Received: July 30, 2023. Revised: September 25, 2023.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Abnormal Activity Recognition with Residual Attention-based ConvLSTM Architecture for Video Surveillance

Anagha Deshpande[1]* **Krishna Warhade[1]** **Pratap Sanap[2]**

[1]*School of Electronics and Communication Engineering, Dr. Vishwanath Karad MIT World Peace University, Pune, Maharashtra, India*
[2]*Head - Research & Innovation @Neilsoft, Pune, Maharashtra, India*
* Corresponding author's Email: anagha.deshpande@mitwpu.edu.in

Abstract: Human activity recognition (HAR) has become a highly researched area with numerous practical applications in public safety. Deep learning has revolutionized HAR by introducing novel approaches to tackle its challenges. Abnormal activity recognition enables prompt intervention and enhances public safety. Presently vision-based activity recognition techniques mainly use recurrent neural network (RNN) architectures like LSTM to handle sequential data dependency. However, this approach struggles to capture the spatial information between consecutive frames in video data, limiting their ability to learn spatiotemporal patterns. To address this issue, we introduce a layer called ResAttenConvLSTM2D, a variant of the ConvLSTM layer, and propose a novel architecture for solving the abnormal activity recognition problem. In the residual attention model, attention is applied to the residual connections, enabling the network to concentrate on portions of the input by calculating the attention score at each time iteration during model training. In addition, the proposed approach addresses the challenge of limited resources in handling video data by employing robust key frame extraction methods using an unsupervised K-Means algorithm. The proposed architecture is tested for benchmark datasets, i.e., AIRT Lab, hockey fight, and abnormal human activity with a classification accuracy of 90%, 96%, and 99% respectively, showing comparable accuracy or complexity compared to the state-of-the-art (SOTA) approaches.

Keywords: Abnormal activities, Intelligent video surveillance, Convolutional long short-term memory, Self-attention, Deep learning.

1. Introduction

In today's era, technology plays a prominent role in shaping and influencing human lives. As we strive for safer living environments, automated smart surveillance emerges as a crucial need of the hour. Given the challenges and complexities of modern society, there is a growing demand for automated smart surveillance systems. Human activity recognition is a key contributing research area in the development of flawless automated smart surveillance systems [1]. In the context of human activity recognition, normal activities are typically defined as common and frequently occurring activities, like walking, running, and jumping. On the other hand, abnormal activities are defined as infrequent or potentially threatening activities, such as fighting, falling, or engaging in violent behaviour. However, the concept of abnormal or anomalous actions varies depending on the subject and context.

Abnormal activity detection is beneficial for identifying unusual behaviour that could pose a threat to individuals or groups. It can be helpful in detecting early warning signs in patients with Alzheimer's or dementia, allowing caretakers to provide proper assistance [2]. Embedding intelligence for abnormal human activity in automated surveillance [3] helps in improving accuracy. This is because abnormal activities often account for a small fraction of the total activities, and may be easily missed by traditional activity recognition systems.

The research on identifying abnormal or anomalous activities in surveillance videos is an arduous and complex task due to various factors like the subjective nature of defining what constitutes abnormal behaviour, limited availability of annotated data for training models, low resolution in surveillance footage, intra/inter-class variations [4].

Traditional approaches for abnormal activity detection rely on handcrafted features and rule-based algorithms [5], which often require domain expertise and extensive tuning to achieve satisfactory results. However, with the recent advances in deep learning [6], there has been a growing interest in developing data-driven models for abnormal activity detection that can automatically learn relevant features and patterns from the data.

In this context, various deep learning methods were proposed, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and autoencoders which resulted in promising outcomes in detecting abnormal events and activities from videos and sensor data.

The motivation behind implementing an AI-powered automated surveillance system to detect human abnormal behavior is to enhance safety in the lives of the elderly and patients, minimize criminal activities and public asset loss resulting from theft, mitigate workplace harassment, and reduce instances of violence.

In our research, we devised a model to recognize abnormal human activity. This model has been constructed using two-dimensional LSTM convolution and a novel layer which is an amalgamation of attention mechanisms and residual connections. In our work, important frames from video sequences were extracted using unsupervised techniques, and these key frames were then used as input to our model. This strategy was employed for making the model less complex and resource-intensive. The proposed technique is compared with two strong baseline methods, the two-dimensional LSTM convolution model, and the LSTM model. We carried out a comparison study using multiple metrics, like accuracy, precision, F1 score, area under the ROC curve, etc., to evaluate the effectiveness of our suggested approach for recognizing abnormal human activities.

Our findings show that the proposed ResAttenConvLSTM2D neural network design offers performance on par with standard baseline topologies, rendering it an acceptable and viable option for abnormal human activity identification.

The following points make up the author’s contribution:

1. A modified unsupervised approach for extracting keyframes from videos to provide a better representation of the video.
2. The development of the ResAttenConvLSTM2D residual attention convolutional recurrent layer, a variant of the standard ConvLSTM layer that combines both the residual and attention concepts with the ConvLSTM layer.
3. Creating an architecture employing the recently developed ResAttenConvLSTM2D layer to solve the issue of abnormal human activity recognition.
4. The approach suggested incorporates autonomous feature extraction using deep learning techniques, which eliminates the necessity for conventional feature extraction engineering.
5. The proposed model's testing on three diverse datasets reveals its promising ability to generalize learned patterns and achieve accurate prediction of abnormal human activities across various situations.

The detailed study is arranged as follows: Section 2 reports a summary of present approaches in the literature for abnormal human activity recognition, section 3 focuses on baseline architecture design and proposed residual-attention architecture and dataset details used in this study; section 4 provides the experimental findings and analysis; section 5, author conclude the paper with a conclusion and future scope.

2. Literature review

Abnormal activity detection is a significant component of human activity recognition, as it can provide early detection of anomalies, improve accuracy, and enhance security in several domains, such as healthcare, surveillance, and security. Many research efforts have been done with the goal of identifying abnormal activity, covering a range of issues, including model architecture, input representation, feature selection, hyperparameter tuning, and dataset characteristics. Researchers explore different approaches and techniques to improve accuracy and effectiveness, leading to a diverse set of tailored solutions for specific contexts and datasets.

This section encompasses a summary of the prominent notable and relevant works aimed at solving the HAR problem. Since this paper focuses on video data, we only consider vision-based HAR techniques. The HAR approaches may be broadly categorized into the below-mentioned strategies, according to published research.

2.1 Traditional methods

Traditional methods typically involve methods for extracting handcrafted features, encoding algorithms for feature representations, and machine learning algorithms for classification tasks. The feature extraction methods broadly cover the local and global features. The local features capture distinctive characteristics within an image, such as corners, edges, or textures. Techniques like SIFT and SURF compute these features and offer robustness against variations in scale, rotation, and illumination [7]. Global features like color histograms, texture descriptors, or statistical moments provide a higher-level representation of the image content [8]. In the feature extraction part, previous methods primarily relied on low-level trajectories, image directs, and consistent patterns. These methods aimed to capture basic visual or spatiotemporal information from the data [9]. Trajectory-based methods suffer from the challenges like occlusion, shadows, and crowded scenarios. To address these issues the researchers employed histogram of optical flow (HOF), histogram of motion direction (HMD), and spatiotemporal gradient techniques for feature extractions [10]. Working with UMN dataset, C. Wu et. al. [11] employed GMM (Gaussian mixture models), optical flow analysis, and fuzzy C-means clustering techniques. Most of the hand-crafted features like VLDA or BOW [12] are specific to the datasets and lack a generalized feature extractor model for human activity recognition. In order to improve the efficacy of human activity systems, researchers investigated conventional machine learning approaches and utilized classification algorithms such as KNN (K-nearest neighbours) [13] and SVM (support vector machines) [14]. However, these algorithms face challenges such as time-consuming processes, labour-intensive requirements, and the intricate nature of feature engineering.

Nowadays researchers have transitioned towards deep learning approaches with the advancement of technology and increased availability of data, to tackle the limitations inherent in traditional methods.

2.2 Deep learning methods

In computer vision, recent research emphasizes deep learning's automatic feature extraction and classification, employing end-to-end architectures for complex tasks like human activity recognition. Gholamrezaii et al. [15] used a 2D CNN with stride-based pooling to enhance human activity recognition by reducing computation time while maintaining or improving model performance. Researchers have suggested using 3D filters instead of 2D filters to extract spatiotemporal features from video frames, enhancing feature learning [16].

Roberta Vrskova introduced a novel dataset for abnormal activity recognition, with less voluminous than UCF crime, and proposed a novel ConvLSTM architecture. Recognition accuracy quoted was 96.19%, leaving room for improvement due to the dataset's novelty [17]. 3D CNNs and Conv-LSTMs exhibit the ability to capture spatio-temporal features within videos and have showcased their effectiveness in accurately detecting instances of violence. The author presented comprehensive end-to-end architectures for both approaches, yielding impressive recognition accuracy. However, it is worth noting that these architectures are characterized by their considerable model complexity. Videos are segmented into chunks of 16 frames each. These chunks are then further split into training and test sets. This division sometimes leads to the inclusion of similar frames in both the training and test sets, which can contribute to achieving higher accuracy in the results [18].

Researchers often employ pre-trained networks such as VGG16, VGG19, and inception V2 to overcome limitations related to limited training data or to expedite the training process. A ResNet-50 model that has been pre-trained is employed to capture features from the video frames. Subsequently, these extracted features are channelled into a ConvLSTM block for further processing. The accuracy results for binary classification for the Hockey fight dataset was 89%, which can be further improved using better convolution LSTM architectures [19]. In recent years generative models have become popular for abnormal activity recognition from video sequences. Sabokroul et al. [20] proposed an innovative technique for anomaly detection in surveillance settings by harnessing the power of generative adversarial networks. The approach of a generative models for human activity recognition may struggle with complex motions, require large datasets, lack transparency, and risk generating unrealistic data.

2.3 Attention-based mechanism

Attention-based mechanisms have emerged as a prominent technique in deep learning, specifically within the domains of sequence modelling and natural language processing. These mechanisms empower models to concentrate on specific components of the input data, imparting varying levels of significance or attention to individual elements. Recent studies have shown that deep earning approaches combined with attention mechanisms work well for various applications,

Received: July 30, 2023. Revised: September 25, 2023.

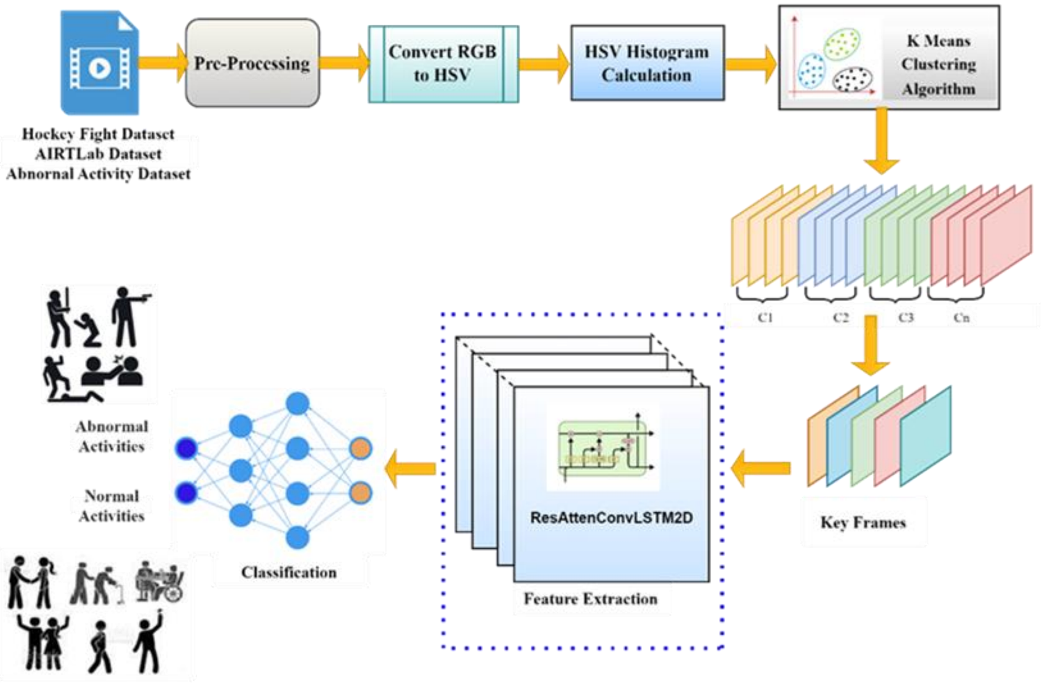


Figure. 1 Flow diagram for the proposed system

lincluding, speech recognition, language translation, image captioning, and video classification.

For identifying human activity, several research emphasize choosing the RGB video's focus. A soft attention method proposed by S. Sharma et.al. [21] learns to concentrate exclusively on video data and can identify the action after a few speedy glances at the RGB video.

The application of attention processes to the detection of human activity from sensor data has produced noteworthy outcomes [22]. The action recognition in videos employs a two-stream attention-based LSTM architecture which effectively tackles the issue of disregarding visual attention. Violence detection was implemented using multi-head self-attention and bidirectional convolution LSTM with better accuracy and inference time [23].

3. Proposed methodology

This section delves into the proposed architecture for human activity recognition, highlighting its essential components: Keyframe extraction, LSTMCONV, attention module, and residual connections.

We recognize the abnormal human activities from the video sequence feed with the help of LSTM convolution and attention mechanism for better feature learning. In the proposed model, firstly, the keyframes from the videos are extracted employing the unsupervised K-means algorithm [24]. Secondly, the four-dimensional tensor is applied as input to the first ConvLSTM2D layer, and then the output is fed to the ResAttenConvLSTM2D layer, a novel layer introduced for better feature extractions. Further fully connected layer and SoftMax probability functions are used for the activity classification. Fig. 1 illustrates the overall flow of the proposed system for abnormal human activity recognition.

3.1 Keyframe extraction

A video feed is composed of a series of static frames that are displayed in rapid succession, resulting in the appearance of continuous motion. Since most frames in a video demonstrate a high degree of correlation with their adjacent frames, in recent times, there has been a growing research emphasis on developing methods that can effectively extract and capture motion information present in the temporal dimension of videos.

The preprocessing step involves extracting video files, processing with OpenCV tools, cropping, adding frames to a list, and changing color channels to standard order for deep learning libraries. HSV instead of RGB in key frame extraction can help to enhance the efficacy of the key frame selection process by concentrating on the most crucial visual aspects of the video. RGB frames are further converted to HSV (hue, saturation, value), and the calculated HSV histogram is shown in Fig. 2, using the number of bins 256 and the range of pixel values

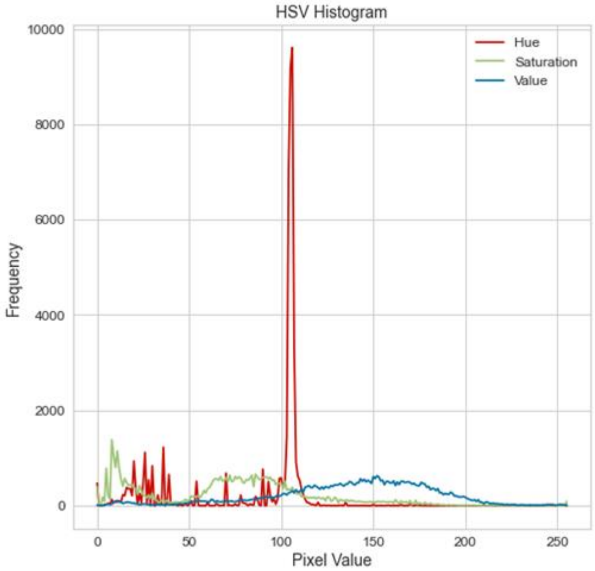
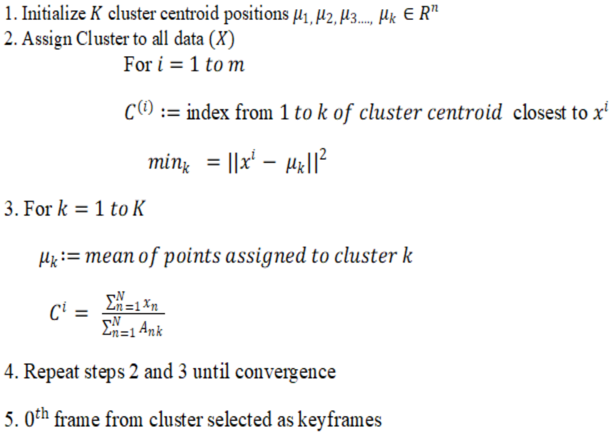
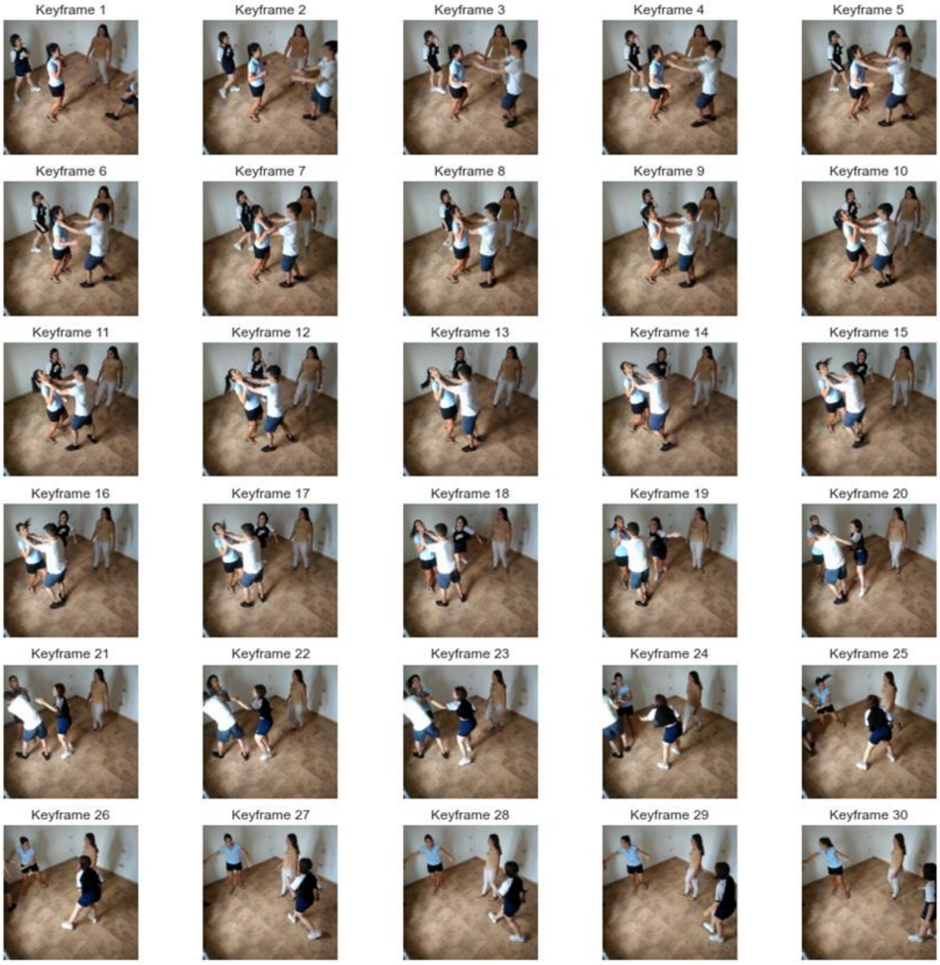


Figure. 2 HSV histogram



```
1. Initialize $K$ cluster centroid positions $\mu_1, \mu_2, \mu_3, ..., \mu_k \in R^n$
2. Assign Cluster to all data $(X)$
        For $i = 1$ to $m$
            $C^{(i)}$ := index from 1 to $k$ of cluster centroid closest to $x^i$
              $min_k = ||x^i - \mu_k||^2$
3. For $k = 1$ to $K$
    $\mu_k$:= mean of points assigned to cluster $k$
    $C^i = \frac{\sum_{n=1}^{N} x_n}{\sum_{n=1}^{N} A_{nk}}$
4. Repeat steps 2 and 3 until convergence
5. $0^{th}$ frame from cluster selected as keyframes
```

Figure. 3 Key frame extraction algorithms



Sorted Keyframes: [0 7 13 15 19 21 26 30 33 37 40 44 49 57 60 61 63 67 73 79 85 88 90 96 103 107 117 120 127 132]

Figure. 4 Sample frames for the violence detection video

(0,256). The HSV histogram values are concatenated and normalized values of the concatenated histogram are calculated which represents the probability of a specific bin. These probability mass function values are fed to the K-means clustering algorithm, to reduce the high-dimensional video data into measurable low-dimensional data and lower the computational complexity while capturing the features of high-dimensional abstract video images.

Fig. 3 shows the basic outline of how the K-means algorithm is used for keyframe extraction, step 1: Set the number of desired keyframes, step 2: Randomly initialize K cluster centroids, step 3: Assign each frame to a cluster based on Euclidian distance matrix, step 4: Update the cluster centroids by recalculating the mean value and then repeat steps 3-4 until convergence. The value of the K is selected empirically based on datasets used in experimentation. The algorithm is going to iterate its steps 300 times before it reaches halting conditions. In our approach, we choose the 0th index keyframe for subsequent feature extraction from the cluster to ensure the inclusion of at least one frame from each cluster. Fig. 4 shows the sample keyframes extracted.

3.2 Convolution LSTM architecture

In this subsection, we justify the significance of Convolution LSTM for human activity recognition applications.

CNNs are widely used for processing image data due to the capability to extract important features from images by applying filters to identify edges, shapes, and textures. RNNs are effective in modelling temporal dependencies in sequential data. The FC-LSTM model consists of a stack of LSTM layers, trailed by one or more dense layers that predict the output. By combining the two, ConvLSTM networks can model both spatial and temporal information simultaneously, making them ideal for videos, and analyzing sequential image data. ConvLSTM proved an extremely effective approach for forecasting air pollution [25].

ConvLSTM layers are a type of recurrent layer that utilizes convolution operations instead of matrix multiplications. This makes the data flow through the ConvLSTM cells and maintains its original dimensions, which is particularly useful for sequential image data [26]. Fig. 5 displays the internal structure of the ConvLSTM cell.

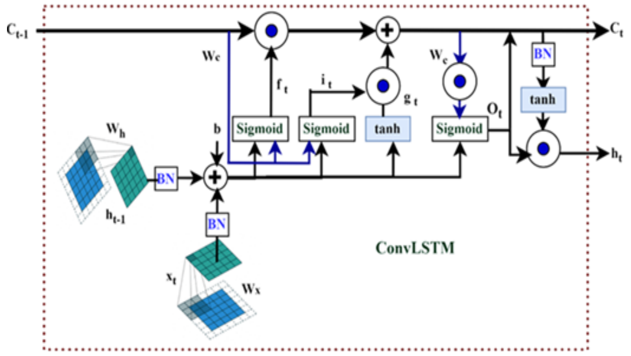


Figure. 5 Internal structure of ConvLSTM cell

$$i_t = \sigma(W_{zi} * z_t + U_{hi} * h_{t-1} + V_{ci} \circ c_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma\left(W_{zf} * z_t + U_{hf} * h_{t-1} + V_{cf} \circ c_{t-1} b_f\right) \quad (2)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ tanh\ (V_{cz} * x_t + U_{hc} * h_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma\ (W_{zo} * x_t + U_{ho} * h_{t-1} + V_{co} \circ c_t + b_o) \quad (4)$$

$$h_t = o_t \circ \tanh(c_t) \quad (5)$$

Eqs. (1) to (5) represent the mathematical operations for the ConvLSTM cell. Here i_t, f_t, o_t are the input, forget, and output gates at time t respectively. The weight metrics are indicated by "W_{zi}", "U_{hi}", "V_{ci}" and "W_{zf}", "U_{hf}", "V_{cf}" and "W_{zo}", "U_{ho}", "V_{co}". "σ" is the nonlinear sigmoid function, and "∘" is Hadamard product, "*" is the convolution.

3.3 Proposed model

The proposed model is conceptualized using the ConvLSTM and fusion of attention and residual. Fig. 6 shows the layered stacking diagram for the proposed novel approach for human activity recognition. The architecture is arranged as follows:

The initial step involves utilizing a conventional Conv2DLSTM layer with 32 filters and a kernel size of 3x3 to capture low-level features in the data. Following this, the output tensor is passed through batch normalization, which helps in normalizing the values and enhancing the stability of the network. Then, two different processes are applied to the normalized tensor.

(a) A second Conv2DLSTM layer with 64 filters and a kernel size of 3x3 is employed, followed by batch normalization and a dropout layer. This combination allows for the extraction and learning of new features in subsequent layers.

(b) Simultaneously, a residual layer is introduced, consisting of a Conv2DLSTM layer with 64 filters and a kernel size of 3x3, followed by a custom attention layer. This residual layer is responsible for capturing important information and highlighting significant patterns. (a) and (b) are combined by element-wise addition.

The output tensors obtained from steps (a) and (b) are combined by element-wise addition. The resulting tensor is then fed into a third Conv2DLSTM

Received: July 30, 2023. Revised: September 25, 2023.

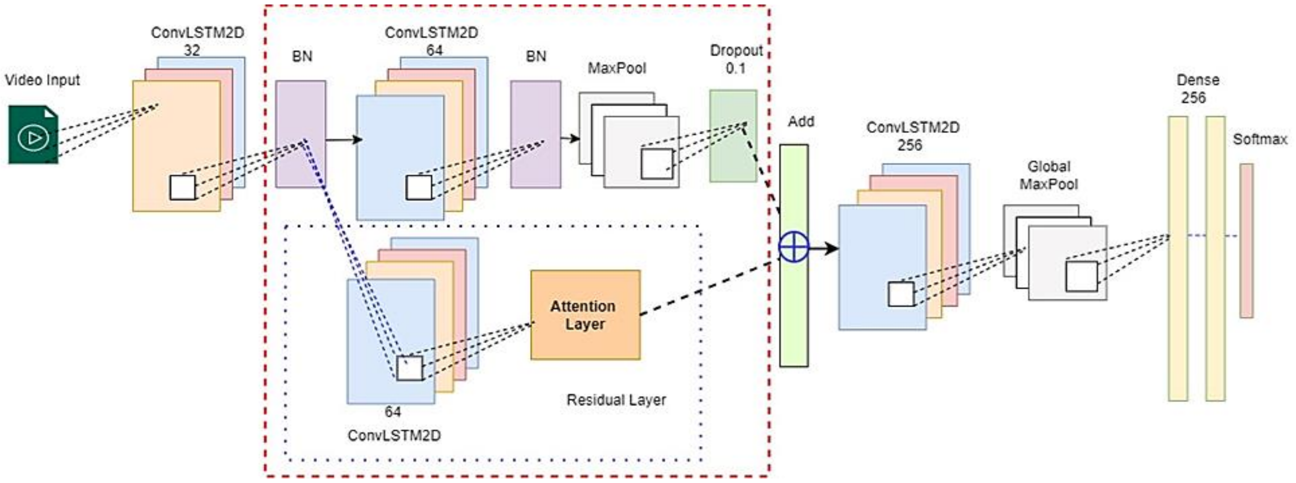


Figure. 6 Proposed architectures for abnormal activity recognition

layer resulting tensor is then fed into a third Conv2DLSTM layer with 256 filters and a kernel size of 3x3, which aims to extract high-level features from the input data. Finally, a global average pooling operation is applied to obtain a one-dimensional tensor. This tensor is subsequently applied as an input to dense layers for the purpose of classification.

3.4 Custom attention layer

Attention mechanisms in deep learning refer to computational techniques that enable models to focus on exact portions of input data selectively. Attention mechanisms in deep learning encompass various types, such as additive attention, dot product attention, and self-attention [27] (or scaled dot product attention). Attention mechanisms in these models work by calculating weighted sums of input data to generate context-aware representations for each input token. Fig. 7 shows the proposed attention layer structure where the input tensor will be residual input x,he self-attention mechanism algorithm with three key phases: Initialization, build, and the call functions. During the Initialization step, the custom attention layer performs non-trainable operations on the input data, such as reshaping or cropping, without involving any trainable weights or biases.

Moving on to the build method, it takes the input shape as an argument and uses this information to create two trainable weights: w and b. These weights will be learned during the training process.

Finally, in the call method, an input tensor x is taken as input. The attention scores for each time step in x are computed by performing a dot product between x and the w weight, and adding the b biases. These attention scores subsequently pass through a hyperbolic tangent activation function (tanh) as in Eq. (6) and a SoftMax activation function Eq. (7). These activations are applied to calculate the attention weights, which are used to weigh the input tensor x. This context vector C_t is then used to compute the attended representation as shown in Eq. (8).

$$e_{t,i} = K.tanh(dot(x.W) + b \qquad (6)$$

$$\propto_{t,i}= \left(\frac{e^{e_{t,i}}}{\sum_{j=1}^{k} e^{e_{t,i}}}\right) \qquad (7)$$

$$c_t \ = \sum_{i=1}^{T} x.\propto_{t,i} \qquad (8)$$

4. Results and discussion

This section discusses the implemented approaches and the findings of the experiments. demonstrating the importance and impact of the proposed study. The model was developed using Keras with the TensorFlow backend and deployed on a Windows 10 platform. The experiments were conducted on a system equipped with an Intel Core i7 8700k CPU running at 3.70GHz, 32GB of RAM, and an Nvidia graphics processing unit (GPU) Geforce GTX1080Ti.

4.1 Datasets

We evaluated the proposed method by conducting experiments on three publicly available datasets that are commonly used for recognizing abnormal activities. These datasets include the AIRT Lab violence dataset [28], hockey fight dataset [29], and Abnormal Activities dataset [17]. The AIRT Lab dataset was created primarily to evaluate the viability of violence detection methods against false positives in fast-moving, nonviolent videos (such as hugs, clapping, mocking, etc.).

The hockey fight dataset is a collection of videos from the Hockey game in the national hockey league. An abnormal activities dataset covers activities such

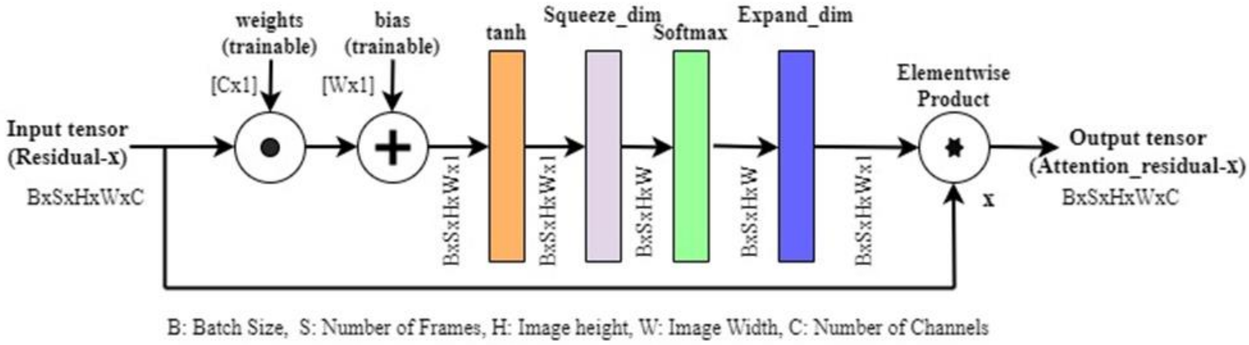


Figure. 7 Attention layer structure

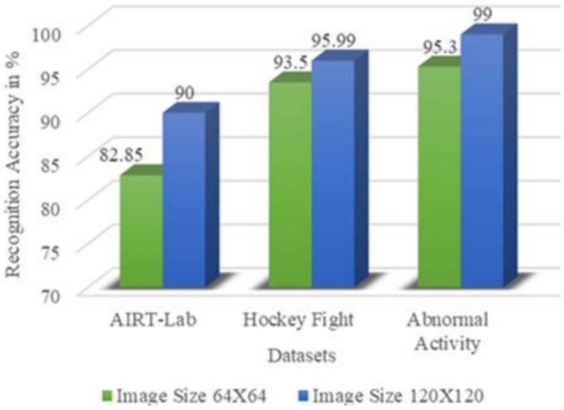


Figure. 8 Recognition accuracy versus input image size

as beggaring, drunkenness, fighting, harassing, kidnapping, the threat of a knife, property damage, robbery, and terrorism. This dataset provides a broader range of aberrant activities for enhanced event recognition, overcoming the issue of short video clips in the UCF-crime dataset. Table 1 shows further details for the mentioned datasets.

4.2 Experimentation

This section discusses the experimentation and results obtained for baseline and proposed models. The performance was examined using a range of evaluation measures such as the confusion matrix, F1 score, recall, precision, class-wise accuracy, the area under the curve (AUC), and receiver operating curve.

The base models and proposed model were trained for 25 epochs with a batch size of 8 using Adam optimizer (learning rate=1e-4, decay=1e-5) and sparse categorical cross-entropy loss function. The train test split used is 80:20. In the experiment, 30 frames are extracted from each video using keyframe selection techniques and are resized to either 64x64 or 120x120 pixels. The larger size (120x120) leads to better recognition accuracy, as shown in Fig. 8.

Therefore, all subsequent evaluations were related to 120x120 pixel images.

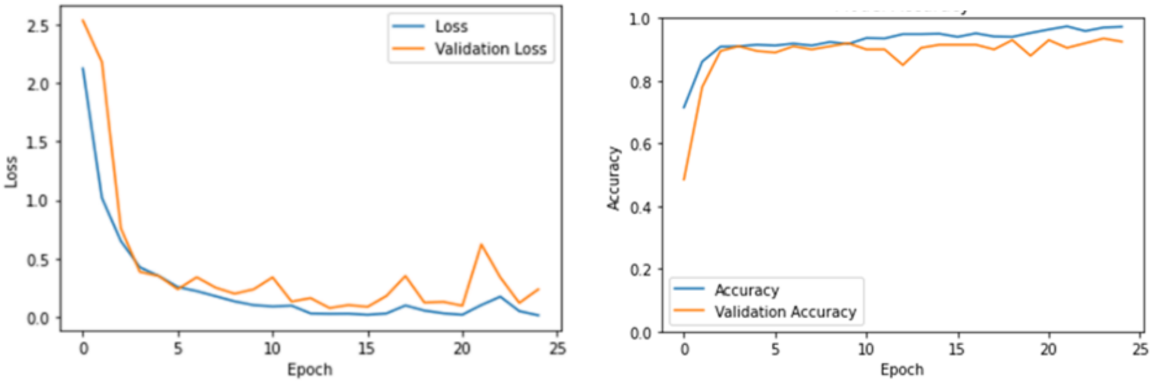
Experimental investigations were done using three versions of the ConvLSTM: ConvLSTM2D, ConvLSTM2D with a residual connection, and our proposed ResAttenConvLSTM2D model.

Table 2 shows the outcomes for the two base models and the proposed model in terms of recall, precision, F1 score, and accuracy. The proposed residual attention-based convolution 2DLSTM model produced impressive outcomes when compared to baseline approaches. Fig. 9 shows the test and validation loss graph for the proposed model for the hockey fight dataset. The confusion matrices of the baseline ConvLSTM2D model and the proposed model, as well as the ROC-AUC curves for all tested databases, are shown in Figs. 10 and 11. The model's accuracy improved when attention was directed toward the residual connections, in contrast to the traditional approach of applying it at the layer before the dense layer. This novel approach enabled the model to compute learnable weights for each input sequence step, leading to enhanced predictions by focusing on specific input elements.

The prime aim of this study was to design a robust and accurate model for the application of abnormal human activity detection. Table 3 shows the comparative analysis of the proposed model with state-of-the-art methods using various standard datasets in terms of model complexity and accuracy. It is noted that some existing methods perform better in regard to accuracy, but they are heavier in terms of computational complexity compared to the proposed method. The proposed method’s accuracy results are comparable with existing methods and perform well on the abnormal activity dataset.

5. Conclusions

Abnormal activity recognition is a significant component of smart surveillance systems as it can provide early detection of anomalies, and enhance security in numerous domains like healthcare, surveillance, and security. In the presented study, we



(a) (b)

Figure. 9 Validation and test loss for hockey fight dataset (a) model loss and (b) model accuracy

Table 1. Dataset details

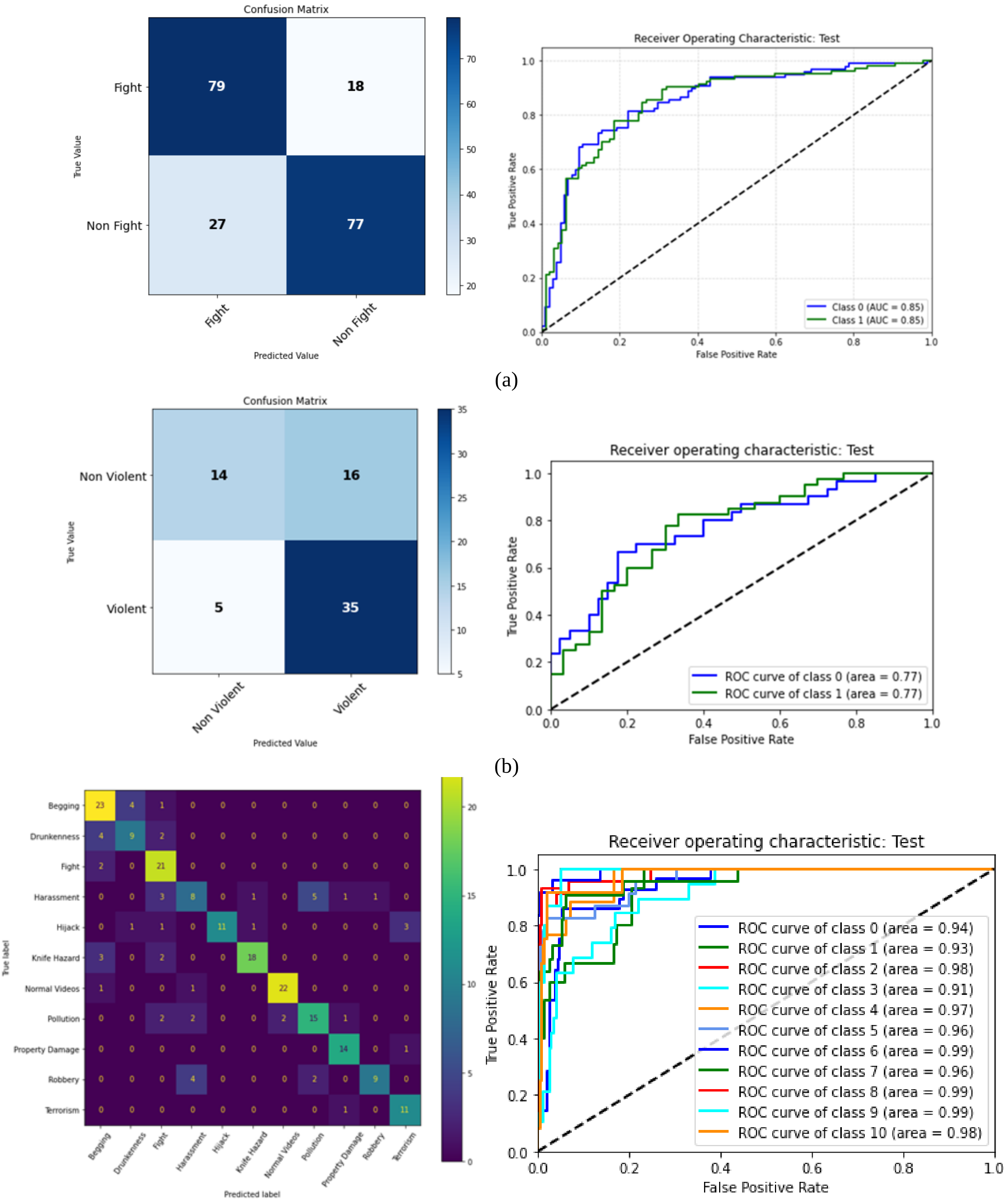
Dataset	Video count	Video length	Other Specification	Challenges
AIRT Lab	350	2-14 Seconds	1920x1080×3, 30 FPS No. of Activity:2	Viewpoint variation, Unbalance Data
Hockey Fight	1000	5 Seconds	360 × 288 ×3, 25 FPS No. of Activity:2	Two persons, Multiple persons, crowd the scene
Abnormal activities	1069	1- 30 Seconds	720×480×3, 30 FPS No. of Activity:11	Differences in camera movement, visual characteristics, and varying lighting conditions.

Table 2. Performance metrices for the base and proposed models

Models	Datasets	Recall	Precision	F1 Score	Accuracy
ConvLSTM2D	Hockey Fight	78	78	78	77.61
ConvLSTM2D+Residual		81	82	81	81.87
Proposed Model		**96**	**96**	**96**	**95.99**
ConvLSTM2D	AIRT Lab Violence Detection	70	71	68	70
ConvLSTM2D+Residual		82	81	82	82.44
Proposed Model		**90**	**90**	**90**	**90**
ConvLSTM2D	Abnormal Activity	76	77	75	75.58
ConvLSTM2D+Residual		84	85	84	84.75
Proposed Model		**99**	**99**	**99**	**99.06**

Table 3. Performance comparison of proposed models with existing methods

Methods	Abnormal Activity % Accuracy	AIRT Lab, %Accuracy	Hockey fight, %Accuracy	Parameters
ConvLSTM [17]	96.19	98.16	-	5.12M
3D Resnet152[17]	90.47	91.42	-	Not Specified
ConvLSTM [18]	-	97.15	97.86	198.4M
ResNet50+ ConvLSTM [19]	-	-	89	Not Specified
Conv2D LSTM [30]		-	94.5	Not Specified
ViolenceNet Optical Flow[31]	-	-	99.2	4.5M
MobileNet V2 +LSTM [32]	-	-	96.1	4.074 M
Proposed Model ResAttenConvLSTM2D	**99**	**90**	**95.99**	**3.833M**



(a)

(b)

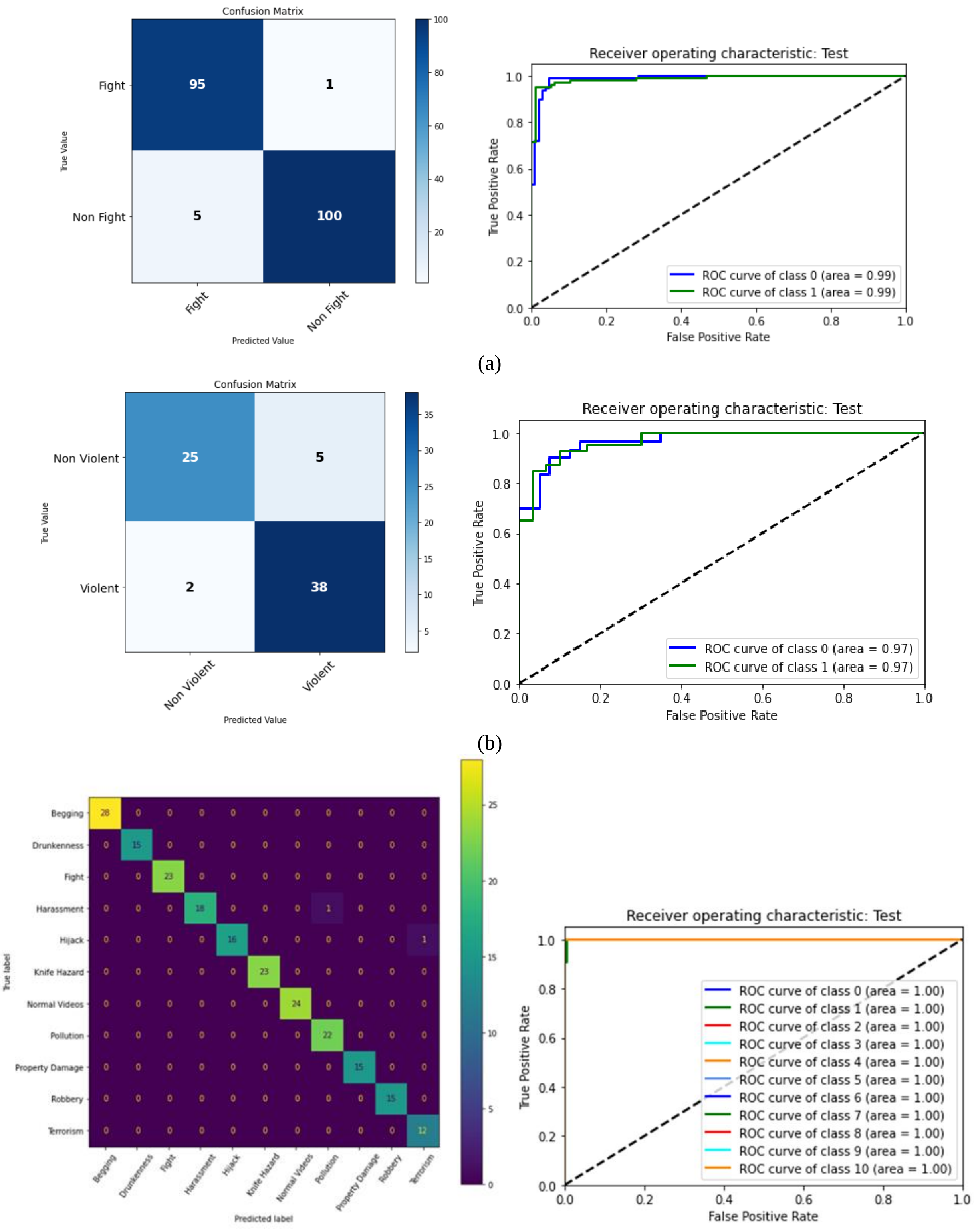
(c)

Figure. 10 ConvLSTM2D base model confusion matrix and ROC curves for: (a) hockey fight, (b) AIRT lab, and (c) abnormal activity

introduced a novel and efficient residual and attention-based ConvLSTM2D model to recognize abnormal human activities from the smart surveillance system.

Our proposed model extracts the keyframes from the video feed and then uses convolution LSTM for extracting spatial and temporal features from a frame sequence; then, it uses a residual attention layer to improve the recognition accuracy of abnormal activities in a surveillance system.

We validated the proposed framework using multiple measurement metrics using diverse datasets.



(a)

(b)

(c)

Figure. 11 Proposed model confusion matrix and ROC curves for: (a) hockey fight, (b) AIRT lab, and (c) abnormal activity

The model is trained and tested on three diverse datasets with multiclass classification ability to handle real-world data. The average recognition accuracy obtained for the proposed model is 95% which is sufficient to identify the nonstandard behaviour of people in public places to enhance security.

In the future, investigation and research can be conducted for selecting meaningful keyframes from clusters. we can explore other deep learning models for video data like 3DCNN or generative models for better human activity recognition.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, Methodology A.Y.D.; Software, A.Y.D.; Validation, A.Y.D.; Writing—DraftResources, Review and Editing, Supervision, K.K.W and P. S.

Acknowledgment

The authors are thankful to Dr. Vishwanath Karad MIT World Peace University, Pune, Maharashtra, India for the support in laboratory and hardware availability.

References

[1] A. Jain, S. Basantwani, O. Kazi, and Y. Bang, "Smart surveillance monitoring system", In: *Proc. International Conference on Data Management, Analytics and Innovation (ICDMAI)*, Pune, India, 2017, pp. 269-273, 2017.

[2] G. Cicirelli, R. Marani, A. Petitti, A. Milella, and T. D. Orazio, “Ambient assisted living: a review of technologies, methodologies and future perspectives for healthy aging of population”, *Sensors*, Vol. 21, No. 10, May 2021.

[3] M. Kashef, A. Visvizi, and O. Troisi, “Smart city as a smart service system: Human-computer interaction and smart city surveillance systems”, *Computers in Human Behaviour*, Vol. 124, 2021.

[4] G. Sreenu and M. A. S. Durai, “Intelligent video surveillance: a review through deep learning techniques for crowd analysis”, *Journal of Big Data*, Vol 6, No. 48, 2019.

[5] Dhiman and D. K. Vishwakarma, “A review of state-of-the-art techniques for abnormal human activity recognition”, *Engineering Applications of Artificial Intelligence*, Vol. 77, pp. 21-45, 2019.

[6] S. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, “Cover the violence: A novel Deep-Learning-Based approach towards violence-detection in movies”, *Applied Sciences*, Vol. 9, No. 4963, 2019.

[7] Y. Hsueh, W. Lie, and G. Y. Guo, “Human behaviour recognition from Multiview videos”, *Information Sciences*, Vol. 517, No. 6, 2020.

[8] M. Elhoseny, et al., “A hybrid model of internet of things and cloud computing to manage big data in health services applications”, *Future Generation Computing System*, Vol. 86, pp. 1383-1394, 2018.

[9] F. Tung, J. S. Zelek, and D. D. A. Clausi, “Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance”, *Image and Vision Computing*, Vol. 29, pp. 230-240, 2011.

[10] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection”, In: *Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, San Diego, CA, USA, pp. 886-893, 2005.

[11] C. Wu and C. Cheng, “A novel detection framework for detecting abnormal human behavior”, *Mathematical Problems*, Vol. 2020, pp. 1-9, 2020.

[12] R. Bhuiyan, S. Tarek, and H. Tian, “Enhanced bag-of-words representation for human activity recognition using mobile sensor data”, *Signal Image and Video Processing*, Vol. 15, pp. 1739–1746, 2021.

[13] S. Mohsen, A. Elkaseer, and S. Scholz, “Human Activity Recognition Using K-Nearest Neighbor Machine Learning Algorithm”, In: *Proc of Sustainable Design and Manufacturing. KES-SDM 2022. Smart Innovation, Systems and Technologies*, Vol. 262, Springer, Singapore, 2022.

[14] Y. Zheng, “Human Activity Recognition Based on the Hierarchical Feature Selection and Classification Framework”, *Journal of Electrical and Computer Engineering*, Vol. 2015, Article ID 140820, 9 pages, 2015.

[15] M. Gholamrezaii and S. M. T. Almodarresi, "Human Activity Recognition Using 2D Convolutional Neural Networks", In: *Proc. of 27th Iranian Conference on Electrical Engineering (ICEE)*, Yazd, Iran, pp. 1682-1686, 2019.

[16] T. Lee, J. Yoon, and I. Lee, “Motion sickness prediction in stereoscopic videos using 3D convolutional neural networks”, *IEEE Transaction on Visual Computing Graphics*, Vol. 25, No. 5, pp. 1919–1927, 2019.

[17] R. Vrskova, R. Hudec, P. Kamencay, and P. Sykora, “A New Approach for Abnormal Human Activities Recognition Based on ConvLSTM Architecture”, *Sensors*, Vol. 22, 2022.

[18] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, and A. Dragoni, “Deep Learning for Automatic Violence Detection: Tests on the AIRTLab data-set”, *IEEE Access*, Vol. 9, pp. 160580–160595, 2021.

[19] M. Sharma and R. Baghel, "Video Surveillance for Violence Detection Using Deep Learning", *Lecture Notes on Data Engineering and Communications Technologies*, Vol. 37, 2020.

[20] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection", In: *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 3379–3388, 2018.

[21] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention", In: *Proc. of 33rd International Conference Mach. Learn.*, New York, USA, pp. 2101-2112, 2016.

[22] X. Yin, Z. Liu, D. Liu, and X. Ren, "A Novel CNN-based Bi-LSTM parallel model with attention mechanism for human activity recognition with noisy data", *Scientific Reports*, Vol. 12, 2022.

[23] C. Dai, X. Liu, and J. Lai, "Human action recognition using two-stream attention-based LSTM networks", *Applied Soft Computing*, Vol. 86, 2020.

[24] Y. Li and H. Wu, "A Clustering Method Based on K-Means Algorithm", *Physics Procedia*, Vol. 25, pp. 1104-1109, 2012.

[25] W. Liu, Y. Wang, D. Zhong, S. Xie, and J. Xu, "ConvLSTM Network-Based Rainfall Nowcasting Method with Combined Reflectance and Radar-Retrieved Wind Field as Inputs", *Atmosphere*, Vol. 13, No. 3, 2022.

[26] M. Haque, R. Hafiz, A. Azad, Y. Adnan, S. Mishu, A. Khatun, and M. Uddin, "Crime detection and criminal recognition to intervene in interpersonal violence using a deep convolutional neural network with transfer learning", *International Journal of Ambient Computing and Intelligence*, Vol. 12, No. 4, pp. 154-167, 2021.

[27] Y. Tan, S. Poh, C. Ooi, and W. Tan, "Human activity 35recognition with self-attention", *Journal of Electrical and Computer Engineering (IJECE)*, Vol. 13, No. 2, pp. 2023-2029, 2023.

[28] M. Bianculli, N. Falcionelli, P. Sernani, S. Tomassini, P. Contardo, M. Lombardi, and A. F. Dragoni, "A dataset for automatic violence detection in videos", *Data in Brief*, Vol. 33, 2020.

[29] E. Bermejo, O. Deniz, G. Bueno, and R. Sukthankar, "Violence detection in video using computer vision techniques", In: *Proc. of International Conference on Computer Analysis of Images and Patterns*, Germany, pp. 332-339, 2011.

[30] M. M. Moaaz and E. H. Mohamed, "Violence Detection In Surveillance Videos Using Deep Learning", *Informatics Bull. Fac. Comput. Artif. Intell. Helwan Univ.*, Vol. 2, No. 2, pp. 1-6, 2020.

[31] F. R. Segador, J. Á. García, F. Enríquez, and O. Deniz, "Violencenet: Dense multi-head self-attention with bidirectional convolutional lstm for detecting violence", *Electronics*, Vol. 10, No. 1601, 2021.

[32] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient Violence Detection in Surveillance", *Sensors*, Vol. 22, 2022.