU activation, and max pooling (kernel size: 1×2×2).
            ▪ Block 2: 3D Convolution with 128 filters (3×3×3, stride 1, padding 1), Batch Normalization,
              ReLU activation, and max pooling (kernel size: 1×2×2).
        o Rearrange output such that the temporal dimension becomes the sequence axis.
    3. Temporal Dynamics Modelling
        o Flatten per-frame feature maps and feed the sequence into an LSTM (2 layers, hidden size = 256) to
          capture long-range temporal dependencies.
    4. Attention Mechanism Integration
        o Apply a linear transformation to compute attention scores over LSTM outputs.
        o Normalize scores via SoftMax and obtain a weighted context vector summarizing salient temporal
          features.
    5. Classification
        o Feed the attention-weighted context vector into a fully connected layer.
        o Use a SoftMax layer to produce final classification probabilities for the binary classes (Usual vs.
          Unusual).
```

4. RESULTS AND ANALYSIS

The experimental evaluation of the proposed method ran on PyTorch through a workstation, which included an Intel Core i7 processor and 32 GB of RAM, and an NVIDIA GeForce RTX 2080 GPU. The implemented training software used Python 3.8 and obtained the newest versions of NumPy and PyTorch, together with the required deep learning libraries. The training process was conducted for 50 epochs using 16 sample batches and an Adam optimizer, which started training from 1×10^{-4} learning rate. Multiple simulations with varying combinations emerged from extensive parameter optimization initiatives to reach optimal training performance while maintaining adequate processing speed. Figure 5 shows the overall training accuracy and loss across the combined datasets.

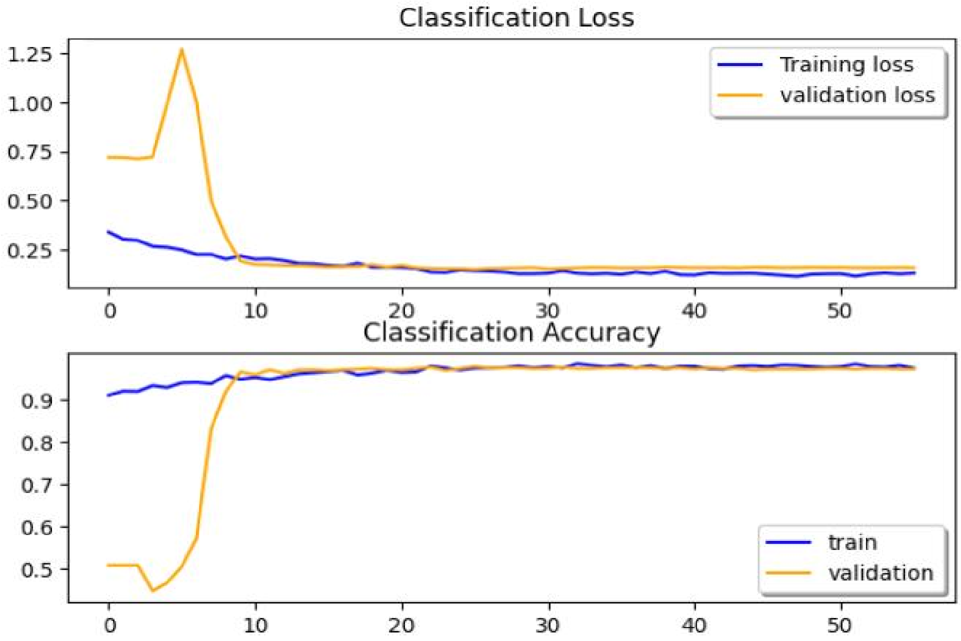

Figure 5: The overall training accuracy and loss across the combined datasets.

The proposed framework employs 3D CNNs for space-time feature extraction, followed by LSTM networks to capture long-term relations and attention modules to focus on important features, thus creating a strong system for detecting anomalous conduct in security tapes. The performance assessment for the proposed methodology took place on UCF-Crime benchmarks and XD-Violence, along with CCTV-Fights. The performance metrics, Accuracy and Precision, together with Recall and F1-Score, appear in Table 4 for all datasets. Figure 6 displays a depiction of the model performance metrics to enable visual assessment of metric changes between datasets.

Table 4: Performance Metrics on Three Datasets

Dataset	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
UCF-Crime	97.41%	97.75%	97.21%	97.47%
XD-Violence	98.11%	98.35%	97.95%	98.15%
CCTV-Fights	98.50%	98.79%	98.25%	98.51%

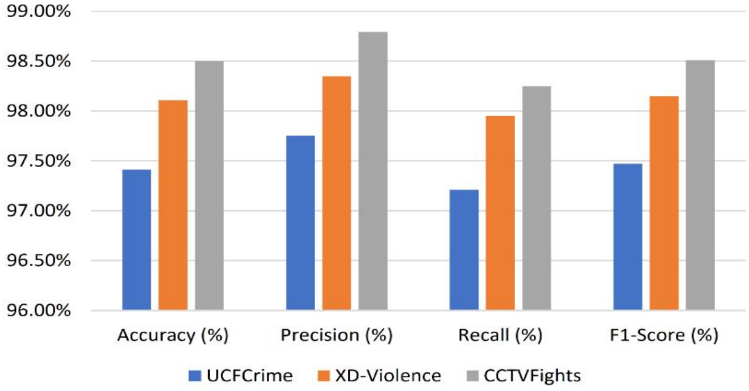

Figure 6: Performance metrics on three datasets.

The proposed method shows consistent performance excellence in various datasets through the combined analysis of spatial and temporal features. The attention mechanism integration was an essential component that directed analysis to the most crucial video segments to remedy any adverse impact of reduced temporality from 3D CNN features. The method's focus maintained an ideal balance between precision and recall by resolving typical class imbalance problems that occur in surveillance data circumstances. Performance evaluation demonstrates that the proposed system provides robust operation and generalizes effectively to new data.

Research involved a thorough evaluation that established the proposed technique's performance level against established methods when using the UCF-Crime dataset. The UCF-Crime dataset was selected because it functions as the benchmark in this field, which enabled us to measure the proposed approach against current cutting-edge surveillance video anomaly detection methods. The proposed method displayed superior results than related approaches because it combined an end-to-end training mechanism with integrated spatiotemporal analysis and context-aware attention models. Table 5 presents a comparison between the proposed method and related works on the UCF-Crime dataset.

Table 5: Comparison with Related Methods

Method	Accuracy	Method	Year
[21]	95.30%	CNN-based anomaly detection	2021
[27]	95.33%	Mobile-Net + BiLSTM	2024
[28]	95.51%	Deep learning–based IoT surveillance	2024
[29]	96.12%	YOLOv5 + Motion Feature Map	2024
Proposed method	**97.41%**	**3D CNN + LSTM + Attention**	**2025**

The results demonstrate that the proposed approach achieves better than related techniques in accuracy measurements and generates remarkable improvements in precision, along with recall and F1 score metrics. The proposed method proves its worth as an enhanced solution for detecting anomalies in security footage obtained from camera surveillance systems. A performance comparison between these methods appears in Figure 7, which visually illustrates the performance differences between them.

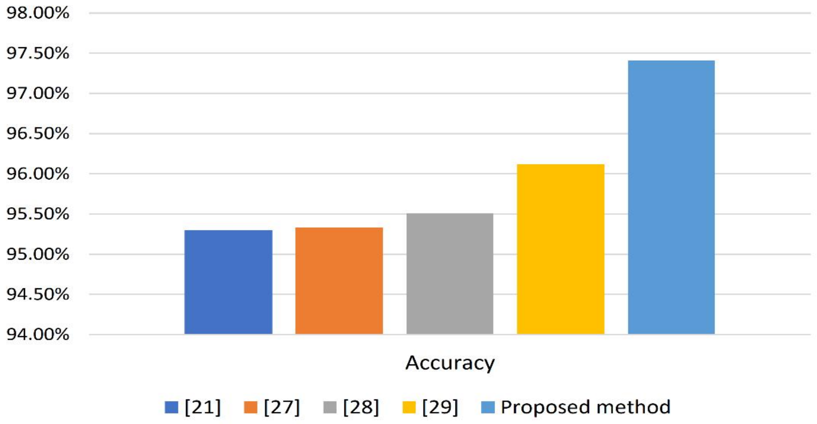

Figure 7 Comparison of the proposed method with some related works.

The experimental results demonstrate that the proposed deep learning model effectively detects abnormal events in many types of surveillance environments. The proposed technique generates superior results by combining high accuracy performance with equal strength between precision and recall measurements compared to existing approaches. Learned spatiotemporal features through 3D CNNs paired with temporal modelling capabilities of LSTMs and attention mechanisms strengthens the system's ability to identify both major and tiny abnormal patterns in complex dynamic spaces. The modifications of this system produce anomaly detection capabilities alongside reliable operations. The model achieves enhanced performance metrics while offering practical operational scalability for security system deployment, which enhances public security measures in various operational environments.

5. Conclusions

This paper introduced a deep learning framework to detect suspicious activity patterns throughout surveillance video recordings. The proposed model uses 3D convolutional layers to obtain spatial features and short-term temporal information, together with an LSTM network that supports modelling extended temporal dependencies. Adding an attention mechanism further refines this process by dynamically weighting the most informative segments of each video, thereby mitigating the dilution of critical anomaly cues, through extensive experiments on the UCF-Crime, XD-Violence, and CCTV-Fights datasets. The proposed method achieved superior accuracy, precision, recall, and F1-score metrics. This represents a substantial improvement over traditional handcrafted-feature approaches and prior CNN- or LSTM-only models, which struggled with limited temporal context or class imbalance issues. The proposed framework exceeds strong quantitative results by offering practical deployment features for real-time implementation. The end-to-end system design allows quick inference processing, which makes it compatible for real-time integration with surveillance systems. The model achieves improved generalizability through combining data augmentation techniques with weighted loss function training, which makes it perform well across various activity patterns and lighting conditions. The ablation studies conducted confirm the individual contributions of the 3D CNN, LSTM, and attention modules to overall performance gains. Future work may explore adaptive temporal resolutions and multi-stream fusion strategies to capture anomalies occurring at various time scales. Extending the framework to handle multi-class anomaly categorization and incorporating unsupervised pre-training could further advance its applicability. Nonetheless, the current study lays a strong foundation for practical, high-performance anomaly detection in security footage, offering a scalable solution for enhancing public safety and operational efficiency in surveillance environments.

Acknowledgement: The authors thank the Middle Technical University, Institute of Medical Technology Al-Mansur, Baghdad, Iraq, for supporting this work.

Conflict of Interest: The authors declare that there are no conflicts of interest associated with this research project. We have no financial or personal relationships that could potentially bias our work or influence the interpretation of the results.

References

[1] A. Khandekar, V. Meneni, H. Uddin, B. Nikhil, M. Tejeshwar, and R. Al–Fatlawy, "Sustainable abnormal events detection and tracking in surveillance system," E3S Web Conf., vol. 529, p. 04009, 2024.
[2] J. Gao, J. Shi, P. Balla, A. Sheshgiri, B. Zhang, H. Yu, et al., "Camera-based crime behaviour detection and classification," *Smart Cities*, vol. 7, pp. 1169–1198, 2024.
[3] K. K. Verma, B. M. Singh, and A. Dixit, "A review of supervised and unsupervised machine learning techniques for suspicious behaviour recognition in intelligent surveillance systems," *Int. J. Inf. Technol.*, vol. 14, pp. 397–410, 2022.
[4] P. G. I. M. Chandrasekara, L. G. Chathuranga, K. A. A. Chathurangi, D. M. K. N. Seneviratna, and R. M. K. T. Rathnayaka, "Intelligent video surveillance mechanisms for abnormal activity recognition in real-time: a systematic literature review," *KDU J. Multidiscip. Stud.*, vol. 5, p. 1, 2023.
[5] D. R. Patrikar and M. R. Parate, "Anomaly detection using edge computing in video surveillance system," *Int. J. Multimed. Inf. Retr.*, vol. 11, pp. 85–110, 2022.
[6] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and prospects," *Comput. Biol. Med.*, vol. 149, p. 106060, 2022.
[7] M. S. Mahdi, A. J. Mohammed, and M. M. Jafer, "Unusual activity detection in surveillance video scene," *J. Al-Qadisiyah Comput. Sci. Math.*, vol. 13, p. 92, 2021.
[8] W. Ullah, A. Ullah, T. Hussain, K. Muhammad, A. A. Heidari, J. Del Ser, et al., "Artificial Intelligence of Things-assisted two-stream neural network for anomaly detection in surveillance Big Video Data," *Future Gen. Comput. Syst.*, vol. 129, pp. 286–297, 2022.
[9] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad, and S. W. Baik, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimed. Tools Appl.*, vol. 80, pp. 16979–16995, 2021.
[10] D. Bouchabou, S. M. Nguyen, C. Lohr, B. LeDuc, and I. Kanellos, "A survey of human activity recognition in smart homes based on IoT sensors algorithms: Taxonomies, challenges, and opportunities with deep learning," *Sensors*, vol. 21, p. 6037, 2021.
[11] S. Zhang, Y. Li, S. Zhang, F. Shahabi, S. Xia, Y. Deng, and N. Alshurafa, "Deep learning in human activity recognition with wearable sensors: A review on advances," *Sensors*, vol. 22, p. 1476, 2022.
[12] E. Ramanujam, T. Perumal, and S. J. Padmavathi, "Human activity recognition with smartphone and

wearable sensors using deep learning techniques: A review," *IEEE Sens. J.*, vol. 21, pp. 13029–13040, 2021.

[13] A. Chatterjee and B. S. Ahmed, "IoT anomaly detection methods and applications: A survey," *Internet Things*, vol. 19, p. 100568, 2022.

[14] A. Nadeem, A. Jalal, and K. Kim, "Automatic human posture estimation for sport activity recognition with robust body parts detection and entropy Markov model," *Multimed. Tools Appl.*, vol. 80, pp. 21465–21498, 2021.

[15] G. Diraco, G. Rescio, P. Siciliano, and A. Leone, "Review on human action recognition in smart living: Sensing technology, multimodality, real-time processing, interoperability, and resource-constrained processing," *Sensors*, vol. 23, p. 5281, 2023.

[16] A. Gupta, A. Tickoo, N. Jindal, and A. K. Shrivastava, "Unusual activity detection using machine learning," in *Proc. Int. Conf. Recent Trends Comput. (ICRTC)*, 2022, pp. 551–559, 2023.

[17] A. Sunil and M. H. Sheth, "Usual and unusual human activity recognition in video using deep learning and artificial intelligence for security applications," in *2021 Fourth Int. Conf. Electr., Comput. Commun. Technol. (ICECCT)*, 2021, pp. 1–6.

[18] U. Singh, P. Gupta, and M. Shukla, "Activity detection and counting people using mask-RCNN with bidirectional ConvLSTM," *J. Intell. Fuzzy Syst.*, vol. 43, pp. 6505–6520, 2022.

[19] M. Malekar, "Detecting criminal activities of surveillance videos using deep learning," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, pp. 188–193, 2021.

[20] A. Feroze, A. Daud, T. Amjad, and M. K. Hayat, "Group anomaly detection: past notions, present insights, and future prospects," *SN Comput. Sci.*, vol. 2, p. 1, 2021.

[21] M. S. Mahdi, A. J. Mohammed, and W. Abdulghafour, "Detection of unusual activity in surveillance video scenes based on deep learning strategies," *J. Al-Qadisiyah Comput. Sci. Math.*, vol. 13, p. 4, 2021.

[22] S. Sharma, B. Sudharsan, S. Naraharisetti, V. Trehan, and K. Jayavel, "A fully integrated violence detection system using CNN and LSTM," *Int. J. Electr. Comput. Eng.*, vol. 11, pp. 3374–3380, 2021.

[23] M. Ramzan, A. Abid, and S. Awan, "Automatic unusual activities recognition using deep learning in academia," *Comput. Mater. Continua*, vol. 70, pp. 1829–1844, 2022.

[24] R. Kalshetty and A. Parveen, "Abnormal event detection model using an improved ResNet101 in context-aware surveillance system," *Cogn. Comput. Syst.*, vol. 5, pp. 153–167, 2023.

[25] A. Shukla, "Detecting impersonation episode using teaching learning–based optimization and support vector machine techniques," *Expert Syst.*, vol. 40, p. 7, 2023.

[26] Á. Jaramillo-Alcázar, J. Govea, and W. Villegas-Ch, "Anomaly detection in a smart industrial machinery plant using IoT and machine learning," *Sensors*, vol. 23, p. 8286, 2023.

[27] D. Mane, "Real-time anomaly detection in video surveillance: A mathematical modelling and nonlinear analysis perspective with MobileNet and Bi-LSTM," *CANA*, vol. 31, pp. 306–319, 2024.

[28] M. H. Wani and A. R. Faridi, "EASAD: efficient and accurate suspicious activity detection using deep learning model for IoT-based video surveillance," *Int. J. Inf. Technol.*, vol. 16, pp. 4309–4321, 2024.

[29] U. Gawande, K. Hajari, and Y. Golhar, "Novel person detection and suspicious activity recognition using enhanced YOLOv5 and motion feature map," *Artif. Intell. Rev.*, vol. 57, p. 16, 2024.

[30] M. U. Meraj, "Anomaly-Detection-Dataset-UCF," Kaggle Dataset, 2021. [Online]. Available: https://www.kaggle.com/datasets/minhajuddinmeraj/anomalydetectiondatasetucf.

[31] P. Wu, J. Liu, Y. Shi, F. Shao, Z. Wu, and Z. Yang, "XD-Violence," Original Metadata [Dataset], TIB LDM Service, 2024. DOI: 10.57702/mt1rr0km. [Online]. Available: https://roc-ng.github.io/XD-Violence/.

[32] M. Perez, A. C. Kot, and A. Rocha, "Detection of real-world fights in surveillance videos," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brighton, UK, May 12–17, 2019. [Online]. Available: http://rose1.ntu.edu.sg/dataset/cctvFights/.