Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.075960

ARTICLE

LRCN-Enabled UAV Surveillance System for Suspicious Human Activity Recognition in Smart Cities

Armaghan Azam[1,#], Arshad Iqbal[1,2,*,#], M. Mohsin Khan[1,2], Naveed Ahmad[3] and Mohamad Ladan[3]

[1]School of Computing Sciences, Pak-Austria Fachhochschule, Institute of Applied Sciences and Technology (PAF-IAST), Mang, Haripur, Pakistan

[2]Sino-Pak Center for Artificial Intelligence (SPCAI), Pak-Austria Fachhochschule, Institute of Applied Sciences and Technology (PAF-IAST), Mang, Haripur, Pakistan

[3]College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia

*Corresponding Author: Arshad Iqbal. Email: arshad.iqbal@spcai.paf-iast.edu.pk

[#]These authors contributed equally to this work

Received: 11 November 2025; Accepted: 06 January 2026; Published: 08 May 2026

ABSTRACT: Public safety and security remain critical concerns in urban environments. Detecting suspicious activities in densely populated areas poses significant challenges for modern smart cities due to occlusions, limited fixed-camera coverage, and the dynamic nature of large crowds. To address this problem, this paper proposes a Artificial Intelligence (AI)-driven unmanned aerial surveillance framework for proactive monitoring and abnormal activity recognition. The system leverages an Long-term Recurrent Convolutional Network (LRCN)-enabled architecture capable of extracting spatiotemporal patterns from aerial video streams, allowing it to detect suspicious behavior with high precision. Three deep learning models are comparatively evaluated: (i) Visual Geometry Group 16 (VGG-16) with Long Short-Term Memory (LSTM) network, (ii) a Motion Influence Map (MIM)-based approach, and (iii) LRCN. Among them, the LRCN model demonstrated superior performance, achieving an accuracy of upto 88% and outperforming the comparative methods in F1-score, Precision, Area Under the Curve (ROC-AUC), and Matthews Correlation Coefficient (MCC). Unlike traditional ground-based CCTV systems, the proposed Unmanned Aerial Vehicle (UAV)-based framework provides wider field-of-view, higher scalability, and improved visibility in dense crowds. Extensive experiments were performed to validate the practicality and robustness of the approach. The empirical findings confirm that the LRCN model effectively identifies and categorizes suspicious activities in real-world, densely populated smart-city environments. Overall, this study presents a novel and scalable aerial surveillance solution that enhances situational awareness, strengthens public-safety infrastructure, and contributes to the development of safer smart cities.

KEYWORDS: Deep learning; suspicious action detection; AI; UAS detection

1 Introduction

Artificial Intelligence (AI)-enabled Unmanned Aerial System (UAS) technology has transformed the domain of public safety and security, specifically in smart cities [1,2]. The utilization of UAS equipped with state-of-the-art AI algorithms has redefined the landscape of surveillance, monitoring, and rapid response [3]. This technological advancement is immensely significant in terms of identifying and reacting to suspicious activity in crowded areas, where traditional monitoring methods are frequently ineffective [4]. Custom-built drones to collect large video data are crucial for surveillance flights in densely populated

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

smart cities [5]. These drones are enabled with high-definition video cameras to record clips utilizing proper attention to detail and excellent sharpness of video in various crowded environments. This data is utilized as the key source for developing machine learning models to handle complex real-world dynamic situations [6,7].

The collected video clips require rigorous data annotation process, which enhances data quality and sensitivity. Deep learning models are trained through this data with the capacity to identify and discriminate a diverse set of patterns, e.g., maneuvers, and activities, in the crowd. The context varies in smart cities from congested public spaces to transport hubs. This variation in the dataset help in learning different situation related to security [8]. The inclusion of this contextual variety in the dataset ultimately results in robust and flexible training of the machine learning models. The extensive strategy guarantees that the models are appropriately prepared to these situations of vulnerabilities and contrasts in swarm environments. The trained models on this data improves the capabilities of surveillance and security in challenging smart cities [9,10].

Long-Term Recurrent Convolutional Network (LRCN), a suitable deep learning model in learning spatiotemporal characteristics. The model is trained to analyze annotated video data. For this purpose, individual frames are processed at specific time intervals. The training process of the LRCN model is repeated which is a resource-intensive and cumbersome that allows the LRCN model to update its capacity to interpret sophisticated human behaviors and human operations on video continuously [11].

The strength of LRCN originates from its simultaneous approach to handling spatial and temporal characteristics. It enables the capture of static characteristics and dynamic changes in real-world congested conditions that develop over time in smart cities. While the model continuously learns features from the annotated video which distinctly enhances the efficiency of pattern recognition.

The analysis of suspicious activity gathered through the surveillance video via UAS embedded with state-of-the-art machine learning models is a challenging task [12]. Specifically, the two components, i.e., UAS and learning models, interweave in one system used to detect and classify suspicious activities in the densely-populated premises. The primary strength of the integration is the global extensiveness in prototypes of threat identification. The data streams collected during UAS flights require extensive training and testing capability to enhance the detection performance of learning models. The interaction between the two elements generates a continuous learning cycle. The learning models gradually gain information about large numbers of suspect actions taking place in tight quarters and become better at recognizing and categorizing these dependent actions [13,14]. Machine learning-enabled UAS systems possess the potential to employ learning algorithms leading to enhanced performance in the identification and classification of suspicious actions. This approach provides a background for aerial surveillance. For this purpose, a solution is proposed using annotated video data gathered from drones for action recognition [15]. Similarly, a cooperative intelligent system for human action detection in dynamic and hostile urban areas is proposed aiming to deploy it on edge devices [16]. However, the collection of local datasets, preparation of the intelligent system for the indigenous scenarios, model training and testing on the collected data, and its deployment on the UAS is a challenging task. Considering this, an aerial system with embedded machine-learning models is required for suspicious activity detection in crowded areas. Therefore, the LRCN-enabled UAS framework is proposed to identify suspicious activity in densely populated environments.

Customized UAS equipped with video capturing and embedded edge devices enabled with LRCN are developed for critical surveillance in densely populated environments. In addition to that two other models, i.e., Visual Geometry Group-16 (VGG-16) with Long Short-Term Memory (LSTM), and the Motion Influence Map (MIM), are employed to analyze the effectiveness of suspicious action identification. The performance of these models is compared with the conventional methods using key metrics including F1 score, Precision,

the Receiver Operating Characteristic Area Under the Curve (ROC_{AUC}), and the Matthews Correlation Coefficient (MCC) leading to the identification of an effective model for identification of suspicious actions within crowded environments. Specifically, this paper proposed machine learning enabled customized UAS embedded with edge devices to use real-time as well as collected video to identity suspicious action. The proposed method shows significant enhancement over the conventional methods.

Our key contributions are listed as follows:

- An LRCN-enabled unmanned aerial surveillance framework is proposed to identify suspicious activity in densely populated environments of smart cities, considering videos' spatiotemporal features. The proposed AI surveillance system collects and processes data to identify suspicious activity in smart cities.
- Three distinct machine learning models, namely LRCN, VGG- 16 with LSTM, and the MIM, are integrated with the aerial surveillance system to analyze the activity for suspicion detection in densely crowded areas of smart cities.
- A rigorous performance comparison of the models is performed using key performance metrics, i.e., Accuracy, Precision, ROC, MCC and detection rate to identify an enhanced model for activity detection.
- The proposed scheme significantly contributes by introducing a novel approach that combines a localized video capture stage using unmanned aerial stations, enhancing data collection in crowded areas of smart cities. Additionally, the systematic testing and comparison of three distinct models architectures succh LRCN, VGG-16 with LSTM, and MIM comparatively enhance the surveillance systems in smart cities.

2 Related Work

Significant contributions have been made for the suspicious activity classification using various machine learning models. A VGG-16 and Support Vector Machine (SVM) models is proposed for Human Activity Recognition (HAR) using Convolutional Neural Network (CNN)'s VGG16 pre-trained model [17]. For this purpose, a UniMiB dataset's accelerometer data is used where the model has achieved up to 79.55% accuracy and 71.63% F Score. Similarly, a novel two-stream CNN model is proposed aiming to enhance abnormal human behavior recognition by combining motion history images and RGB data. The model leverages VGG-16 for motion history image training and Faster R-CNN with Kalman filter-assisted data annotation for RGB image training. The results achieved through models across diverse datasets demonstrate superior accuracy compared to existing schemes [18]. Furthermore, a fusion convolution architecture is proposed using semi-CNN, for efficient spatio-temporal feature learning in video action recognition. By combining 1D, 2D, and 3D convolutions, it outperforms equivalent 3D models on the UCF-101 dataset, achieving a 16%–30% boost in top-1 accuracy with reduced parameters and mitigating overfitting.

Semi-CNN integrates pre-trained 2D CNNs is proposed by employing VGG-16, ResNets, DenseNets for spatial extraction and temporal encoding [19]. The proposed method tackles GPU memory constraints in end-to-end learning for video action recognition using CNN features. By treating deep networks as local feature extractors and aggregating local features, the framework improves robustness to noisy labels, achieving significant performance gains on HMDB51 and UCF101 datasets with a straightforward maximum pooling approach for sparse local features [20]. Yao et al. have extensively explored advancements in video action recognition through CNN integration, addressing the challenge of extending 2D spatial features to 3D spatio-temporal signals. Strategies include 3D CNN, incorporating motion-related information, and fusion methods [21].

Deep learning-based human identification is proposed where 2D-CNN, VGG-16, and ResNet50 are employed to achieve an accuracy of 82.96%, 81.84%, and 80.03% [22]. However, the 2-layer CNN on Kaggle/real-time video achieved higher accuracy than pre-trained VGG-16, which is 72.20% while ResNet50

with transfer learning performed better than VGG-16 transfer learning with an accuracy of 99.18% vs. 98.36%, resulting in overfitting of the model. Gawande et al. proposed a robust deep learning system for pedestrian detection, tracking, and suspicious activity recognition to address growing security threats [23]. The pedestrian dataset encompasses various behaviors, enabling uniform annotation analysis with state-of-the-art recent deep-learning approaches in vision-based surveillance.

A YOLOv5 model-based method is proposed for crime detection, aiming to mitigate shoplifting crime activities [24]. The method enhances up to 8.45-fold detection, which is much faster than the Robust Temporal Feature Magnitude learning (RTFM) baseline utilizing the UCF Crime dataset. It performs 3% better in F1 score, irrespective of expensive data augmentation or image feature extraction. Drones are an essential component in smart cities by providing an accidental response and reconnaissance. A Hybrid Data Augmentation Method (Wasserstein GAN+CNN-LSTM) approach for Human Activity Recognition is proposed, considering aerial surveillance that achieves an accuracy of 84.83% [25]. The method has jointly applied CNN and LSTM human intention prediction, which captures spatial as well as temporal features that outperform the conventional methods [26]. Advanced AI and Deep Neural Network algorithms enable smart video surveillance to automatically detect any suspicious activity in a frame with enhanced reliability [27]. Bakirci demonstrated a YOLOv8-enabled drone system on a Jetson Nano platform to monitor vehicle mobility [28].

Deep learning model-based methods are proposed with a slow-fast algorithm to automatically predict anomalies in monitored regions of interest in CCTV systems intended for augmenting cyber-physical security [29,30]. This model focuses on identifying objects, suspect motion and activities based on previously recorded feeds and real-time videos. Abdullah and Jalal proposed a method by leverages semantic segmentation to extract images from the background extraction, a combination of object activity analysis along with crowd counting in public spaces and tracking the objects, and optimally describes the spatiotemporal activity to enhance anomaly detection [31]. Therefore, the proposed method achieved high accuracy rates over UCSD (84.8%), Mall (89.16%), UMN (82.5%), and MED datasets.

Global-Local Attention (GLA) mechanism is proposed for improving video representation in the context of temporal action detection [32]. This strategy can cleanly separate between action classification and localization tasks, which achieves improved results on benchmark datasets such as THUMOS'14 and ActivityNet-1.3. The GLA model as well, with 1373 frames per second (FPS) inference speed on a single Nvidia Titan Xp GPU [33]. Human Activity Recognition (HAR) is a challenging task. For this purpose, a new transfer learning framework is proposed to recognize user-invariant and subject-specific characteristics allowing accuracy improvements reaching 43% and 66.6% gain in training time. On the Hardware side, the model trained from early layers using sensor measurements achieves up to 43% increased efficiency [34,35]. Deep learning methods is proposed where Conv-LSTM and LRCN are employed for effective prediction of human action considering time series data. The CNN-LSTM network enhanced accuracy up to 82% that is also implementable on mobile and wearable computing applications [36].

Based on these increased criminal and suspicious activity incidences, this paper proposed a deep learning-based behavior recognition model-namely the LRCN model by combining CNN with LSTM techniques. The method can efficiently sort suspicion activities with a realistic accuracy level close to 80.55% [37]. Using the CNN, LSTM, and LRCN models, this proposal successfully distinguishes and discovers the aberrant and criminal activities in the videos with an accuracy of 87.8%. The proposed model can be deployed in real-time fields such as public places and hospitals, where it ensures accurate surveillance and timely action, improving security and safety [38]. Veenu et al. proposed an intelligence drone system to put together a fast, inexpensive method for detecting human behavior appropriate for processions and other surveillance scenarios [39].

AI-equipped UAS can enhance law enforcement's situational awareness of public events. Enhancing crowd assessment and response methods can assist in the peaceful conduct of events and possible threat identification, by real-time monitoring [40]. The human action detection dataset was recorded through a drone in an outdoor environment, where it includes 240 high-definition video clips from 13 dynamic actions with over 66,919 frames [41]. It can be used for various potential applications, including action recognition, surveillance, and situational awareness in poor visual conditions with low altitudes and speeds or gait analysis. A real-time spatial-temporal depth separable CNN for crowd analysis in videos is introduced, where a rapid processing speed of 3.4 ms per frame is achieved [42]. Using a dataset of 41,000 manually annotated frames, the proposed model accurately captures crowd dynamics [43]. Despite its effectiveness, manual frame annotation is noted as time-consuming. However, a significant advancement in crowd analysis implications for real-world applications is important [44], where an automated annotation process improves the model development [45]. Recently, a Transformer-enhanced Dual-Stream Network (TDS-Net) is proposed where a Dual Stream Transformer Network that jointly learns RGB and motion features for accurate video anomaly detection on benchmark datasets [46]. Similarly, an Artificial Intelligence of Things (AIoT)-based anomaly detection framework is proposed that combines on-edge lightweight classification with cloud-assisted transformer analysis, achieving high accuracy across multiple benchmark datasets [47].

3 Proposed AI-UAS Surveillance System

The proposed method utilizes machine learning framework to extract spatiotemporal features to identify suspicious activity within multi dimensional datasets. The architecture of the proposed framework is depicted in Fig. 1, integrates UAS equipped with edge devices used to operate the AI model and records videos during surveillance operations, which is refined to prepare it for analysis. Before training the model, the received and refined videos are labeled according to their corresponding classes. The primary objective is to correctly label the actions into two main classes (i) fighting actions, which are considered suspicious activity, and (ii) non-fighting activities, including walking and running, comprehending public safety, and security patrolling tasks, as illustrated in Figs. 2 and 3. Deep learning models such as LRCN, VGG-16 & LSTM model, and MIM are employed to analyze and interpret the dataset to identify patterns linked to activities in the target environment. A thorough comparison of performance metrics is conducted after training and testing of the models. These models are deployed to be utilized for detecting suspicious activities in video detection to enhance an AI-driven UAS system.

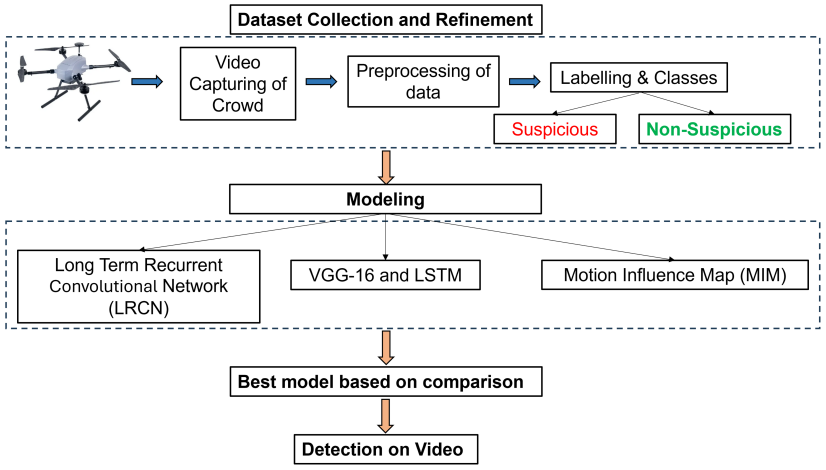

Figure 1: Architecture and implementation of the proposed framework.

Figure 2: Videos dataset containing fighting, running, and walking labeled classes [48,49].

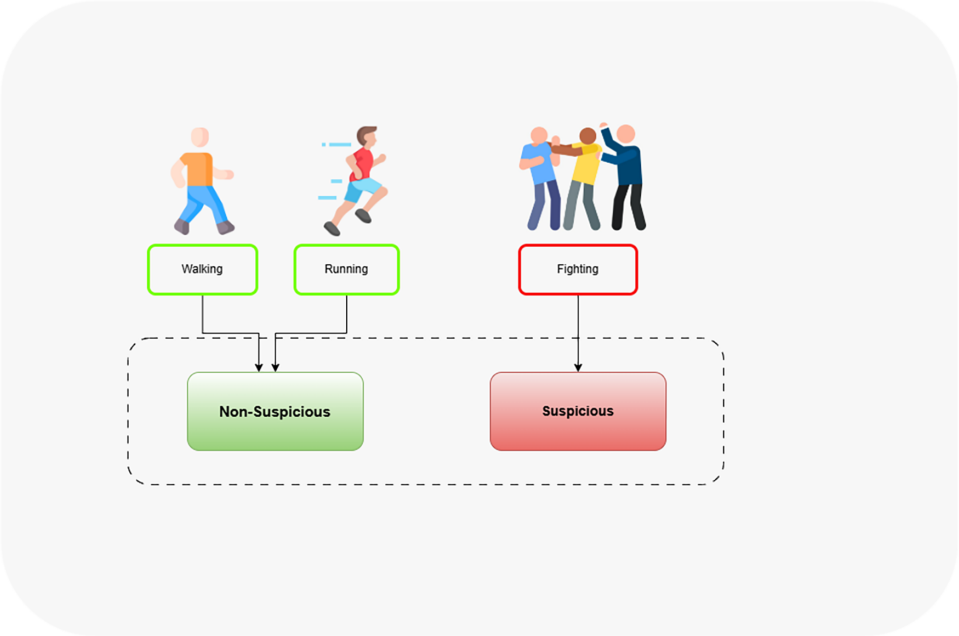

Figure 3: An example of how the classes are labeled into suspicious and non-suspicious action.

As UAS requires low-latency while processing the information, the machine learning model is optimized for low-latency outcomes. The proposed mechanism processing time while embedded on edge device is about 1 to 2 s per frame, due to UAS capability constraints. Different techniques can be used to enable this mechanism, for example, model optimization, pruning and utilizing GPUs. This techniques assisted the real-time processing and prediction of activities recorded via UAS. This helps in developing an alert system for timely detection of suspicious activity in a congested environment. The proposed mechanism is then embedded on Jetson devices placed on drones. The system also consists of YOLOv8 based object detection to process recorded clips. The YOLOv8 models helps in identifying objects in recorded video clips. Then, the LRCN model assists in actions recognition. The NVIDIA Jetson has accelerated GPU processing times and high frame rate. This helps achieving low latency enabling nearly real-time surveillance and activity classification. This setting enables UAS framework to detect suspicious behavior in a crowded area.

3.1 Dataset Collection and Description

The dataset consists of three primary classes of activities fighting, walking, and running. For the robustness and diversity of the dataset, videos are gathered from distinct sources. The fighting class is

employed from the movies-fight detection dataset [48], and walking and running are obtained from the KTH dataset [49]. The KTH dataset is a collection of sequences consisting of six distinct actions. Each action class encompasses 100 sequences, with each sequence comprising nearly six hundred frames with a frame rate of 25 frames per second (See Fig. 2). There are, approximately, 100 videos per class including fighting, running, and walking. Also, the impact of the proposed framework is validated on a locally developed dataset collected through a UAS camera in a crowded environments. The main aim of this testing is to observe the effectiveness of the proposed framework in practical situations.

3.2 Data Preprocessing and Feature Extraction

The first step in this process is an extensive frame-by-frame analysis of the video to create a feature representation. To enable useful temporal analysis, the frames are aligned and synchronized. The frames are aligned and synchronized to analyze the temporal behaviors and action recognition. The OpenCV package is employed for video processing. A sequence of 30 frames is made by choosing frames at regular intervals. Subsequently, the images are resized to meet the pixel dimension of the dataset. The frames are resized to a uniform resolution of 64 × 64 pixels to ensure architectural consistency. In addition, strategies are applied to improve frame clarity by minimizing noise and artifacts. Similarly, filters are used to suppress background noise while retaining essential visual features. Contrast normalization is also applied to ensure consistent pixel intensity values across frames, resulting in higher quality inputs. The preprocessing stage further includes background subtraction, which highlights foreground objects relevant to action prediction. Temporal smoothing methods are used to reduce jitter in frame sequences, enabling smoother video analysis. Color normalization is applied to improve the model's generalization across diverse visual environments by aligning color distributions. Together, these preprocessing steps transform the raw input data and establish a robust foundation for the subsequent phases of model training and evaluation in AI-driven surveillance systems.

Each clip in the dataset is analyzed, labeled, and assigned to one of the three main classes i.e., running, walking, or fighting. Proper Labeling is required in creating the ground truth for model training and identification of the classes, Fig. 3 shows the distribution of classes with labels. Main features are extracted from the video clips, where Deep CNN models are used to extract the spatial features, including VGG-16, which enables visual details for each frame [50]. Furthermore, the video frames' temporal features are extracted using Recurrent Neural Networks (RNNs) to capture the development of actions over time sequence. Particularly, the LSTM architecture is used for temporal analysis of events which generates the features of sequential data.

3.3 Learning Models Training, Testing and Evaluation Metrics

Deep Learning models such as LRCN, and VGG-16 with LSTM, and MIM are employed to analyze the spatial temporal analysis of video clips. LRCN is a distinctive model architecture developed to focus on maximizing the strength of spatial and temporal aspects. It can effectively detect the activities and distinguish the actions since it considers the temporal dynamics and visual context. The hyperparameters used to optimize the models are illustrated in Table 1 where the comparison of the training environment for the LRCN, VGG-16 + LSTM, and MIM models. To ensure a fair comparison, all models use the same input size, sequence length, optimizer, loss function, metrics, number of epochs, and batch size. Early stopping is applied only to the LRCN model, where training stops if accuracy does not improve for 10 consecutive epochs, helping prevent overfitting. In the training process, the models learn various features from the training set. The training set is 75% of the total dataset where internal parameters are fine-tuned. Similarly, the validation set is 25% of the total dataset used to validate the outcomes with the target results. It is achieved through the

number of iterations aiming to minimize the difference between the models' predictions and the ground truth labels. The models are evaluated using various metrics to measure their performance including accuracy, F1 score, Matthews Correlation Coefficient, Precision, and Receiver Operating Characteristic Area Under the Curve (ROC_{AUC}). These metrics are used to accurately identify suspicious action, and minimize false positive and false negative suspicious action.

Table 1: Comparative training hyperparameters for LRCN, VGG-16 + LSTM, and MIM models.

Hyperparameter	LRCN Model	VGG-16 + LSTM Model	MIM Model
Input Resolution (Height × Width)	64 × 64	64 × 64	64 × 64
Temporal Sequence Length (Frames per Clip)	30	30	30
Optimization Algorithm	Adam	Adam	Adam
Objective/Loss Function	*categorical_crossentropy*	*categorical_crossentropy*	*categorical_crossentropy*
Evaluation Metric	accuracy	accuracy	accuracy
Number of Training Epochs	50	50	50
Mini-Batch Size	4	4	4

3.3.1 Accuracy

It counts the model's percentage of accurate predictions out of all its forecasts or it refers to the ratio of the proportion of the true positives, or correct predictions of positive instances, and the predicted negatives, or the true negatives, to the total number of predictions. Mathematically,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where *TP*, *TN*, *FP*, and *FN* are true positives, true negatives, false positives, and false negatives, respectively.

3.3.2 Precision

It refers to the ability of the model to correctly predict the favorable outcome. It is the number of true positive results divided by the total of the true positive and the false positive results. In other words, when a model has a high precision, it generates few false positive predictions. Mathematically,

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

3.3.3 F Score

It is a combination of recall and precision where the former quantifies the percentage of true positives among all real positive occurrences, while the latter shows the percentage of true positives among all positive predictions. F1 score is a useful metric since it finds a compromise between precision and recall when false positives and false negatives have distinct outcomes. Mathematically,

$$F1 - Score = \frac{2 \cdot (Precision \cdot Recall)}{Precision + Recall} \tag{3}$$

where

$$Recall = \frac{TP}{TP + FN}. \tag{4}$$

3.3.4 (MCC)

It is a global metric because all four elements of the confusion matrix are considered, i.e., True Positives, True Negatives, False Positives, and False Negatives. It provides a better sense of how our model performs, particularly in scenarios involving imbalanced datasets. The MCC scores range from −1 (representing entirely incorrect predictions) to +1 (indicating flawless predictions), with 0 suggesting random predictions. Mathematically,

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

3.3.5 ROC_{AUC}

It is used to assess a model's ability to distinguish between positive and negative instances in binary classification tasks. This metric utilizes the ROC curve, which plots the true positive rate, i.e., sensitivity, against the false positive rate, i.e., specificity, at different decision thresholds. The AUC quantifies the area under the curve where a larger AUC value indicates superior model performance. The calculation of $\mathrm{ROC}_{\mathrm{AUC}}$ entails summing the areas of trapezoids formed by connecting consecutive points on the ROC curve. Mathematically,

$$\mathrm{ROC}_{\mathrm{AUC}} = \sum_{i=1}^{n} [TPR(i) \cdot (FPR(i) - FPR(i-1))] \tag{5}$$

where $TPR(i)$ represents the true positive rate, i.e., sensitivity, at the i-th threshold, and $FPR(i)$ is the false positive rate at the i-th threshold.

4 Deep Learning Models Architecture

Three models such as LRCN, VGG-16 with LSTM, and MIM are employed to analyze the performance of the proposed framework. The architectural and analytical detail of each model is presented below.

4.1 LRCN Model

This is a type of neural network architecture that combines CNN and RNN for analyzing sequential data, i.e., video frames. It is used to learn both spatial features as well as temporal dependencies in the video frame. The hybrid architecture of LRCN assists in the accurate recognition of actions within crowded environments. Specifically, it can extract particular features such as fighting actions from video clips labeled as suspicious and distinguish walking and running actions labeled as non-suspicious. The CNN part in the architecture of LRCN extracts spatial features which is essential for understanding the visual context of actions within individual video frames. The convolution layers in the network use filters to extract features from the input image. The operation for a single filter is expressed as,

$$C(i, j, k) = \sum_{m,n} I(i + m, j + n, k) \cdot K(m, n, k) \tag{6}$$

where $C(i, j, k)$ is the convoluted value at at i and j location in the feature map of k-th layer, $I(i + m, j + n, k)$ is a small pixel value $i + m$ and $j + n$ in k-th channel, and $K(m, n, k)$ is the weight of the internal filter at m and n in k-th channel.

The second part of LRCN, i.e., RNN, is used to extract temporal features, considering how actions evolve. For this purpose, LSTM networks are integrated in LRCN. It is a form of RNN that is well-suited for processing sequential data to capture temporal dependencies over time, such as video frames. It is an important part of identifying dynamic actions with intricate temporal dynamics. The mathematical representation of LSTM is

$$i_t = \sigma(W_i \cdot [lstm_t, h_{t-1}] + b_i), \tag{7}$$

$$f_t = \sigma(W_f \cdot [lstm_t, h_{t-1}] + b_f), \tag{8}$$

$$o_t = \sigma(W_o \cdot [lstm_t, h_{t-1}] + b_o), \tag{9}$$

$$g_t = \tanh(W_g \cdot [lstm_t, h_{t-1}] + b_g), \tag{10}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot g_t, \tag{11}$$

and

$$h_t = o_t \cdot \tanh(C_t) \tag{12}$$

where i_t represents the input gate, f_t denotes the forget gate, o_t signifies the output gate, g_t is the cell input, C_t represents the cell state, h_t is the hidden state, and W_i, W_f, W_o, W_g are weight matrices, and b_i, b_f, b_o, b_g are the bias vectors. Similarly, $lstm_t$ is the input at time t, and h_{t-1} is the previous memory.

During training phase, the LRCN model learns different actions using the hyperparameters by the iterative exposure of the frames. The layer wise LRCN architecture is listed in the Table 2. During the testing phase, the model is embedded on an edge device placed on UAS. The framework evaluate the video frames through the spatiotemporal analysis to predict different actions. The classification of different actions into suspicious and non-specious is based on fighting, and normal walking and running, respectively.

Table 2: Sequential model architecture for LRCN model.

LRCN-Layers	Operations	Input Matrix	Output Matrix	Parameters
Step-1	Conv2D(1)	30, 64, 64, 3	30, 64, 64, 3	896
Step-2	Max Pooling 2D(1)	30, 64, 64, 3	30, 16, 16, 32	0
Step-3	Conv2D(2)	30, 16, 16, 32	30, 16, 16, 64	18,496
Step-4	Max Pooling 2D(2)	30, 16, 16, 64	30, 4, 4, 64	0
Step-5	Conv2D(3)	30, 4, 4, 64	30, 4, 4, 128	73,865
Step-6	Max Pooling 2D(3)	30, 4, 4, 128	30, 2, 2, 128	0
Step-7	Conv2D(4)	30, 2, 2, 128	30, 2, 2, 256	131,328
Step-8	Max Pooling 2D(4)	30, 2, 2, 256	30, 1, 1, 256	0
Step-9	Flatten	30, 1, 1, 256	30, 256	0
Step-10	LSTM	30, 256	32	36,992
Step-11	Dense	32	3	99

4.2 VGG-16 with LSTM Model

In this scenario, the VGG-16 is followed by the LSTM model for action recognition. The spatiotemporal analysis is performed using CNN network VGG-26 and LSTM in dynamic crowded area. The CNN network learns the spatial features while the LSTM network learns the temporal features. The hierarchical structure of VGG-16 network learns the spatial feature in the frames. Lower layer focuses on edges and colors. Higher layers focus on shapes and edges. This helps the network the learning of the contextual behaviors in frames. Therefore, it leads the model to learn the contextual understanding available in frames. The middles layers use the Rectified Linear Unit (ReLU) activation function to extract positive values. The output of ReLu is

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, \text{otherwise} \end{cases} \tag{13}$$

The max pooling operation is used to conserve prominent features such as edges and textures whereas max pooling

$$P(i, j, k) = \max(I(2i, 2j, k), I(2i+1, 2j, k), I(2i, 2j+1, k), I(2i+1, 2j+1, k)). \tag{14}$$

The LSTM learns the temporal features in the video streams. The temporal aspect helps in understanding how the action evolves with time. This model uses the spatial features learned through the CNN network to learn the actions evolved with time. It preserves the dynamic natures of videos and assist model recognizes beyond static visuals.

The models are trained on the dataset consisting of fighting, walking and running classes. The data is spitted in training set and testing set with the ratio of 75% and 25%, respectively. In the iterative training, the models learns the features to reduce the gap between their predictions and targets. The architectural details of VGG-16 and LSTM models are given in Tables 3 and 4, respectively. In the models testing phase, the 25% of the dataset is used asses model generalization on unseen data. It acts as a benchmark for determining the models' capability to accurately classify actions. The models performance can be identied using various metrics including accuracy, F1 score, MCC, precision, and ROC_{AUC}. These metrics provide insights into the models' effectiveness in correctly identifying actions and their overall performance on unseen data.

Table 3: Sequential model architecture for VGG-16 model.

VGG-16 Layers	Operations	Input Matrix	Output Matrix	Parameters
Step-1	Conv2D(1)	64, 64, 3	64, 64, 64	1792
Step-2	Max Pooling 2D(1)	64, 64, 64	32, 32, 64	0
Step-3	Conv2D(2)	32, 32, 64	32, 32, 128	73,856
Step-4	Max Pooling 2D(2)	32, 32, 128	16, 16, 128	0
Step-5	Conv2D(3)	16, 16, 128	16, 16, 256	295,168
Step-6	Max Pooling 2D(3)	16, 16, 256	8, 8, 256	0
Step-7	Conv2D(4)	8, 8, 256	8, 8, 512	1,180,160
Step-8	Max Pooling 2D(4)	8, 8, 512	4, 4, 512	0
Step-9	Conv2D(5)	4, 4, 512	4, 4, 1024	2,359,808
Step-10	Max Pooling 2D(5)	4, 4, 1024	2, 2, 1024	0
Step-11	Average Pooling (2D)	2, 2, 1024	1024	0

Table 4: Sequential model creation for LSTM model.

LSTM Layers	Operations	Input Matrix	Output Matrix	Parameters
Step-1	Functional Model (VGG-16)	30, 64, 64, 3	30, 512	14,714,788
Step-2	LSTM	30, 512	256	787,456
Step-3	Dense (1)	256	1024	263,158
Step-4	Dense (2)	1024	3	3075

4.3 MIM Model

This model is used to extract dynamic motion information in video data to recognize the required action. It can be visualized as a grid representing how the motion would affect the probability that a specific action would happen at a specific location on the video frame. The feature extraction process is performed by this model by acquiring the motion features from the picture frames. It identifies how minutely the motion affects the relative occurrence of an action. The extraction may involve complex processes such as optical flow and motion

$$M(i,j) = M_m(F(i,j)) \tag{15}$$

represents the features of motion extracted at a specific location for the frame $F(i,j)$ at i and j spatial coordinates, and $M_m(.)$ is a function that calculates the magnitude of motion at a given location. The *MIM* is generated as the motion features are extracted. This map assigns a weight to each spatial location in the frame, indicating the influence of motion on the likelihood of specific actions occurring. The *MIM* is calculated as a weighted sum of the motion features from nearby locations and

$$MIM(i,j) = \sum[w(k,l) \cdot M(i+k, j+l)] \tag{16}$$

represents the value of the *MIM* at i and j, $M(i+k, j+l)$ denotes the motion feature at k and l coordinates, and $w(k,l)$ is a weight matrix that determines the influence of motion features on the *MIM*. The *MIM* enriches the model's understanding of actions by providing context-aware insights into the role of motion in crowded environments. Table 5 explains the sequential model architecture of the *MIM*.

Table 5: Sequential model architecture for MIM.

MIM Layers	Operations	Input Matrix	Output Matrix	Parameters
Step-1	Conv3D(1)	30, 64, 64, 3	28, 62, 62, 32	2624
Step-2	Max Pooling 3D(1)	28, 62, 62, 32	14, 31, 31, 32	0
Step-3	Conv3D(2)	14, 31, 31, 32	12, 29, 29, 64	55,360
Step-4	Max Pooling 3D(2)	12, 29, 29, 64	6, 14, 14, 64	0
Step-5	Conv3D(3)	6, 14, 14, 64	4, 12, 12, 128	221,312
Step-6	Max Pooling 3D(3)	4, 12, 12, 128	2, 6, 6, 128	0
Step-7	Flatten	2, 6, 6, 128	9, 216	0
Step-8	Dense (1)	9, 216	1, 024	9,438,208
Step-9	Dense (2)	1, 024	3	3075

The apparent motion of the objects and surfaces in video frames is crucial. Optical flow is a fundamental technique to quantify this analysis. The model employs this method to measure how the pixel movement

between the successive frames is depicted dynamically to ensure the motion within the scene. This process measures the changes in the image features so that the velocity and direction between the motions are determined. Therefore, the action in a dynamic setting is better understood as the pixel changes in a very detailed manner which enhances the model's understanding of how the dynamics are changing in the given scene.

Pixel Motion Tracking (PMT) Measures the movements of pixels from one frame to the next and assists in capturing the displacements of individual pixels. This technique provides the ability to fully describe the pixel dynamics of the motion and is critical for applications such as tracking of objects, vision and estimation of motion, and video editing. Dynamic Representation (DR) provides a lively representation of the movements in the environment graph, a sequence of moving objects and surfaces as they change over time. The feature allows for analysis of pixel displacements and provides insight into the spatial dynamics and temporary evolution of motion in the observed scene. Optical flow helps in analyzing the displacements of objects from consecutive frames to understand the temporal evolution of the movement using frame-to-frame analysis. This method contributes to scene understanding and recognition of events.

5 Results and Discussion

LRCN performs well as compared to other models by achieving upto 88% accuracy. On the contrast, the VGG-16 with LSTM Model achieves an accuracy of 85.3% while the accuracy of the MIM reaches upto 68%, as illustrated in Table 6. The high performance of the LRCN model validates the precisely recognizing suspicious activities. The enhanced performance of the LRCN model enabled due to the spatiotemporal feature extraction. The CNN part in this model enables the learning of spatial feature extraction while the LSTM networks enables the learning of temporal dynamics. Therefore, the LRCN model potentially learn features of the video both spatially and temporally, helps in activity recognition in congested environments.

Table 6: Comparison of LRCN, VGG-16 + LSTM, and MIM.

Model	LRCN	VGG-16 + LSTM	MIM
MCC	0.8219	0.7796	0.6155
F1 score	0.8801	0.8535	0.5680
Accuracy	0.8800	0.8533	0.6800
Precision	0.8841	0.8541	0.5077
ROC_{AUC}	0.9670	0.9591	0.7806

The accuracy over the epochs of different models such as LRCN model, MIM, and VGG-16 with LSTM are shown in Fig. 4. The LRCN outperforms the other models in term of training and validation accuracy. The training process is executed over 50 epochs, with the dataset dividied into training and validation ratio of 75% and 25%, respectively. Notably, the model demonstrates a balanced generalization, avoiding overfitting and under-fitting. This significant results indicates model has successfully learned the feature validated by the validation accuracy. The confusion matrix of the LRCN model achieved further strengthen this as mentioned in Fig. 5. The VGG-16 with LSTM model faces a potential issue of overfitting, where the model have shown a significant performance on training data but comparatively lower performance on unseen data, as illustrated in Fig. 4. The confusion matrix for VGG-16 with LSTM model has shown similar impact as shown in Fig. 6. In case of MIM model, a potential concern of underfitting is observed where a model fails to capture the underlying patterns in the data, resulting in suboptimal performance on both the training and validation sets. The confusion matrix achieved for MIM model validates this in Fig. 7.

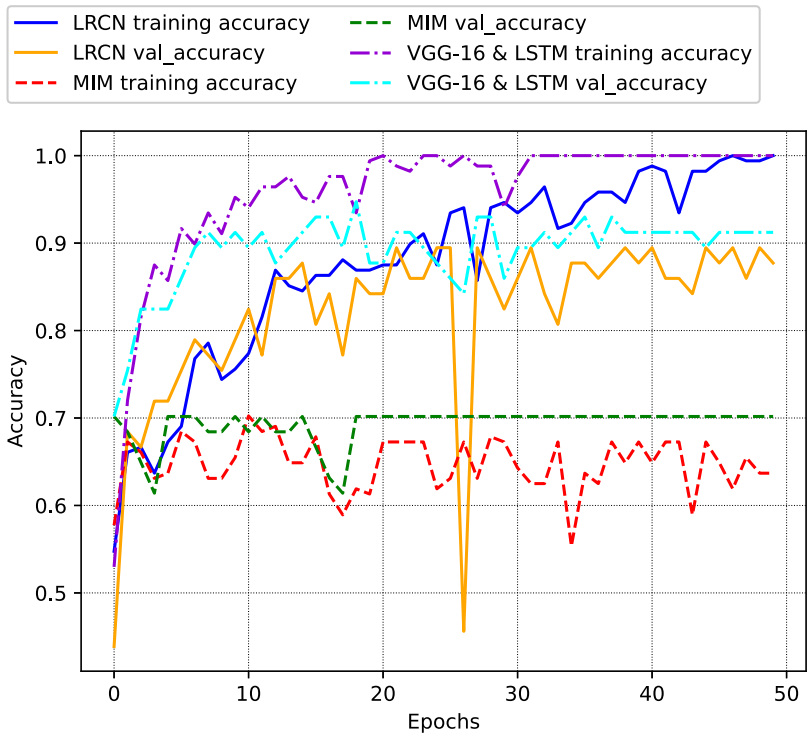

Figure 4: Combined model accuracy over epochs.

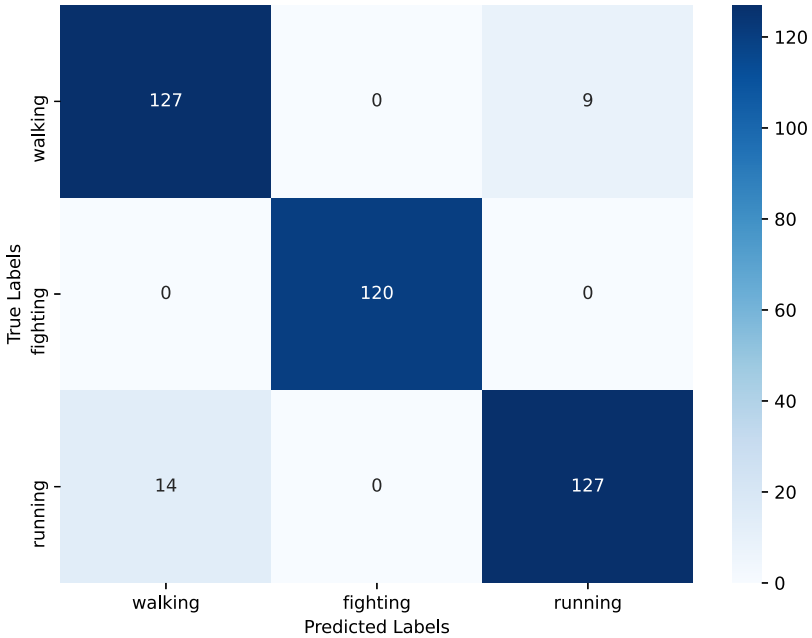

Figure 5: Confusion matrix of LRCN Model.

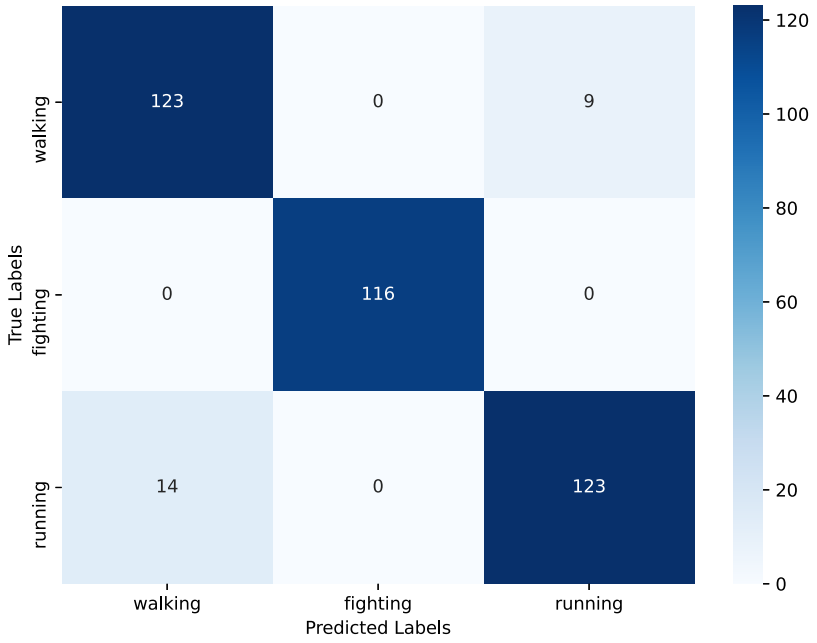

Figure 6: Confusion matrix of VGG-16 with LSTM model.

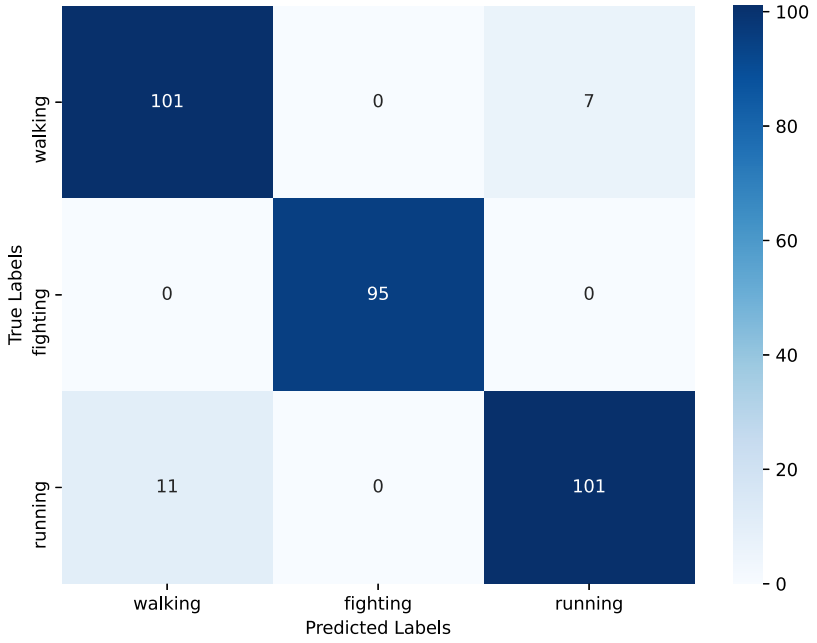

Figure 7: Confusion matrix for MIM model.

Using a structured approach to train and evaluate the drone-captured video data, the LRCN model achieves a high confidence score of 0.99 in recognizing different actions. The model reliably identifies the 'fighting' class as suspicious, while consistently classifying activities such as 'walking' and 'running' as non-suspicious on a locally collected drone-based crowd dataset. The visual results, illustrated with red and green bounding boxes in the Fig. 8, further improve interpretability by clearly showing how effectively the model distinguishes between suspicious and non-suspicious behaviors.

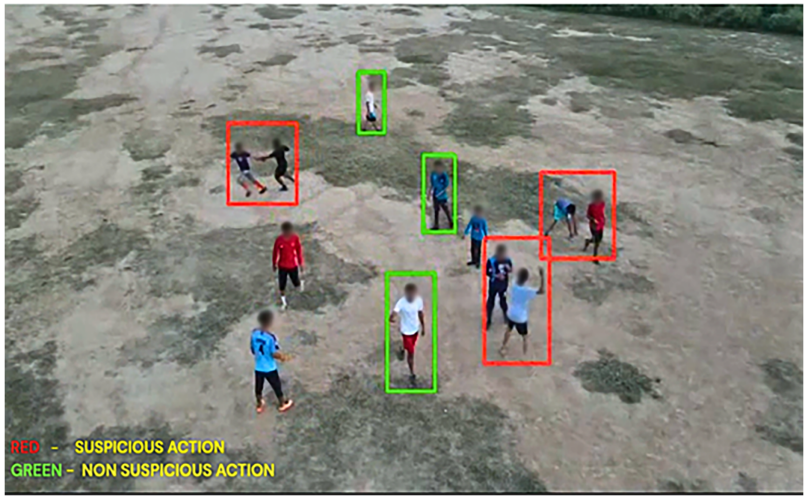

Figure 8: Recognizing suspicious and routine actions in videos recorded by drones.

6 Model Deployment and Limitation

The deep learning models such as LRCN model, are deployed on a Jetson AGX Orin 64 GB edge module, specification is listed in Table 7 integrated with the UAV platform with specification listed in Table 8. The trained LRCN model was exported to Open Neural Network Exchange (ONNX) and optimized via TensorRT with FP16, INT8 calibration evaluated. Inference was executed using a clip ingest pipeline (camera -> Model detection -> per-object buffer -> recognition). The performance on the training workstation (high-end GPU) is measured approximately 11.90 ms per sample (upto 84 samples/s). On the Jetson AGX Orin 64 GB, average inference time per frame for the LRCN pipeline is upto 105.38 ms (upto 9.49 FPS), with observed min/max values of upto 74.09 and upto 2081.61 ms, respectively; the long-tail maxima were attributable to intermittent IO/tracking overhead. These results show that LRCN can be deployed on modern embedded AI platforms with system-level optimizations, achieving near real-time performance in many single-target scenarios. For multi-target or strict-latency applications, a lightweight model configurations is required and reduced temporal windows, or further model compression (pruning/INT8) to meet operational constraints. The LRCN model requires significant computation, so its speed decreases in multi objects environment. Occasional delays are expected caused by tracking and video input fluctuations, which can momentarily slow the system during flight. Additionally, challenging conditions like low light, rapid drone movement, or dropped frames can affect recognition accuracy.

Table 7: NVIDIA Jetson AGX Orin 64 GB specifications.

Category	Specification
Model	Jetson AGX Orin 64 GB
AI Performance	275 TOPS (INT8)
GPU	NVIDIA Ampere, 2048 CUDA cores, 64 Tensor Cores
Max GPU Frequency	1.3 GHz
CPU	12-core Arm Cortex-A78AE v8.2, 3 MB L2 + 6 MB L3
CPU Max Frequency	2.2 GHz
DL Accelerator	2× NVDLA v2.0, 1.6 GHz max
Vision Accelerator	PVA v2.0

(Continued)

Table 7 (continued)

Category	Specification
Memory	64 GB LPDDR5, 256-bit, 204.8 GB/s
Storage	64 GB eMMC 5.1
CSI Camera Support	6 cameras (16 virtual), 16 lanes MIPI CSI-2 D-PHY 2.1/C-PHY 2.0
Networking	1× GbE, 1 × 10 GbE
Power	15–60 W

Table 8: Configuration and operational specifications of the UAV utilized for recording surveillance video.

Category	Drone Specification
Maximum Take-off Weight	Up to 16 kg
Holding Capacity	Up to 5 kg
Endurance	45 min
Flying Radius	15 km
Mix Height	1500 m AGL, 4500 m AMSL
Navigation System	GNSS (GPS, RTK supported)
Failsafe Features	Return-to-Home (RTH), Auto-Land
Operating Temperature	–15°C to +55°C
Camera Sensor	1/2.8" Exmor R CMOS
Optical Zoom	30×
Digital Zoom	12×
Effective Pixels	2.13 MP
Gimbal	3-axis stabilization
Recognition Range	Car: 800 m, Human: 500 m
Control Distance	15 km

The LRCN model is very small compared with the AGX Orin. It has 261,676 parameters, so in FP32 the weights require $261{,}676 \times 4 = 1{,}046{,}704$ bytes (≈ 1.05 MB). In FP16 this is ≈ 0.52 MB, and in INT8 it is ≈ 0.26 MB. A 30-frame clip of size 64×64 RGB contains $30 \times 64 \times 64 \times 3 = 368{,}640$ values. In FP32 this becomes $368{,}640 \times 4 = 1{,}474{,}560$ bytes (≈ 1.47 MB), while in uint 8 it is 368,640 bytes (≈ 0.37 MB). The AGX Orin memory bandwidth is 204.8 GB/s. Moving a 1.47 MB clip through memory takes $(1.47 \times 10^6)/(204.8 \times 10^9) \approx 7.2 \times 10^{-6}$ s (≈ 0.007 ms), which is essentially negligible. For a 2-min video (120 s → 3600 frames) with clip_len = 30 and stride = 10, the number of windows is $(3600 - 30)/10 + 1 = 357$. If each inference takes 5 ms, the total time is $357 \times 0.005 \approx 1.79$ s; even at 10 ms it is $357 \times 0.01 \approx 3.57$ s. For a 5-min video (300 s → 9000 frames), the number of windows is $(9000 - 30)/10 + 1 = 897$. At 5 ms per inference the total is $897 \times 0.005 \approx 4.49$ s, and at 10 ms it is $897 \times 0.01 \approx 8.97$ s. The AGX Orin also provides 275 TOPS (INT8). Even if one inference needed 1 Giga Floating-point Operations per Second (GFLOP), the device could theoretically run $275 \times 10^{12}/10^9 = 275{,}000$ such inferences per second. Overall, the small LSTM (32 units) adds little cost, most work is the CNN over 30 frames, and with FP16 TensorRT and stride 10 the system achieves faster-than-real-time performance. Thus, the LRCN fits easily in edge device, for instance AGX orin, and can process multi-minute UAV videos well within real-time limits.

7 Conclusions and Future Work

Public safety and security are significant challenges in an expeditiously changing environment, particularly, suspicious activity detection is a challenging task in smart cities. This paper tackles these challenges by utilizing cutting-edge AI-powered Unmanned Aerial Systems (UAS) for proactive surveillance and activity detection in smart cities. We evaluated three distinct models: the VGG-16 network combined with Long-Short Term Memory (LSTM), a Motion Influence Map (MIM)-based approach, and a Long-Term Recurrent Convolutional Network (LRCN). The LRCN model has achieved up to 88% accuracy in action classification as compared to VGG-16 with LSTM, and MIM models in a crowded smart city. Similarly, it has achieved an MCC of 0.82, reflecting the learning of dynamic activities. The LRCN has achieved an F1 score of 0.88 and an ROC_{AUC} of 0.97, which shows its effectiveness in a dynamic situation. Comparatively, the VGG-16 with LSTM and MIM models have achieved up to 85.33% and 68% accuracy, respectively. The comparatively lower results explain that these models are ineffective in suspicious action detection in a crowded environment. In the future, real-time detection of suspicious behavior is a challenging task. Furthermore, identifying the intent before it occurs, e.g., fighting, is also very interesting to work on. For this purpose, a high computational system is required to instantly save the video stream, detect the suspicious action, and alert the security.

Acknowledgement: The authors would like to thank Prince Sultan University for paying the Article Processing Charges (APC) of this publication. They would also like to thank Prince Sultan University for their support.

Funding Statement: The Article Processing Charges (APC) of this publication were financed by Prince Sultan University.

Author Contributions: Armaghan Azam, Arshad Iqbal—Conceptualization, methodology, experiments, writing—original draft. Armaghan Azam, Arshad Iqbal, M. Mohsin Khan—Data collection/curation, analysis, visualization. Arshad Iqbal, Naveed Ahmad, Mohamad Ladan, M. Mohsin Khan—Supervision, writing—review & editing. Naveed Ahmad, Mohamad Ladan, Arshad Iqbal—funding acquisition. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study can be accessed through the following links: KTH Actions Dataset: https://www.csc.kth.se/cvap/actions/. Movies Fight Detection Dataset: https://www.kaggle.com/datasets/naveenk903/movies-fight-detection-dataset.

Ethics Approval: The dataset employed in this study is publicly available and fully anonymized. No identifiable personal information was directly involved, and the research complies with ethical and data-privacy requirements. For real-world deployment, the system should adhere to relevant privacy regulations, such as obtaining informed consent where applicable, minimizing the collection of unnecessary personal data, and ensuring that all recorded footage is securely stored and used solely for authorized purposes. Additionally, any practical deployment must comply with local drone operation laws, surveillance regulations, and ethical guidelines to protect individual privacy and prevent misuse of the technology.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hashesh AO, Hashima S, Zaki RM, Fouda MM, Hatano K, Eldien AST. AI-enabled UAV communications: challenges and future directions. IEEE Access. 2022;10:92048–66. doi:10.1109/ACCESS.2022.3202956.
2. Wasim M, Ahmed I, Abbas N, Saba T, Elyassih A, Rehman A. Content oriented 3D-CNN sequence learning architecture for academic activities recognition using a realistic CAD dataset. Sci Rep. 2025;15(1):25250. doi:10.1038/s41598-025-07620-3.

3. Cheng N, Wu S, Wang X, Yin Z, Li C, Chen W, et al. AI for UAV-assisted IoT applications: a comprehensive review. IEEE Inter Things J. 2023;10(16):14438–61. doi:10.1109/jiot.2023.3268316.
4. Gohari A, Ahmad AB, Rahim RBA, Supa'at ASM, Abd Razak S, Gismalla MSM. Involvement of surveillance drones in smart cities: a systematic review. IEEE Access. 2022;10:56611–28. doi:10.1109/access.2022.3177904.
5. Srivastava A, Badal T, Garg A, Vidyarthi A, Singh R. Recognizing human violent action using drone surveillance within real-time proximity. J Real Time Image Process. 2021;18(5):1851–63. doi:10.1007/s11554-021-01171-2.
6. Li T, Liu J, Zhang W, Ni Y, Wang W, Li Z. UAV-human: a large benchmark for human behavior understanding with unmanned aerial vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 16266–75.
7. Tahir NUA, Long Z, Zhang Z, Asim M, ELAffendi M. PVswin-YOLOv8s: UAV-based pedestrian and vehicle detection for traffic management in smart cities using improved YOLOv8. Drones. 2024;8(3):84.
8. Husman MA, Albattah W, Abidin ZZ, Mustafah YM, Kadir K, Habib S, et al. Unmanned aerial vehicles for crowd monitoring and analysis. Electronics. 2021;10(23):2974. doi:10.3390/electronics10232974.
9. Papaioannidis C, Mademlis I, Pitas I. Autonomous UAV safety by visual human crowd detection using multi-task deep neural networks. In: 2021 IEEE International Conference on Robotics and Automation (ICRA); 2021 May 30–Jun 5; Xi'an, China. p. 11074–80.
10. Alshehri M, Zahoor L, AlQahtani Y, Alshahrani A, AlHammadi DA, Jalal A, et al. Unmanned aerial vehicle based multi-person detection via deep neural network models. Front Neurorobot. 2025;19:1582995. doi:10.3389/fnbot.2025.1582995.
11. Mittal P. A comprehensive survey of deep learning-based lightweight object detection models for edge devices. Artif Intell Rev. 2024;57(9):242. doi:10.1007/s10462-024-10877-1.
12. Nasir R, Jalil Z, Nasir M, Alsubait T, Ashraf M, Saleem S. An enhanced framework for real-time dense crowd abnormal behavior detection using YOLOv8. Artif Intell Rev. 2025;58(7):202. doi:10.1007/s10462-025-11206-w.
13. Chavan R, kanamarlapudi A, Rani G, Thakkar P, Dhaka VS. CrowdDCNN: deep convolution neural network for real-time crowd counting on IoT edge. Eng Appl Artif Intell. 2023;126:107089. doi:10.1016/j.engappai.2023.107089.
14. Al-Ghanem WK, Qazi EUH, Faheem MH, Quadri SSA. Deep learning based efficient crowd counting system. Comput Mat Conti. 2024;79(3):4001–20. doi:10.32604/cmc.2024.048208.
15. Choi J, Sharma G, Chandraker M, Huang JB. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2020 Mar 1–5; Snowmass Village, CO, USA. p. 1706–15.
16. Cob-Parro AC, Losada-Gutiérrez C, Marrón-Romera M, Gardel-Vicente A, Bravo-Muñoz I. A new framework for deep learning video based human action recognition on the edge. Expert Syst Appl. 2024;238:122220. doi:10.1016/j.eswa.2023.122220.
17. Athavale V, Kumar D, Gupta S. Human action recognition using CNN-SVM model. Adv Sci Technol. 2021;105:282–90. doi:10.4028/www.scientific.net/ast.105.282.
18. Liu C, Ying J, Yang H, Hu X, Liu J. Improved human action recognition approach based on two-stream convolutional neural network model. Vis Comput. 2021;37(6):1327–41. doi:10.1007/s00371-020-01868-8.
19. Leong MC, Prasad DK, Lee YT, Lin F. Semi-CNN architecture for effective spatio-temporal learning in action recognition. Appl Sci. 2020;10(2):557. doi:10.3390/app10020557.
20. Lan Z, Zhu Y, Hauptmann AG, Newsam S. Deep local video feature for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops; 2017 Jul 27–26; Honolulu, HI, USA. p. 1219–25.
21. Yao G, Lei T, Zhong J. A review of Convolutional-Neural-Network-based action recognition. Pattern Recogn Letters. 2019;118(8):14–22. doi:10.1016/j.patrec.2018.05.018.
22. Indhumathi J, Balasubramanian M, Balasaigayathri B. Real-time video based human suspicious activity recognition with transfer learning for deep learning. Int J Image Graph Signal Process. 2023 2;15(1):47–62. doi:10.5815/ijigsp.2023.01.05.
23. Gawande U, Hajari K, Golhar Y. Real-time deep learning approach for pedestrian detection and suspicious activity recognition. Procedia Comput Sci. 2023;218(4):2438–47. doi:10.1016/j.procs.2023.01.219.

24. Nazir A, Mitra R, Sulieman H, Kamalov F. Suspicious behavior detection with temporal feature extraction and time-series classification for shoplifting crime prevention. Sensors. 2023;23(13):5811. doi:10.3390/s23135811.
25. Bousmina A, Selmi M, Ben Rhaiem MA, Farah IR. A hybrid approach based on GAN and CNN-LSTM for aerial activity recognition. Remote Sens. 2023;15(14):3626. doi:10.3390/rs15143626.
26. Saif S, Wollega ED, Kalevela SA. Spatio-temporal features based human action recognition using convolutional long short-term deep neural network. Int J Adv Comput Sci Applicat. 2023;14(5). doi:10.14569/ijacsa.2023.0140501.
27. Wastupranata LM, Kong SG, Wang L. Deep learning for abnormal human behavior detection in surveillance videos—a survey. Electronics. 2024;13(13):2579. doi:10.3390/electronics13132579.
28. Bakirci M. Vehicular mobility monitoring using remote sensing and deep learning on a UAV-based mobile computing platform. Measurement. 2025;244(3):116579. doi:10.1016/j.measurement.2024.116579.
29. Agarwal M, Parashar P, Mathur A, Utkarsh K, Sinha A. Suspicious activity detection in surveillance applications using slow-fast convolutional neural network. In: Verma P, Charan C, Fernando X, Ganesan S, editors. Advances in data computing, communication and security. Singapore: Springer Nature Singapore; 2022. p. 647–58. doi:10.1007/978-981-16-8403-6_59.
30. Ding X. A deeply-recursive convolutional network for crowd counting. IEEE Trans Pattern Anal Mach Intell. 2020;11:317–29.
31. Abdullah F, Jalal A. Semantic segmentation based crowd tracking and anomaly detection via neuro-fuzzy classifier in smart surveillance system. Arab J Sci Eng. 2023;48(2):2173–90. doi:10.1007/s13369-022-07092-x.
32. Elmadany N, He Y, Guan L. Information fusion for human action recognition via biset/multiset globality locality preserving canonical correlation analysis. IEEE Trans Image Process. 2022;31(11):4321–9. doi:10.1109/TIP.2018.2855438.
33. Tang Y, Zheng Y, Wei C, Guo K, Hu H, Liang J. Video representation learning for temporal action detection using global-local attention. Pattern Recognit. 2023;134(11):109135. doi:10.1016/j.patcog.2022.109135.
34. Shao W, Bouazizi M, Tomoaki O. Depth video-based secondary action recognition in vehicles via convolutional neural network and bidirectional long short-term memory with spatial enhanced attention mechanism. Sensors. 2024;24(20):6604. doi:10.3390/s24206604.
35. An S, Bhat G, Gumussoy S, Ogras U. Transfer learning for human activity recognition using representational analysis of neural networks. ACM Trans Comput Healthcare. 2023;4(1):1–21. doi:10.1145/3563948.
36. Bhatia S, Chauhan T, Gupta S, Gambhir S, Panchal JH. An approach to recognize human activities based on ConvLSTM and LRCN. In: 2023 6th International Conference on Information Systems and Computer Networks (ISCON); 2023 Mar 3–4; Mathura, India. p. 1–6.
37. V BV, Indhuja V, Reddy MV, Nikhitha N, Pramila P. Suspicious activity detection using LRCN. In: 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT); 2023 Jan 23–25; Tirunelveli, India. p. 1463–70.
38. Suhas S, Kusuma S, Kiran P, Sindhu Yadav S, Singh S, Sahani V. A deep learning approach for detection and analysis of anomalous activities in videos. In: 2023 International Conference on Network, Multimedia and Information Technology (NMITCON); 2023 Sep 1–2; Bengaluru, India. p. 1–8.
39. Veenu Dr, Vikas V, Katiyar A. Human activity recognition vision based pose detection. In: Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2022; 2022 Feb 19–20; Delhi, India. doi:10.2139/ssrn.4366737.
40. Simpson T. Real-time drone surveillance system for violent crowd behavior unmanned aircraft system (UAS)–human autonomy teaming (HAT). In: 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC); 2021 Oct 3–7; San Antonio, TX, USA. p. 1–9.
41. Perera AG, Law YW, Chahl J. Drone-action: an outdoor recorded drone video dataset for action recognition. Drones. 2019;3(4):82.
42. Qaraqe M, Yang YD, Varghese EB, Basaran E, Elzein A. Crowd behavior detection: leveraging video swin transformer for crowd size and violence level analysis. Appl Intell. 2024;54(21):10709–30. doi:10.1007/s10489-024-05775-6.

43. Liu Z, Yan D, Cai Y, Song Y. Spatio-temporal human action localization in indoor surveillances. Pattern Recognit. 2024;147(2):110087. doi:10.1016/j.patcog.2023.110087.
44. Azmat U, Alotaibi SS, Abdelhaq M, Alsufyani N, Shorfuzzaman M, Jalal A, et al. Aerial insights: deep learning-based human action recognition in drone imagery. IEEE Access. 2023;11:83946–61. doi:10.1109/ACCESS.2023.3302353.
45. Tripathy S, Shanmugam P. Real-time spatial-temporal depth separable CNN for multifunctional crowd analysis in videos. Int J Image Graph. 2025;25(5):2550047. doi:10.1142/s0219467825500470.
46. Hussain A, Ullah W, Khan N, Khan ZA, Kim MJ, Baik SW. TDS-Net: transformer enhanced dual-stream network for video Anomaly Detection. Expert Syst Appl. 2024;256:124846. doi:10.1016/j.eswa.2024.124846.
47. Hussain A, Khan N, Khan ZA, Yar H, Baik SW. Edge-assisted framework for instant anomaly detection and cloud-based anomaly recognition in smart surveillance. Eng Appl Artif Intell. 2025;160(1):111936. doi:10.1016/j.engappai.2025.111936.
48. Movies fight detection dataset. Kaggle Dataset. 2022 [cited 2026 Jan 5]. Available from: https://www.kaggle.com/datasets/naveenk903/movies-fight-detection-dataset/.
49. KTH Center for Autonomous Systems. KTH actions dataset. [cited 2024 Jan 28]. Available from: https://www.csc.kth.se/cvap/actions/.
50. Tammina S. Transfer learning using VGG-16 with deep convolutional neural network for image classification. Int J Scient Res Publicat. 2022;9:143–50. doi:10.29322/ijsrp.9.10.2019.p9420.