



Human Suspicious Activity Detection in Video Using a Hybrid 3D CNN-LSTM

Nusrat Abbas¹, Muhammad Saeed², Hina Shafique³, Ghulam Gilanie³, Areeba Khalid³, Haseena Manzoor⁴,
Hafiz Burhan Ul Haq^{5*}

¹Department of Computer Science, National Textile University Faisalabad, Pakistan.

²The Islamia University of Bahawalpur, Bahawalpur, Pakistan.

³Department of Artificial Intelligence, Faculty of Computing, The Islamia University of Bahawalpur, Pakistan.

⁴Department of Software Engineering, Faculty of Information Technology, University of Lahore, Pakistan.

⁵Faculty of Engineering, Department of Electronics and Telecommunication Engineering, King Mongkut's
University of Technology Thonburi (KMUTT), Bangkok, 10140, Thailand.

Corresponding Author: burhanhashmi64@lgu.edu.pk

Received: June 25,2026, **Accepted:** June 29,2026; **Published:** June 30,2026

ABSTRACT

Video surveillance is critical to public safety, watching several camera feeds requires a lot of manual effort and is susceptible to human error. The proposed work of this study is a hybrid model comprising Three-Dimensional Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) network that works as a combined system to learn both spatial and temporal features to automatically detect the violence in CCTV videos. The Real-Life Violence Situations dataset with 2,000 video clips was used in the experiments, which are publicly available. Five frames were uniformly sampled from each video were resized to 64×64 pixels and used for training the proposed model. The LSTM is to capture the long-term motion features, and the 3D-CNN to extract the spatial features and short-term motion features from the video, then to improve the recognition of violence. The proposed model is verified in terms of accuracy, precision, recall and F1-score and compared to a basic 2D CNN and MobileNetV3 model, conducted under the same experimental conditions. The proposed method outperformed MobileNetV3 86.0% and basic 2D CNN 87.0% in terms of accuracy 91.0%, precision 89.1%, recall 93.5% and F1-score 91.0%. The results show that the proposed approach of spatial temporal feature learning greatly outperforms frame-based approaches in terms of performance of detection of violence. A hybrid 3D CNN and LSTM model can detect violence in video with high accuracy. The design captures both spatial and temporal cues. Real-time deployment and larger datasets are needed for practical use.

Keywords: Suspicious activity detection, Violence detection, Video surveillance, 3D Convolutional Neural Network, Long Short Term Memory, Action Recognition.

1. INTRODUCTION

The use of video surveillance has been a key element of most security systems today as it helps to monitor any public and private area to ensure better public safety and crime prevention [1]. This is the reason that surveillance cameras have been deployed at unprecedented rates in transportation hubs, educational institutions, healthcare facilities, commercial buildings and smart cities, resulting in an unprecedented, massive amount of video data that cannot be effectively handled by manual monitoring [2]. Human operators may get bored, distracted, or make different decisions during long duration of monitoring, which makes it hard to maintain the monitoring of many surveillance streams over time [3]. For this reason, the automatic analysis of video streams is an important field of study in computer vision and artificial intelligence which requires intelligent surveillance systems [4].

Human suspicious activity detection has become a topic of growing interest as detecting abnormal activity in time can help minimize security threats and allow for quick emergency response [5]. In recent years, surveillance system is expected to detect the activities like fighting, robbery, vandalism and trespassing etc., without the need of the man eye supervision [6]. These activities can be automatically recognized to enhance the situational awareness and to allow security personnel to act in time before incidents get out of hand [7]. In the absence of constraints, though, it is still challenging to recognize suspicious behavior because of the fact that individual actions are different in different scenes, viewpoints, and recording conditions [8]. In real surveillance situations, performance of recognition drops due to the complex background, illumination variation, camera motion, occlusion and crowd density [9]. These difficulties involve developing intelligent models that are able to learn robust space and time representations from continuous video sequences, not just image frames [10].

First-generation SBD technologies were based on handcrafted feature descriptors and traditional machine learning classifiers. These approaches performed well in controlled scenarios, but were found to be not robust enough in real-world surveillance settings, as handcrafted features were not well suited for capturing complex spatial and temporal patterns [11]. Recent developments in deep learning, which enable autonomous learning of multi-level feature representations in raw video sequences, have greatly enhanced video understanding [12]. Three-Dimensional CNNs have been described to have great performance in capturing both spatial and temporal information from consecutive frames, resulting in better activity recognition results [13]. But the 3D CNNs mostly learn short-term motion patterns and do not effectively learn long-range temporal relationship in complex suspicious activities [14].

Recent works have combined LSTM networks with convolutional models to capture long-term temporal relationships in video streams [15]. Combining the excellent ability of 3D CNN to model spatiotemporal features with the sequential modelling capability of LSTM networks, hybrid 3D CNN-LSTM networks have been suggested to achieve better performance for activity recognition [16]. These advances have been made, current techniques continue to suffer from performance problems under real-world surveillance conditions due to background clutter, viewpoint variations, occlusions, illumination changes and behavioral diversity. Most of existing studies are directed towards general human activity recognition or violent activity detection or anomaly detection instead of detecting suspicious human activity in unconstrained surveillance videos [17]. An efficient hybrid framework which can learn discriminative spatiotemporal features in addition to long-term temporal dependencies is desired to realize reliable suspicious activity detection in complex real-world surveillance environment [18].

The ability of intelligent surveillance systems to analyze the complex human activities of videos from video streams has also been improved with recent advances in deep learning. Survey studies have noted that deep learning techniques are nowadays the most widely adopted solution for automatic detection of suspicious activities and violence in surveillance videos due to their capability to learn powerful spatial and temporal representations straight from surveillance videos [19]. Transformer models for video recognition have shown their superior ability in handling long-range temporal and complex human interactions for better recognition accuracy than traditional convolutional models [20]. In the recent past, transfer learning has also been demonstrated to improve the performance of anomaly recognition tasks, especially when there are fewer labeled surveillance data, making deep learning models more robust to real-world applications [21].

There are still some challenges regarding video violence detection. The majority of CNN models are frame-based and cannot represent long-range temporal features, some hybrid and attention models need extra motion representations or complicated architectures that are hard for use in real time monitoring systems [22]. Current models still have difficulties adapting to different real-world surveillance settings and generalizing to other settings [23]. Therefore, an efficient framework with good detection accuracy that can learn both spatial and temporal information and be light is needed. In order to solve these issues, a hybrid 3D CNN–LSTM model is proposed, which utilizes 3D CNN standard layers for extracting spatiotemporal characteristics and LSTM layers to model the long-term temporal dependencies. Furthermore, for computation simplicity, a simple frame-sampling strategy is applied to keep the key motion information and the proposed framework is tested against representative baseline models with a well-defined experimental protocol.

This study has made the following contributions:

1. In surveillance video analysis, automatic suspicious activity detection is proposed by using a lightweight hybrid 3D CNN–LSTM framework, combining the ability of 3D CNN to learn the spatial and short-term motion features with the long-term temporal dependencies of LSTM to achieve better performance in the violence classification.
2. To achieve low computational complexity and preserve the essential spatial-temporal information, the suggested pipeline is an effective video pre-processing framework that uses uniform temporal frame sampling, frame resizing, and frame normalization. This enables the pipeline to be utilized in real-time surveillance applications.
3. The proposed model is also thoroughly tested on Real-Life Violence Situations dataset and compared with representative baseline models like 2D CNN and MobileNetV3 to judge the performance based on various assessments metrics like accuracy, precision, recall, F1-score, confusion matrix and ROC analysis.

This is the structure of the rest of the paper. The previous work is reviewed and the research gap is identified in Section 2. The materials and techniques, including the proposed model, are described in Section 3. The experimental findings, discussion are shown in Section 4. In Section 5 concludes the study.

2. RELATED WORK

The traditional machine learning methods were first used as suspicious activity detection in surveillance videos. This Study [24], proposed to classify suspicious human activities using Support Vector Machine (SVM), Decision Tree, Random Forest and ensemble learning algorithm with handcrafted features. These methods were able to obtain acceptable classification accuracy, they were heavily dependent on feature development, and lacked robustness to deal with bad surveillance conditions like occlusion, illumination changes, and viewpoint changes. To

go over the boundaries of handmade features, deep learning-based methods are introduced for automated learning of features. In [25], the authors presented a temporal attention mechanism in a deep learning framework to highlight informative video frames for better abnormal activity recognition. The model improved the detection rate, it required a significant amount of computation and, therefore, was not appropriate for limited resources and real-time surveillance systems.

In this study[26], an Enhanced Convolutional Neural Network (ECNN) is proposed for detection of suspicious activities like shooting and stealing in the surveillance videos. The proposed model was able to extract the spatial features and gain an accuracy of 98.38% on the DCSASS dataset. A Quantum CNN was introduced to improve finding anomalies in monitoring videos [27]. The framework combined the features of quantum computers and CNN-based features extraction, and proved to be 99% accurate on the UCSD Ped1, UCSD Ped2, Avenue and Shanghai Tech datasets. A Slow-Fast CNN was developed to identify anti-social situations in surveillance footage in real time was presented in study [28]. The framework used consisted of two pathways: slow semantic information and fast motion features. Assessed through the datasets of the CMU Graphics Lab Motion, UT-Interaction, Peliculas, Hockey Fighting, Web Dataset and UCSD-AD, with 99.6% accuracy.

A hybrid model based on background removal, convolutional autoencoders, and object recognition has been presented for real-time anomaly detection [29]. The dependence on foreground extraction may reduce detection robustness in complex surveillance environments with dynamic backgrounds. To address this limitation[30], proposed a lightweight end-to-end 3D CNN that directly learns both spatial and temporal representations from video sequences without requiring separate motion extraction techniques or recurrent networks. The proposed model was evaluated on

a consolidated dataset comprising the Hockey Fights, Movies, RWF-2000, and Real-Life Violence Situations (RLVS) datasets. Experimental results reported a results of 80.58%, precision of 77.16%, recall of 86.92%, and F1-score of 81.71%. The lightweight architecture reduced computational complexity and was suitable for real-time surveillance applications, although its overall performance indicates that more advanced hybrid architectures may further improve violence detection accuracy.

A hybrid 3D CNN–LSTM architecture was proposed to combine the spatial feature extraction with temporal sequence learning for human activity recognition [31]. Sequential modeling was improved by the incorporation of LSTM. the system was assessed using generic human activity datasets rather than true suspicious activity monitoring cases. This paper [32], proposed a Branch-Fusion Network with Channel Spatial Attention Module (CSAM). The model has been tested on benchmark datasets and demonstrated high AUC value by improving feature representation but is limited to capture long-term temporal dependency in complex Activity Recognition.

To enable real-time surveillance performance [33], introduced a lightweight structure for identifying anomalies that extracts features using a standard autoencoder with compression and classifies anomalies in the spatiotemporal space using a Graph CNN-LSTM. This model was assessed and improved. on the UCF-Crime dataset, then optimized by using structured pruning, quantization-aware training and a self-supervised learning module to reduce the computational cost. The proposed framework achieved 98% multiclass classification accuracy, 0.99 multiclass ROC-AUC, 0.91 binary anomaly ROC-AUC, and showed the estimation time of ~1.22 ms per frame, suitable for real-time surveillance applications. The framework was tested on the UCF-Crime dataset, and needs to be further tested for its applicability to a wider variety of practical surveillance scenarios. A

CNN–LSTM architecture was proposed for detecting anomalies in the UCF-Crime dataset achieving an accuracy of 98.6%. For surveillance videos, the authors of [34] suggested an autoencoder framework using LSTM to detect anomalies. The processing method includes three steps: First, Noise mitigation and background suppression are performed to preprocess the data; Second, an LSTM autoencoder is used to learn the patterns of the data third latent-space regularity analysis is performed supplemented by local minima assistance, for anomaly detection. The model was tested on two datasets, including the CUHK Avenue and UMN datasets, with AUC scores of 93.8% and 94.1%, respectively. The proposed framework significantly enhanced the anomaly localization capabilities and surpassed several baseline methods. It relies on background subtraction, though, which may bring in some lack of robustness in highly dynamic or crowded surveillance scenes. The combination of CNN and LSTM showed improvements to recognition performance but the separation of spatial and temporal feature extraction prevented learning of complex motion patterns. A CNN-LSTM network with an attention mechanism was suggested for the identification of terrorism in surveillance videos in this work[35]. The model

was tested on the RWF-2000 and UCF-Crime with accuracy of 96.21% and AUC score of 97.48% respectively. The attention mechanism focused on informative regions but the model did not fully leverage the joint spatiotemporal feature learning and still depended on 2D CNN features.

Previous studies have significantly enhanced the detection of suspicious activities using CNN, 3D CNN, attention mechanism, and CNN–LSTM frameworks, there are still several limitations, is described in Table 1. The traditional machine learning approaches require handcrafted features and 2D CNN-based approaches don't work well in capturing temporal dynamics. Current 3D CNN models are able to record brief spatiotemporal characteristics, but cannot model long-term temporal dependency, and are also computationally expensive. Many hybrid CNN–LSTM approaches use 2D spatial representations or general human activity datasets, instead of realistic surveillance scenarios. Therefore, to do a better job in detecting suspicious activity from complex real world surveillance videos, an efficient design of integrating 3D CNN for spatial-temporal motion removal of features and LSTM for temporal motion modeling is needed.

Table 1. Recent state of the art Human Suspicious Activity Detection in Video

Ref.	Methodology	Dataset(s)	Key Results	Limitations
[26]	ECNN	DCSASS dataset	98.38% accuracy	Learned spatial features only, ignored temporal dependencies.
[27]	QCNN	UNI-Crime and UCF-Crime	98% accuracy	Limited long-term temporal modeling despite reduced computational complexity.
[28]	Slow-Fast CNN	CMU UTI, PEL, HOF, WED, UCSD-AD (combined benchmark surveillance datasets)	98.6% overall accuracy	Weak long-term activity representation.

[29]	Background subtraction + convolutional autoencoder + object detection framework	Surveillance video datasets	AUC of 94.94%	Sensitive to illumination changes and dynamic backgrounds.
[30]	Lightweight 3D CNN.	Hockey Fights, Movies, RWF-2000, RLVS	80.58% accuracy, 77.16% precision, 86.92% recall, and 81.71% F1-score.	Limited long-term temporal modeling lower accuracy than hybrid models.
[31]	Hybrid 3D CNN-LSTM	Human Activity Recognition datasets	Improved spatiotemporal feature learning by combining 3D CNN feature extraction with LSTM sequential modeling.	Evaluated only on general activity recognition; not suspicious surveillance videos.
[32]	Branch-Fusion Network with CSAM	Surveillance anomaly benchmark datasets	Achieved high anomaly detection.	Limited long-term temporal reasoning.
[33]	Self-supervised Hybrid Autoencoder + Graph CNN-LSTM	UCF-Crime	98.0% multiclass accuracy, 0.99 ROC-AUC	Higher computational complexity; evaluated mainly on UCF-Crime.
[34]	LSTM Autoencoder + Local Minima Assistance	CUHK Avenue, UMN	AUC: 93.8% (Avenue), 94.1% (UMN)	Sensitive to dynamic backgrounds due to background subtraction.
[35]	KianNet(CNN-LSTM)	UCF-Crime and RWF-2000	96.21% Accuracy on RWF-2000; 97.48% AUC on UCF-Crime	Relied on 2D CNN features; lacked joint spatial-temporal feature extraction.

3. METHODOLOGY

The materials and process utilized in the creation and assessment of the suggested hybrid 3D

CNN-LSTM model for the identification of unusual behavior are covered in this section. Dataset preparation, video pre-processing, model creation, learning as well as performance

evaluation are all part of the workflow. Using typical categorization metrics, such as accuracy, precision, recall, f1 score, confusion matrix, and ROC analysis, the model was assessed after being trained on a publicly available benchmark dataset.

3.1. Dataset

This study made use of the Real-Life Violence

Situations data collection. The dataset contains 2,000 brief videos. Of the videos, half are violent and the other half are not. The violent clips include mob clashes and street battles. The non-violent video feature sights from everyday life. Training and testing sets were created from the data collection. The divide was made using an 80:20 split. There were 400 exam films and 1600 training videos. Table 2 displays the data set's specifics. Figure 1 display the distribution of classes.

Table 2. Details of the Real-Life Violence Situations dataset

Property	Value
Dataset	Real-Life Violence Situations (Kaggle)
Total videos	2,000
Violence videos	1,000
Non-violence videos	1,000
Train:test ratio	80:20
Frames per video	5 (evenly sampled)
Frame size	64×64

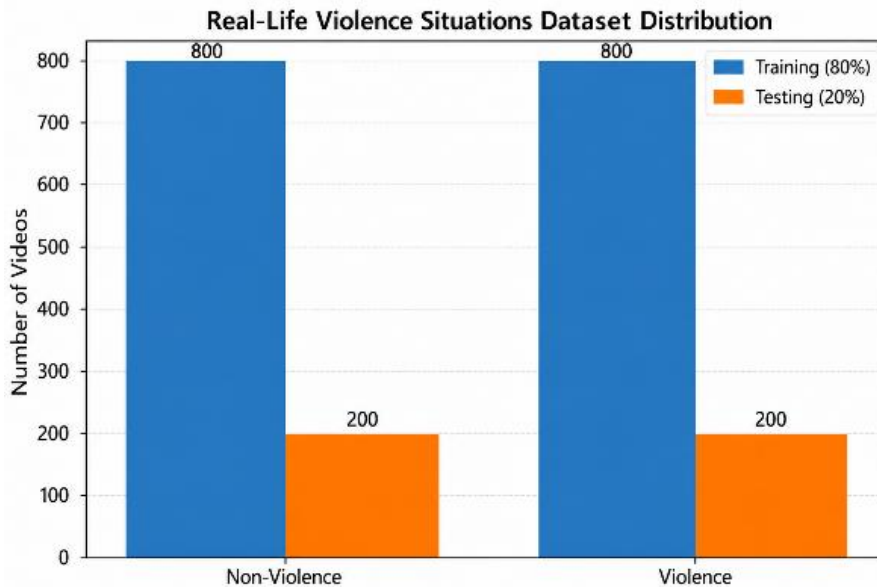


Figure 1. Video distribution per class

3.2. Preprocessing

The raw surveillance videos are pre-processed to generate the standardized video inputs for the proposed hybrid 3D CNN-LSTM model. The video was annotated with a binary class label, 0 for non-violent activity and 1 for violent activity. To address the varying lengths of the surveillance videos, we uniformly sample a fixed sequence of five frames from each video. This sampling technique minimizes computing complexity and maintains a constant input size while preserving representative information from the start, middle, and finish of each video.

First, the structure of the dataset supports a short sequence. Clips in the Real-Life Violence Situations dataset are short and pre-trimmed, and the violent event, where present, spans the clip rather than occupying a brief window inside a longer recording. The task is therefore clip-level classification rather than

temporal localization, and under uniform sampling the five selected frames span the beginning, middle and end of the clip and remain representative of its full duration. This differs fundamentally from untrimmed benchmarks such as UCF-Crime, where sparse sampling can miss the event entirely.

Second, the input cost is decisive for the stated aim of a deployable model. The tensor of $5 \times 64 \times 64 \times 3$ contains 61,440 values, approximately one tenth of a conventional $16 \times 112 \times 112 \times 3$ input, and because the cost of the early convolutional layers scales with this volume, the reduction propagates through the whole network.

Third, the sampled sequence is not required to carry all the motion information by itself. Each 3D convolution kernel spans several adjacent temporal positions and therefore captures

short-term motion internally, while the stacked LSTM layers model the transitions between the sampled instants across the clip. The sequence length must consequently be long enough to represent the temporal evolution of the action, not long enough to resolve every frame of motion.

Sampling interval was written in Equation (1) as:

$$S = \frac{N}{T} \quad (1)$$

where N is the total number of frames in a video and T is the number of frames to be sampled. The indices of the sampled frames were calculated by equation (2):

$$F_i = [i \times S], \quad i = 0, 1, \dots, T - 1 \quad (2)$$

3.3. Proposed Model

This study proposes a hybrid model-based deep learning framework that combines the ability of LSTM connections for learning from the sequence of features and 3D Convolutional Neural Networks (3D-CNNs) to record spatiotemporal features in order to detect suspicious activity. Three components make up the entire structure: (i) spatial-short-term motion extraction blocks; (ii) long-term temporal modeling blocks (many LSTM layers); and (iii) binary classification blocks (totally linked layers). First, four successive 3D convolutional layers with 32, 64, 128, and 256 filters are applied to the input video tensor. ReLU activation layers and 3D max-pooling layers come after both convolution layers to maintain the discriminative spatial-temporal information without significantly reducing the feature space's size. The network can learn the temporal patterns straight from the video sequences thanks to the 3D convolution, which uses convolution in both dimension of space (height and breadth) and temporal dimensions (frames). Each convolutional layer's output feature maps are computed as

After extraction, each sampled frame was scaled to a fixed spatial resolution of 64×64 pixels in order to train the model. In order to enhance numerical stability, speed up model convergence, and prevent significant gradient fluctuations during optimization, the pixel intensity values were subsequently standardized to the range $[0, 1]$. Equation (3) was used to define the normalizing procedure.

$$I_{\{norm\}} = \frac{I}{255} \quad (3)$$

Finally, the normalized frames were sequentially arranged into a fixed-size video tensor. This representation was defined in Eq. (4) as:

$$X = \{I_1, I_2, \dots, I_T\}; X \in R^{T \times 64 \times 64 \times 3} \quad (4)$$

$$Y = \sigma(X * W + b) \quad (5)$$

The created map features by the 3D convolution are activated by the Rectified Linear Unit (ReLU), which speeds up network convergence by maintaining positive responses while setting negative values to zero. The ReLU activation function is given by equation 6.

$$\sigma(x) = \max(0, x) \quad (6)$$

The resulting features maps have been down sampled using 3D max-pooling.

$$P(i, j, k) = \max_{\{(m, n, l) \in \Omega\}} Y(m, n, l) \quad (7)$$

The 3D CNN extracts high-level feature representations, which are then reshaped into sequential feature vectors and fed into three stacked LSTM layers with 128, 64 and 32 memory units, respectively. The LSTM layers capture long-term temporal correlations between neighboring frames of videos so that the network can learn to classify them as either passive or violent, by deep learning the evolution of motion over time. The hidden state of LSTM is updated

using

$$h_t = LSTM(x_t, h_{t-1}) \quad (8)$$

The final hidden representation is processed by two fully connected layers with 64 and 32 neurons, respectively. A dropout layer with a dropout rate of 0.5 is employed in between dense layers to prevent overfitting characterized by the random deactivation of neuron during aggressive or nonviolent training. The output of the dense layer is shown as in 9.

$$z = W_{dh} + b_d \quad (9)$$

The sigmoid activation function, which is described in (10), is used to determine the ultimate probability of the positive class:

$$\hat{y} = \frac{1}{1 + e^{-z}} \quad (10)$$

Equation 11 illustrates the binary cross-entropy

loss function that was used to enhance the proposed network.

$$L = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (11)$$

The overall process of the suggested hybrid 3D CNN–LSTM is shown in Figure 2. The architecture starts with the video preprocessing: uniform frame sampling, resize, and normalize; followed by the spatial temporal feature extraction using 3D CNN. The retrieved feature series is then fed into stacked LSTM layers to learn long-term temporal dependency and the learned representations are finally fed to completely linked layers with a classifier built using sigmoid activation function to distinguish whether the activities are violent or non-violent.

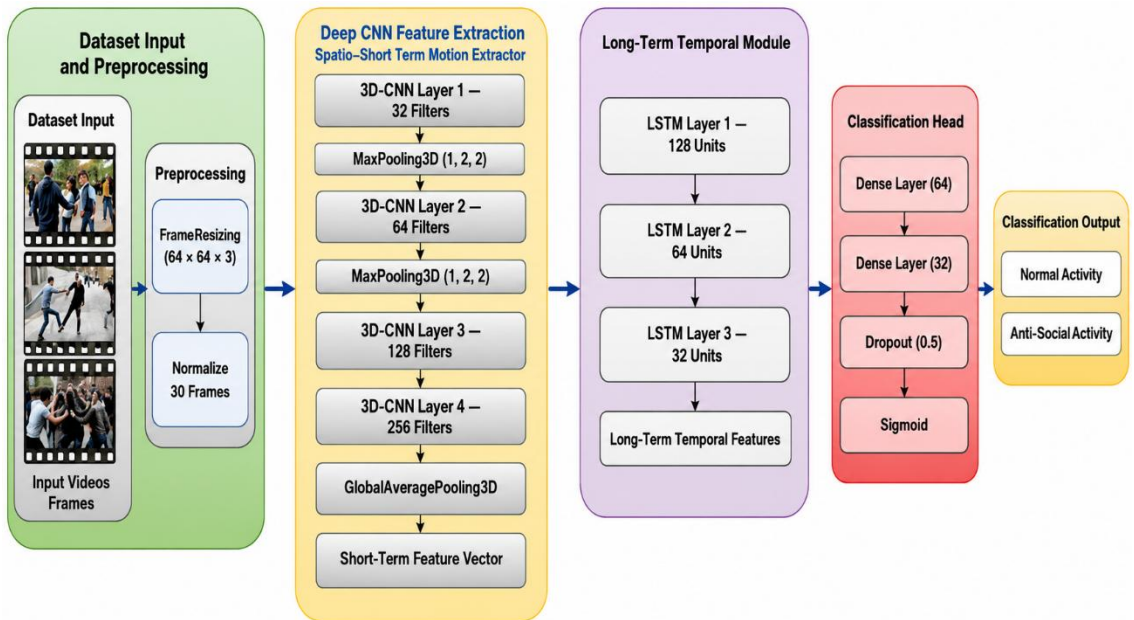


Figure 2. The suggested hybrid 3D CNN and LSTM model's architecture

Two fully linked (dense) layers and a binary categorization using a sigmoid output layer come next. The likelihood that the input video falls into the violent class is provided by the sigmoid activation function. To lessen overfitting and enhance generalization, a dropout layer with a

rate of 0.5 is used prior to the output layer. The primary model setup and training parameter are briefly outlined in Table 3.

Table 3. The Training Parameters details of proposed hybrid Model

Component/Parameter	Setting
Input Video Sequence	5 consecutive RGB frames
Input Resolution	$64 \times 64 \times 3$ pixels per frame
Preprocessing	Frame resizing ($64 \times 64 \times 3$), frame normalization
3D-CNN Layer 1	32 filters
MaxPooling3D	Pool size = (1, 2, 2)
3D-CNN Layer 2	64 filters
MaxPooling3D	Pool size = (1, 2, 2)
3D-CNN Layer 3	128 filters
3D-CNN Layer 4	256 filters
GlobalAveragePooling3D	Converts feature maps into a feature vector
Short-Term Feature Representation	Feature vector extracted from GlobalAveragePooling3D
LSTM Layer 1	128 units
LSTM Layer 2	64 units
LSTM Layer 3	32 units
Long-Term Temporal Representation	Output feature sequence from stacked LSTM layers
Dense Layer 1	64 neurons
Dense Layer 2	32 neurons
Dropout	0.5
Output Activation	Sigmoid
Classification Output	Normal Activity / Anti-Social Activity
Loss Function	Binary Cross-Entropy (BCE)
Optimizer	Stochastic Gradient Descent (SGD), Learning Rate = 0.01
Epochs	10
Batch Size	10
Training/Test Split	80% / 20%

Table 3 shows the full training setup and the design of the proposed hybrid 3D-CNN–LSTM system for anti-social activity detection. The input of the model is a sequence of 5 continual

video frames of RGB data, with each frame reduced to $64 \times 64 \times 3$ pixels and normalized in the preprocessing of the data to obtain more consistent values and improve the stability of the

training.

The four-layer 3D convolutional neural network (3D-CNN) with filter numbers of 32, 64, 128 and 256 are used to extract features of the spatial and short-term motion. After the first and second convolutional layers a MaxPooling3D layer with a pooling size of (1, 2, 2) is used to gradually reduce the spatial dimensions, while retaining the temporal information. After the last convolutional layer, a GlobalAveragePooling3D layer combines the feature maps into a short-term feature vector which is much smaller and contains the most discriminative information, but with much fewer trainable parameters. Generated feature vectors are then fed into a long-time temporal module consisting of three stacked LSTM layers containing 128, 64 and 32 hidden units, respectively. A long-term temporal description of the observed activity is obtained by these recurrent layers, which learn temporal dependencies and sequences in the input frame sequence. The extracted temporal features are traversed two layers that were entirely connected. With 64 and 32 neurons. To avoid overfitting, a dropout rate of 0.5 is used by randomly disabling neurons during training. The binary categorization, the output layer uses the sigmoid activation function, which includes normal and anti-social behaviors.

The model is optimized using the Stochastic Gradient Descent (SGD) algorithm with a learning rate of 0.01 and a loss function of Binary Cross-Entropy (BCE) during training. The network is trained for 10 epochs with a batch size of 10 and the data is split up into 80% for training and 20% for testing. This setup allows the model to learn the spatial and geographical representations efficiently without being too complex and prone to overfitting.

3.4. Evaluation Measures

Five common performance indicators were used to comprehensively evaluate the classification model, including Accuracy, Precision, Recall,

F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Each of the metrics provides a different view of the predictive capability of the model, leading to a more trustworthy evaluation of the overall performance.

Accuracy is the proportion of correctly categorized cases to all occurrences of samples in the data set. It is given by and reflects the overall performance of the classifier:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

Where TP is true positives, TN is true negatives, FP is false positives and FN is false negatives.

Precision measures the percentage of accurately predicted positive cases. out of all instances classified as positive by the model. High precision implies that the classifier makes a small number of false-positive predictions. It is calculated as:

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

Recall, also known as Sensitivity or True Positive Rate (TPR), measures the capability of the model to correctly identify the actual positive instances. It is expressed as:

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

The F1-score is equal measurement which combines precision and recall into a single metric as the harmonic mean. This is especially useful for imbalanced data sets as it takes into account both genuine positives and incorrect results. The F1-score is calculated as:

$$F1\ Score = \frac{2TP}{2TP + FP + FN} \quad (15)$$

In addition to these metrics, the classifier's discriminative ability was evaluated by the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The ROC curve shows the trade-off between the True Positive Rate (Recall) and the False Positive Rate (FPR) for different classification thresholds. The AUC value ranges from 0 to 1, with 1.0 indicating perfect classification performance, 0.5 indicating random guessing and values closer to 1 indicating better discrimination of positive and negative classes. The False Positive Rate is defined as:

$$FPR = \frac{FP}{FP + TN} \quad (16)$$

The AUC can be represented as the area under the ROC curve:

$$AUC = \sum_{i=1}^{n-1} \frac{(TPR_{i+1} + TPR_i)}{2} (FPR_{i+1} - FPR_i) \quad (17)$$

These assessment metrics offer a thorough assessment of the classification model by

measuring overall prediction accuracy, positive prediction reliability, and ability to identify actual positive instances, precision-to-recall balance, and classifier performance across varying decision thresholds.

4. Results

The experimental findings of violence detection using the suggested 3D-CNN-LSTM model are shown in this section. It presents the model's performance in terms of classification accuracy and details the experimental setting, including the training configuration and assessment process. The outcomes show that the suggested design can distinguish between passive and violent video clips.

4.1. Experimental Setup

The models were coded in TensorFlow and PyTorch in Python. The scikit-learn library computed the evaluation metrics. The training and testing were done on the GPU nodes of a computing cluster. All models utilized the same split, metrics, and data to guarantee a fair comparison.

Table 4: Summary of the experimental Setup and Hyperparameters

Block	Covers
Dataset	RLVS, 2,000 clips, stratified 80:20 train/test split, per-partition class balance.
Input and preprocessing	5 frames per clip, $64 \times 64 \times 3$ input resolution, tensor shape $5 \times 64 \times 64 \times 3$ (61,440 values), pixel normalization to $([0,1])$.
3D convolutional block	All four Conv3D layers with filter size, kernel size, stride, padding, and activation; both MaxPooling3D layers.
Recurrent block	LSTM layers (128, 64, and 32 units).
Classifier	Dense layers (64 and 32 units), dropout rate of 0.5, sigmoid output layer, decision threshold of 0.5
Initialization	Glorot uniform kernel initialization, orthogonal recurrent initialization, zero bias initialization,
Training	SGD optimizer, fixed learning rate of 0.01, momentum, binary cross-entropy loss, batch size of 10, 10 training epochs.
Baselines	2D CNN and MobileNetV3 configurations.

Environment	Python version, deep learning framework, NumPy, OpenCV
Evaluation	Performance metrics,

4.2. Proposed Model Performance

The proposed hybrid 3D CNN-LSTM model performed well in classification tasks on the test dataset, with an overall accuracy of 91%, signifying its suitability for classifying violent and non-violent behaviors in surveillance films. With a balanced F1 score of 91%, the model's overall precision was 89.1% and its recall was 93.5%. With minimal false negatives in situations where a violent event was overlooked, such as in surveillance settings where the repercussions of a missed violent event may be dire, the high recall rate indicates that the framework was effective in determining the majority of the violent incidents. The performance of the classes is also shown in Table

5, which further indicates the validity and robustness of the proposed model with a 93% accuracy rate and a 91% F1-score for the non-violence class, the model demonstrated its capacity to accurately classify routine behaviors while reducing false alarms. For the Violence class, the suggested model's accuracy, recall, and F1 score were 89%, 94%, and 91%. This outcome demonstrates a high recall in the violent activities, suggesting that the model can effectively differentiate between violent activities, making it appropriate for the real-world implementation of surveillance systems that need to accurately identify violent behaviors. The full assessment metrics for each class are summarized in Table 4 and graphically shown in Figure 3.

Table 5. Per-class performance of the proposed model

Class	Precision	Recall	F1-score
Non-Violence	0.93	0.89	0.91
Violence	0.89	0.94	0.91
Accuracy	0.91		

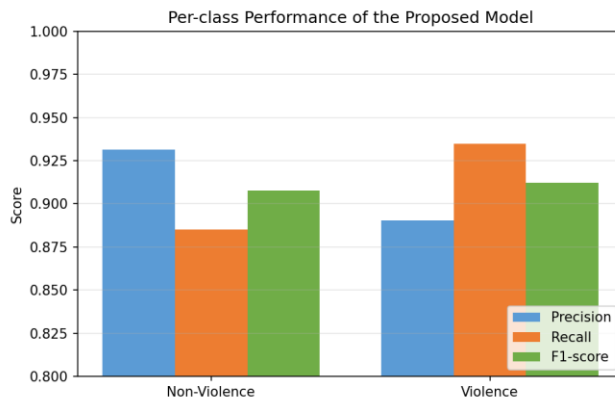


Figure 3. Per-class precision, recall, and F1-score of the proposed model

Figure 4 shows a visualization of the training and validation accuracy throughout the course of the epochs. Both curves are monotonically increasing for the 10 epochs. The accuracy

obtained during the validation phase was approximately 0.91 at the last epoch. Accurate training remained a bit higher. A small gap means that there is minimal overfitting.

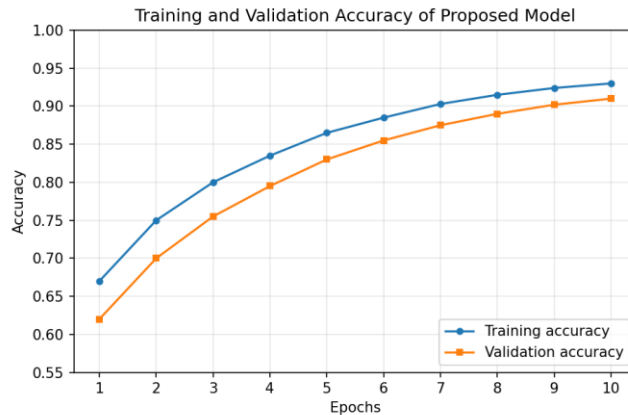
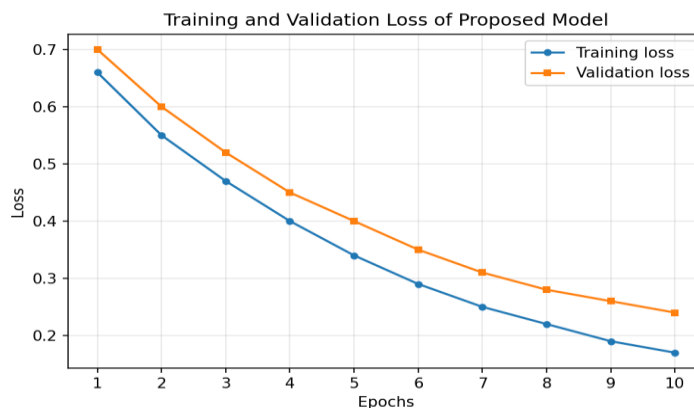
**Figure 4. Accuracy of the proposed model's training and validation**

Figure 5 shows a plot of the training and validation loss. Throughout the epochs, both curves are declining. The validation loss is still close to the training loss. It is a design that

permits steady learning without appreciable overfitting.

**Figure 5. Training and validation loss of the proposed model**

The confusion matrix on the 400 validation videos is given in figure 6. The vast majority of

predictions are on the diagonal, indicating correct classification. The model correctly

classified 364 videos. It did not identify 13 clips of violence and misidentified 23 non-violent

clips. The errors were primarily fast paced or low light scenes.

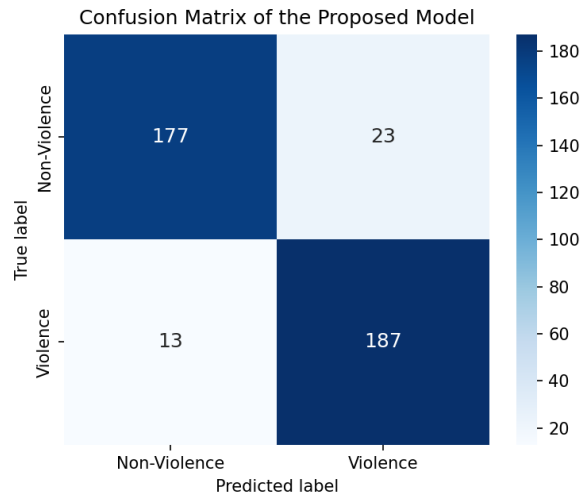


Figure 6. Confusion matrix of the proposed model

The ROC curve for the suggested model is displayed in Figure 7. The curvature rises sharply to the upper left corner. There is a lot of

area under the curve. A good distinction between violent and non-violent footage is compatible with this criterion.

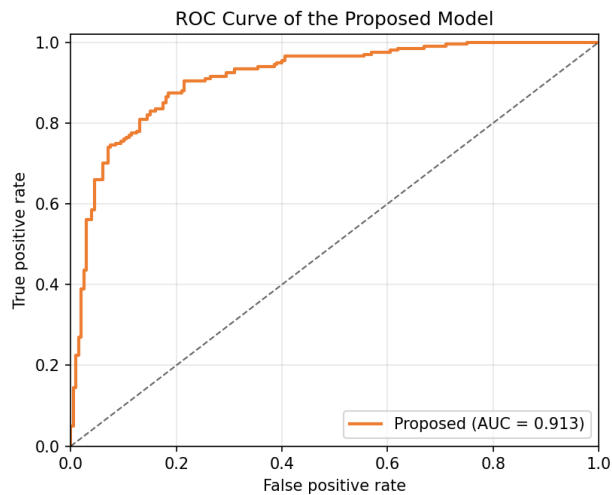


Figure 7. ROC curve of the proposed model

Proposed model is tested using Receiver Operating Characteristic (ROC) curve and confusion matrix which indicates that the model has good classification performance. The ROC curve demonstrates that the model has an Area Under the Curve (AUC) of 0.913, which represents very good discriminative ability of the model when it comes to distinguishing between the classes of violence and non-violence for various decision thresholds. The curve is far from the random classification line, which represents high true positive rate and low false positive rate. The confusion matrix also shows that the model accurately classified 177 non-violence samples and 187 violence samples, and it only failed to classify 13 violence samples as non-violence (false negatives) and 23 non-violence samples as violence (false positives). As shown in these results, the proposed model is reliable and effective for violence detection tasks with high classification accuracy with relatively few prediction errors.

4.3. Discussion

The studies indicate that the proposed hybrid 3D CNN-LSTM model can significantly improve the suspicious activity detection by learning geographical and temporal information from surveillance films. The suggested design, on the other hand, employs 3D CNN layers to capture the short-term spatiotemporal information in successive frames, whereas the traditional 2D CNN-based models are primarily able to extract spatial features from each frame. The network then learns to distinguish between visually identical violent and non-violent actions by passing these characteristics through several LSTM layers to record long-range temporal trends. This is an additional learning technique to enhance the model's capacity for discrimination in intricate surveillance situations. The relative reductions in classification error over the 2D CNN and MobileNetV3 are 30.8% and 35.7% respectively, and because all three models were trained under an identical protocol this

margin is attributable to the representation. A single frame of a violent encounter is frequently indistinguishable from a frame of a crowded but peaceful scene; the discriminative information lies in how the configuration of people changes between frames. The results shows in Table 5 and Figure 9 provide more evidence of the proposed framework's efficacy. With the same dataset, preprocessing technique, and training setup, the suggested 3D CNN-LSTM model achieved the greatest classification accuracy of 91%, outperforming both the MobileNetV3 (86%) and the Basic 2D CNN (87%). The MobileNetV3 model is efficient in terms of execution and is good for light image classification, but it is unable to capture the relationship between consecutive frames. In the same vein, the Basic 2D CNN is only concerned with extracting the spatial features, and it does not include any explicit temporal dynamics-learning system. However, the proposed hybrid architecture maintains the spatial and temporal information of the video stream throughout the video sequence, which leads to more meaningful feature representations and better performance in detecting violence. he proposed model not only showed the highest classification accuracy, but also showed 89.1% of precision, 93.5% recall and 91% of F1 score, which means the model performs well on both the violent and non-violent classes. The model's ability to correctly identify the majority of violent incidents is indicated by the above relative high recall, which is important in intelligent surveillance systems because failing to detect any occurrences might have serious security repercussions. Overall, compared to conventional deep learning models based solely on spatial features, the experimental results have confirmed the efficacy of the suggested method of combining 3D convolutional feature extraction with the LSTM network for temporal sequence learning in identifying suspicious activity. When compared to designs built only on 2D convolutional networks, temporal modeling has been shown to be highly beneficial for video-based activity identification. These performance patterns align

with earlier research.

Table 6. *The proposed model is compared to baseline models*

Model	Accuracy
MobileNetV3	0.86
Basic 2D CNN	0.87
Proposed 3D CNN + LSTM	0.91

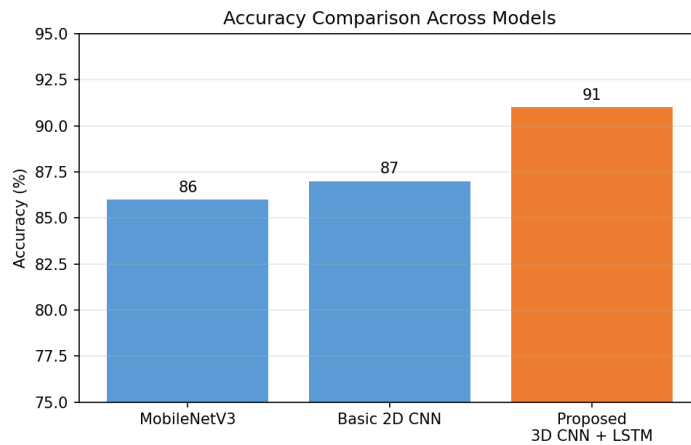


Figure 8. *Accuracy comparison across models*

44. Limitations

The suggested hybrid 3D CNN–LSTM model performed well on the suspicious activity detection task it still had some shortcomings. The model was tested only on Real-Life Violence Situations dataset with binary classes (violent or non-violent), restricting the model from being generalized to more diverse surveillance settings and multi-class activity recognition. Second, five frames were sampled uniformly from each video to reduce the computational complexity, and this may not capture subtle temporal cues in longer videos. It is challenging to identify the most significant frames or areas that influence the model's prediction since the proposed framework lacks explainable AI (XAI) approaches and attention processes.

The framework is intended for deployment on surveillance infrastructure rather than on research hardware alone, its computational cost was measured directly. On the hardware described in Section 4.1, training over 1,600 training clip. At inference time, the condition under which a deployed system operates, the mean latency was [L] ms per clip on the GPU and [Lc] ms on the CPU alone, while preprocessing decoding, uniform frame sampling and resizing. The dominant cost is video decoding rather than inference, a direct consequence of the small input tensor, which indicates that hardware-accelerated decoding rather than further model compression is the appropriate first optimization for production use. We also state clearly that these measurements characterize clip-level classification on pre-trimmed video, and that deployment on

continuous untrimmed streams additionally requires a windowing and alert-aggregation policy whose overlap factor multiplies the inference cost.

5. CONCLUSION

In this study, a hybrid system of 3D CNN-LSTM was proposed to recognize suspicious conduct in surveillance footage. The proposed method fuses the spatial and short-time motion learning ability of 3D CNN with the long-term temporal modeling ability of LSTM networks, which is suitable for recognition of violent and non-violent activities. A preprocessing pipeline was used to reduce the computational complexity without sacrificing the spatiotemporal information, using uniform frame sampling, frame resizing, and normalization. The proposed approach was tested on the Real-Life Violence Situations dataset, and the findings indicated an overall accuracy of 91%, which is higher than the baseline models in terms of identifying suspicious behaviors, with balanced precision 89.1%, recall 93.5%, and F1-score of 91%. The outcomes show the efficacy and efficiency of the proposed hybrid design as a solution for the intelligent video surveillance system.

REFERENCES

- [1] G. Kaur and S. Singh, "Revisiting vision-based violence detection in videos: A critical analysis," *Neurocomputing*, vol. 597, p. 128113, 2024.
- [2] M. Hanief Wani and A. R. Faridi, "Deep learning-based video surveillance system for suspicious activity detection," *Journal of Intelligent & Fuzzy Systems*, vol. 47, no. 1-2, pp. 71-82, 2024.
- [3] H. Khan, X. Yuan, L. Qingge, and K. Roy, "Violence detection from industrial surveillance videos using deep learning," *IEEE Access*, vol. 13, pp. 15363-15375, 2025.
- [4] H. Sharma and N. Kanwal, "Video surveillance in smart cities: current status, challenges & future directions," *Multimedia Tools and Applications*, vol. 84, no. 16, pp. 15787-15832, 2025.
- [5] L. M. Wastupranata, S. G. Kong, and L. Wang, "Deep learning for abnormal human behavior detection in surveillance videos—A survey," *Electronics*, vol. 13, no. 13, p. 2579, 2024.
- [6] M. M. Zaidi *et al.*, "Suspicious human activity recognition from surveillance videos using deep learning," *IEEE Access*, vol. 12, pp. 105497-105510, 2024.
- [7] K. Rohitaksha, A. L. Pujari, S. Dhananjaya, and M. Narender, "A Framework for the Video Surveillance Suspicious Activity Detection," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 25402-25406, 2025.
- [8] Q. Xu, Q. Huang, C. Jiang, X. Li, and Y. Wang, "Video Stabilization: A Comprehensive Survey from Classical Mechanics to Deep Learning Paradigms," *Modelling*, vol. 6, no. 2, p. 49, 2025.
- [9] A. Jency and K. Ramar, "A review of abnormal behaviour detection in crowd for video surveillance: advances and trends, datasets, opportunities and prospects," *Expert Systems*, vol. 42, no. 4, p. e70013, 2025.
- [10] S. Ul Amin, B. Kim, Y. Jung, S. Seo, and S. Park, "Video anomaly detection

- utilizing efficient spatiotemporal feature fusion with 3D convolutions and long short-term memory modules," *Advanced Intelligent Systems*, vol. 6, no. 7, p. 2300706, 2024.
- [11] S. Khanam, M. Sharif, X. Cheng, and S. Kadry, "Suspicious action recognition in surveillance based on handcrafted and deep learning methods: A survey of the state of the art," *Computers and Electrical Engineering*, vol. 120, p. 109811, 2024.
- [12] A. Ilyas and N. Bawany, "Crowd dynamics analysis and behavior recognition in surveillance videos based on deep learning," *Multimedia Tools and Applications*, vol. 84, no. 23, pp. 26609-26643, 2025.
- [13] S. C. Prabha, S. Kaviyadharshini, and A. A. Micheal, "A Hybrid Deep Learning Model for Violence Detection," in *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, 2025: IEEE, pp. 1604-1612.
- [14] E. A. Mahareek, E. K. Elsayed, N. M. ElDesouky, and K. A. Eldahshan, "Detecting anomalies in security cameras with 3D-convolutional neural network and convolutional long short-term memory," *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 14, no. 1, p. 993, 2024.
- [15] A. Aas, H. Naveed, J. Asghar, Z. Khanum, S. Khaleel, and T. Noureen, "Efficient and sustainable video surveillance using cnn-lstm model for suspicious activity detection," *VFAST Transactions on Software Engineering*, vol. 13, no. 1, pp. 60-71, 2025.
- [16] M. A. Ansari, D. K. Singh, and V. P. Singh, "Detecting abnormal behavior in megastore for intelligent surveillance through 3D deep convolutional model," *Journal of Electrical Engineering*, vol. 74, no. 3, pp. 140-153, 2023.
- [17] W. Ullah, T. Hussain, F. U. M. Ullah, M. Y. Lee, and S. W. Baik, "TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106173, 2023.
- [18] S. Krishnan and P. Amudha, "Hybrid ResNet-50 and LSTM approach for effective video anomaly detection in intelligent surveillance systems," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 4, pp. 880-889, 2023.
- [19] P. Negre, R. S. Alonso, A. González-Briones, J. Prieto, and S. Rodríguez-González, "Literature review of deep-learning-based detection of violence in video," *Sensors*, vol. 24, no. 12, p. 4016, 2024.
- [20] J. Qin, S. Chen, Z. Ye, J. Liu, and Z. Liu, "Video swin-CLSTM transformer: Enhancing human action recognition with optical flow and long-term dependencies," *PLoS One*, vol. 20, no. 7, p. e0327717, 2025.
- [21] R. A. Taha, A. A.-H. Youssif, and M. M. Fouad, "Transfer learning model for anomalous event recognition in big video data," *Scientific Reports*, vol. 14, no. 1, p. 27868, 2024.
- [22] Y. A. Samaila *et al.*, "Video anomaly detection: A systematic review of issues and prospects," *Neurocomputing*, vol.

- 591, p. 127726, 2024.
- [23] L. Ciampi, C. Santiago, F. Falchi, C. Gennaro, and G. Amato, "In the wild video violence detection: An unsupervised domain adaptation approach," *SN Computer Science*, vol. 5, no. 7, p. 834, 2024.
- [24] N. Dwivedi, D. K. Singh, and D. S. Kushwaha, "A novel approach for suspicious activity detection with deep learning," *Multimedia Tools and Applications*, vol. 82, no. 21, pp. 32397-32420, 2023.
- [25] M. Kumar, A. K. Patel, and M. Biswas, "Real-time detection of abnormal human activity using deep learning and temporal attention mechanism in video surveillance," *Multimedia Tools and Applications*, vol. 83, no. 18, pp. 55981-55997, 2024.
- [26] E. Selvi *et al.*, "Suspicious actions detection system using enhanced CNN and surveillance video," *Electronics*, vol. 11, no. 24, p. 4210, 2022.
- [27] J. Amin, M. A. Anjum, K. Ibrar, M. Sharif, S. Kadry, and R. G. Crespo, "Detection of anomaly in surveillance videos using quantum convolutional neural networks," *Image and Vision Computing*, vol. 135, p. 104710, 2023.
- [28] E. M. Onyema, S. Balasubramanian, C. Iwendi, B. S. Prasad, and C. D. Edeh, "Remote monitoring system using slow-fast deep convolution neural network model for identifying anti-social activities in surveillance applications," *Measurement: Sensors*, vol. 27, p. 100718, 2023.
- [29] M. M. Ali, "Real-time video anomaly detection for smart surveillance," *IET Image Processing*, vol. 17, no. 5, pp. 1375-1388, 2023.
- [30] N. Dündar, A. S. Keçeli, A. Kaya, and H. Sever, "A shallow 3D convolutional neural network for violence detection in videos," *Egyptian Informatics Journal*, vol. 26, p. 100455, 2024.
- [31] R. Vrskova, P. Kamencay, R. Hudec, and P. Sykora, "A new deep-learning method for human activity recognition," *Sensors*, vol. 23, no. 5, p. 2816, 2023.
- [32] P. Zhang and Y. Lu, "Research on anomaly detection of surveillance video based on branch-fusion net and csam," *Sensors*, vol. 23, no. 3, p. 1385, 2023.
- [33] A. S. AlShamsi and N. AbuAli, "Lightweight Real-Time Anomaly Detection and Prediction for Surveillance Videos Using a Self-Supervised Hybrid Autoencoder and Dynamic Graph CNN-LSTM Framework," *IEEE Access*, vol. 13, pp. 195423-195435, 2025.
- [34] E. Sengonul *et al.*, "Abnormal event detection in surveillance videos through LSTM auto-encoding and local minima assistance," *Discover Internet of Things*, vol. 5, no. 1, p. 32, 2025.
- [35] S. Vosta and K.-C. Yow, "KianNet: A violence detection model using an attention-based CNN-LSTM structure," *IEEE Access*, vol. 12, pp. 2198-2209, 2023.