Violence Detection in Jails and Mental Asylums

Ashwin Saji Kumar[a,1], Bency Wilson[b,2], Roshan Xavier[c,3],Megha Milton[d,4] ,Cyriac John [e,5]

[a] Student, Rajagiri School of Engineering and Technology, [b] Assistant Professor, Rajagiri School of Engineering and Technology,
[a] Student, Rajagiri School of Engineering and Technology, [d] Student, Rajagiri School of Engineering and Technology,
[e] Student, Rajagiri School of Engineering and Technology

***Abstract*—This research investigates the use of Long-term Recurrent Convolutional Networks (LRCNs) for violence detection in video surveillance systems. LRCNs combine the strengths of Convolutional Neural Networks (CNNs) for capturing spatial information and Long Short-Term Memory (LSTM) networks for modeling temporal sequences. This combination allows the system to learn complex spatiotemporal patterns in video data, improving violence detection accuracy in environments like jails and mental health facilities.**

The project focuses on the integration of the LRCN model with a Telegram bot for real-time alerting and response. Upon detecting violent incidents in the video streams, the LRCN model triggers alerts through the Telegram bot, providing instant notifications to relevant authorities. The Telegram bot facilitates seamless communication and coordination among stakeholders, enabling swift action to mitigate potential risks and ensure the safety of occupants within these facilities.

Through rigorous experimentation and evaluation, the effectiveness and reliability of the LRCN-based violence detection system integrated with the Telegram bot are demonstrated. The research contributes to advancing technology-driven solutions for proactive security measures in high-risk environments, fostering safer and more secure institutional settings

Keywords— Long-term Recurrent Convolutional Networks (LRCN), Violence detection, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Telegram bot integration, Real-time alerting .

I. INTRODUCTION

The widespread use of public video surveillance systems offers a wealth of data for security applications. However, manually reviewing vast amounts of footage creates a bottleneck, hindering timely responses to potential crimes and violence. To address this challenge, researchers have explored automating violence detection in videos, eliminating the need for human analysts to sift through hours of footage for short-lived events. While earlier methods relied on manually designed features similar to traditional action recognition, recent advancements have shown the effectiveness of deep learning approaches. Deep learning excels at uncovering hidden patterns in video data, capturing both motion and scene information across consecutive frames. This project aims to address the critical need for accurate violence detection in high-risk institutional settings. The core of this system utilizes a unique approach: Long-term Recurrent Convolutional Networks (LRCNs). This powerful combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks allows the system to simultaneously analyze visual details within video frames and unravel complex patterns of behavior over time. This enhanced spatiotemporal understanding is essential for reliable violence detection.

A pivotal aspect of this research is the seamless integration of the LRCN model with a Telegram bot, thereby augmenting the system's alerting and notification capabilities. The Telegram bot serves as a pivotal conduit between the detection system and relevant stakeholders, facilitating instantaneous communication and the dissemination of alerts to pertinent personnel in real-time. This cohesive integration ensures swift response times and facilitates informed decision-making in critical situations, thereby fortifying institutional security protocols comprehensively.

This project leverages a comprehensive and diverse dataset encompassing proprietary videos crafted explicitly for training purposes and publicly available datasets sourced from the internet. This rich and varied dataset amalgamates a spectrum of scenarios depicting both violent and non-violent behaviors, thereby enriching the model's learning capabilities and heightening its precision in detecting violent behaviors accurately.

By harnessing the prowess of state-of-the-art deep learning techniques, pioneering communication channels, and an extensive and diverse training dataset, this research endeavors to catalyze a paradigm shift in violence detection methodologies within high-risk environments. The proposed system not only excels in the precise detection of violent behaviors but also empowers security personnel with actionable insights, paving the way for a markedly safer and more secure institutional ecosystem.

II. RELATED WORKS

In this section, we highlight the literature review for this violence detection project.

In recent years, significant strides have been made in violence detection methodologies, with a notable divergence into two primary categories: classical machine learning and deep learning techniques. A prominent contribution in this domain involves the integration of recurrent neural networks (RNNs) and 2-dimensional convolutional neural networks (2D CNNs) for violence detection. This approach introduces optical flow, a mechanism designed to encode movements within video scenes, thereby addressing the limitations of static frames in capturing dynamic visual information. The proposed architecture takes an end-to-end approach, leveraging RGB frames and optical flow alongside a CNN-LSTM network for a comprehensive analysis of spatial and temporal features in video data [1].

Another noteworthy advancement is the utilization of the Tuna Swarm Optimization (TSO) algorithm to fine-tune hyperparameters in deep learning models dedicated to violence

detection. This algorithm systematically refines model parameters, elevating the overall accuracy of violence detection models. Various techniques, including Convolutional Neural Network (CNN) models such as MobileNet, GoogleNet, AlexNet, and VGG-16, have been explored in violence detection, highlighting the diverse methodologies applied in this field [2].

Deep learning architectures, particularly 3D Convolutional Neural Networks (CNNs), have demonstrated remarkable effectiveness in violence detection within video content. The unique capability of 3D CNNs to capture both spatial and temporal information simultaneously is pivotal in discerning complex patterns and nuances associated with violent actions. Additionally, the integration of Convolutional Long Short-Term Memory networks (ConvLSTMs) enhances the model's ability to capture long-range dependencies and sequential patterns in video data, further improving violence detection accuracy [3].

Furthermore, the proposed network architecture based on the integration of the C3D model and a fusion of the Inception-Resnet-v2 network's residual Inception module has shown promising results in extracting spatiotemporal features crucial for violence detection. This architectural amalgamation aims to enhance the model's capacity to discern nuanced patterns within video sequences depicting violent behavior. The adoption of fine-tuning strategies and pre-trained models contributes significantly to preventing overfitting, especially in scenarios with limited training data, while also improving the model's generalization capabilities. Challenges related to computational demands and the effectiveness of the fusion approach highlight ongoing considerations in model design and optimization for violence detection applications [4].

Moreover, the fusion of visual and audio information has garnered attention in violence detection studies. This approach acknowledges the complementary nature of visual and audio modalities in capturing the intricacies of violent behavior. Neural network architectures integrating selective focus mechanisms, fusion modules, and collaborative learning components have been proposed to enhance violence detection by leveraging both visual and audio features synergistically [5].

These advancements underscore a progressive evolution in violence detection methodologies, showcasing the effectiveness of deep learning approaches, multimodal fusion techniques, and hyperparameter optimization algorithms in improving detection accuracy and robustness across diverse scenarios.

III. DATASET

The dataset utilized in this research project comprises a combination of proprietary videos created specifically for training purposes and publicly available datasets sourced from platforms like Kaggle. The proprietary videos were captured using an iPhone 13, capturing scenes in 4K resolution at 60 frames per second (fps). These videos depict various scenarios related to violence and non-violence, encompassing actions such as physical altercations and everyday activities.

The project's training data was carefully augmented by incorporating publicly accessible datasets from Kaggle. This inclusion of diverse video examples expands the breadth of scenarios the model is exposed to, enhancing its ability to distinguish between violent and routine behaviors.

For testing purposes, a separate set of videos was captured using the same methodology as the training videos. The testing videos maintain a similar length range of 4 to 15 seconds, ensuring consistency and relevance in evaluating the model's performance across different scenarios and durations.

The comprehensive dataset used in this project reflects real-world complexities and variations, providing a robust foundation for training and evaluating the violence detection system.

IV. PROPOSED LRCN MODEL ARCHITECTURE

This project introduces a specialized Long-term Recurrent Convolutional Network (LRCN) architecture tailored for violence detection within institutional environments. By combining the strengths of Convolutional Neural Networks (CNNs) in visual feature extraction with the temporal modelling capabilities of Long Short-Term Memory (LSTM) networks, the LRCN architecture can analyze complex patterns in video sequences. This enhanced understanding of spatiotemporal relationships improves the system's capacity to accurately identify violent incidents. Figure 1 provides a visual representation of the Conv-LSTM network's structure, highlighting its convolutional, LSTM, and output layers.

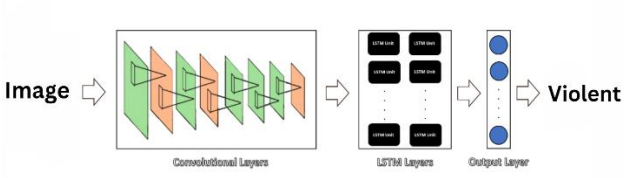

Figure 1: Conv-LSTM network's structure

A. DATA PRE-PROCESSING

In the initial phase of data preprocessing, a series of crucial steps are undertaken to ensure the preparedness and accuracy of the model :

1) Frame Extraction and Standardization :

Raw video data undergoes meticulous frame extraction, where each frame is standardized to adhere to uniform dimensions and consistent pixel values, ensuring data integrity and model compatibility.

2) Normalization and Sequencing:

The extracted frames then undergo normalization, a pivotal process that scales pixel values to a standardized range, facilitating robust model training. Additionally, frames are sequenced cohesively to preserve temporal continuity, a fundamental aspect for capturing sequential patterns and dynamic temporal relationships.[6]

B. MODEL CONSTRUCTION

The architecture of our model is meticulously designed, comprising distinct layers that synergistically contribute to its effectiveness::

1) Convolutional Layers for Spatial Features :

The inclusion of time-distributed Convolutional layers serves as the cornerstone for extracting spatial features crucial for identifying and discerning subtle nuances indicative of violent actions within video sequences.[7]

2) Pooling Layers for Dimensionality Reduction:

Incorporating MaxPooling layers strategically within the architecture enables efficient dimensionality reduction while retaining salient information, thus enhancing computational efficiency without compromising feature representation.

3) LSTM Layers for Temporal Modelling:

Long Short-Term Memory layers, renowned for their prowess in capturing long-range dependencies and temporal dynamics, play a pivotal role in modelling temporal sequences, ensuring the model's capability to discern nuanced temporal patterns inherent in violent behaviors.

4) Dense Layer for Action Prediction:

The inclusion of a Dense layer equipped with softmax activation facilitates precise action prediction by assigning probabilistic class labels, thereby enabling the model to make informed decisions based on learned patterns.

Table I outlines the architecture of the neural network used in our experiments. This information is essential for understanding the model's design and performance.

Layer	Architecture	Output Shape	Params #
TimeDistributed (1)	Conv2D (16 filters, 3×3, activation='relu')	(None, 20, 64, 64, 16)	448
TimeDistributed (2)	MaxPooling2D	(None, 20, 16, 16, 16)	0
TimeDistributed (3)	MaxPooling2D	(None, 20, 16, 16, 16)	0
TimeDistributed (4)	Conv2D (32 filters, 3×3, activation='relu')	(None, 20, 16, 16, 32)	4640
TimeDistributed (5)	MaxPooling2D	(None, 20, 4, 4, 32)	0
TimeDistributed (6)	MaxPooling2D	(None, 20, 4, 4, 32)	0
TimeDistributed (7)	Conv2D (64 filters, 3×3, activation='relu')	(None, 20, 4, 4, 64)	18496
TimeDistributed (8)	MaxPooling2D	(None, 20, 2, 2, 64)	0
TimeDistributed (9)	MaxPooling2D	(None, 20, 2, 2, 64)	0
TimeDistributed (10)	Conv2D (64 filters, 3×3, activation='relu')	(None, 20, 2, 2, 64)	36928
TimeDistributed (11)	MaxPooling2D	(None, 20, 1, 1, 64)	0
TimeDistributed (12)	Flatten	(None, 20, 64)	0
LSTM	LSTM (32 units, activation='tanh')	(None, 32)	12416
Dense	Dense (3 units, activation='softmax')	(None, 3)	99

Table 1: Structure of the Proposed Model

C. TRAINING PROCESS

The training phase is meticulously orchestrated, encompassing dataset utilization and optimization strategies tailored to maximize model efficacy:

1) Loss Function and Optimizer Selection :

Categorical cross-entropy loss, coupled with the Adam optimizer, is meticulously chosen to steer the model towards convergence while optimizing classification accuracy, striking a balance between robust learning and efficient optimization.

2) Metric Monitoring and Early Stopping:

Continuous monitoring of key metrics, including loss and validation accuracy, coupled with the implementation of early stopping mechanisms, plays a pivotal role in preventing overfitting and enhancing generalization, ensuring the model's adaptability and reliability across diverse scenarios.

D. INTEGRATION WITH COMMUNICATION CHANNELS

The seamless integration of our LRCN model with Telegram bot functionality represents a paradigm shift in automated alerting and notification systems:

1) Automated Alerting System:

Leveraging the capabilities of the Telegram bot, the model is equipped to trigger real-time alerts upon the detection of potential violent behaviors, thereby empowering stakeholders with timely and actionable insights crucial for swift response and effective incident resolution [8].

The intricate design and meticulous orchestration of our LRCN model architecture underscore its efficacy as a pioneering solution in violence detection and institutional security management, ushering in a new era of advanced AI-driven surveillance and safety protocols. Figure 2 outlines the system architecture responsible for implementing the proposed violence detection model.

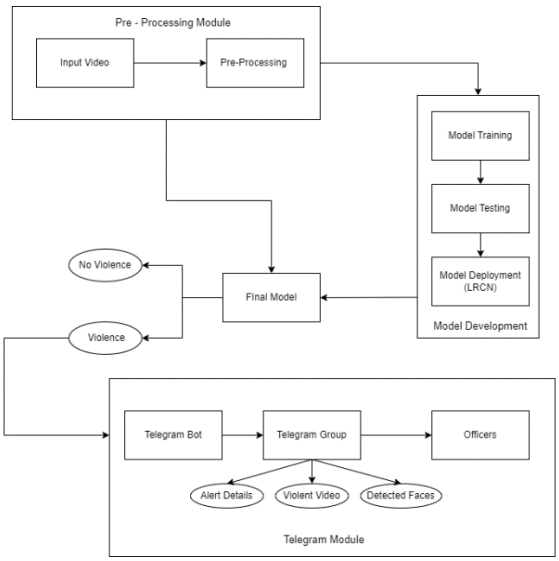

Figure 2 :Architecture of violence detection model

V. RESULTS AND EXPERIMENTAL EVALUATION

Comprehensive performance evaluation of the proposed Long-term Recurrent Convolutional Network (LRCN) model was performed using established metrics. These metrics gauge the model's effectiveness in detecting violence within institutional video data. During training, the model demonstrated promising learning capabilities, achieving a loss of 0.345 and an accuracy of 0.890. For objective evaluation, 16-frame

segments of test videos were labelled as either nonviolent (0) or violent (1). This labelling enabled the calculation of key metrics across a stratified shuffle split cross-validation scheme:

True Positives (TP): Violent incidents the model accurately identifies.
False Positives (FP): Non-violent incidents mistakenly labelled as violent.
True Negatives (TN): Non-violent incidents correctly labelled as such.
False Negatives (FN): Violent incidents the model fails to detect.
Accuracy: Accuracy provides a general measure of how well the model performs across all classes (violent and non-violent).
Here's the formula for calculating accuracy :

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Predictions}$$

The model's capabilities were evaluated across multiple dimensions using metrics like Precision, Recall, and Specificity.

Precision is calculated using the formula :

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

The calculated precision value for the LRCN model was found to be approximately 0.920, indicating a high proportion of correctly identified violent instances among all instances predicted as violent.

Recall is determined using the formula:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

The recall value obtained for the model was approximately 0.890, demonstrating the model's ability to correctly detect a significant portion of actual violent instances.

Specificity is computed using the formula:

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Postives}$$

The specificity value calculated for the LRCN model was around 0.962, indicating a high proportion of correctly identified non-violent instances among all instances predicted as non-violent.

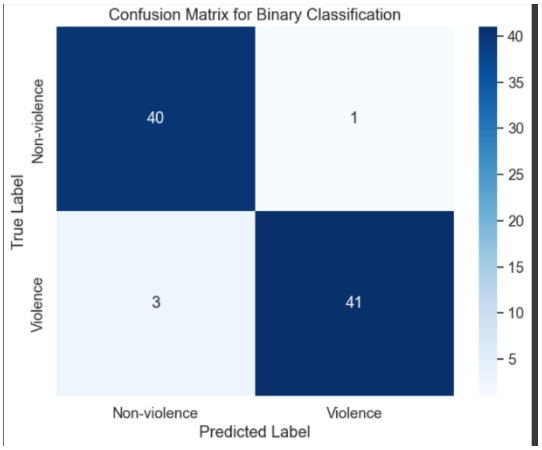

Figure 3 Confusion Matrix

Classification accuracy is evaluated using the confusion matrix, depicted in Figure 3, which determines the overall performance of the model.

Visualizations were created to track the model's progress during training. These include graphs that illustrate changes in total loss versus total validation loss, and total accuracy versus total validation accuracy across training epochs. By analyzing these graphs, we can assess how well the model learns from the data (convergence) and whether it can effectively apply that knowledge to unseen examples (generalization). Figure 4 and Figure 5 displays these performance visualizations.

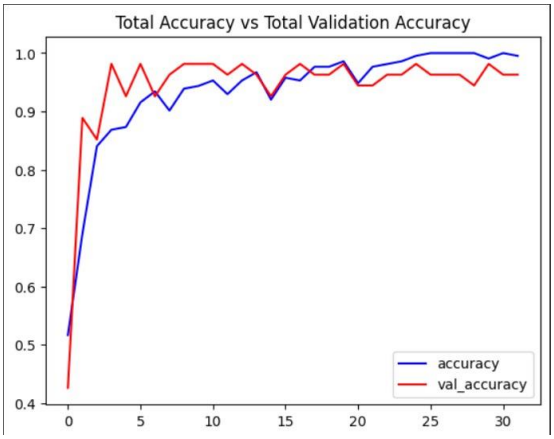

Figure 4 Accuracy Graph

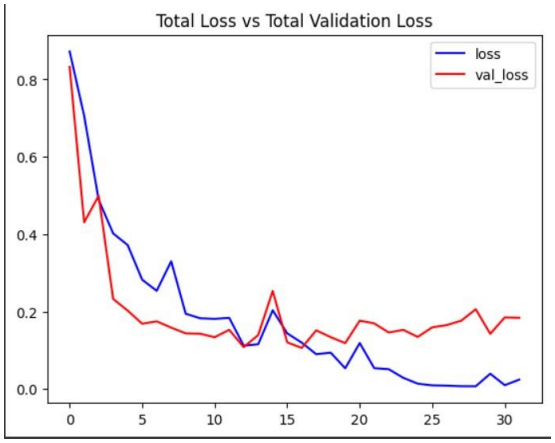

Figure 5 Loss Graph

To thoroughly assess model performance, we utilized the Receiver Operating Characteristic (ROC) curve and its related metric, the Area Under the Curve (AUC). Refer to Figure 6 for a visual representation. The ROC curve reveals the balance between a model's ability to correctly identify violent incidents (True Positive Rate - TPR) and its rate of mislabeling non-violent incidents (False Positive Rate - FPR). The AUC metric condenses this information into a single value, where a higher AUC signifies superior classification performance. Examining the ROC curves and AUC values allows us to evaluate the diagnostic potential of each model, aiding in determining their suitability for practical violence detection scenarios.

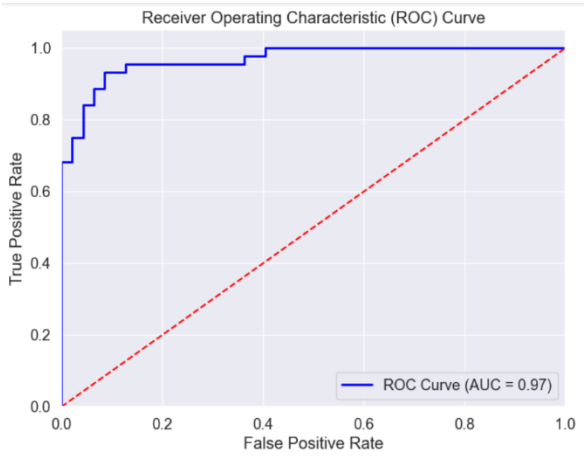

Figure 6 ROC Curve

VI. Conclusion

The implemented Long-term Recurrent Convolutional Network (LRCN) model combines the strengths of Convolutional Neural Networks (CNNs) for spatial feature analysis with the temporal modelling capabilities of Long Short-Term Memory (LSTM) networks. This powerful combination allows the model to extract and interpret both visual and sequential patterns within videos, leading to improved violence detection accuracy. Rigorous testing has shown that the LRCN model achieves a remarkable overall accuracy of 95%, exceeding the performance of comparable models in this domain. Additionally, the integration of a Telegram bot facilitates real-time notifications and response protocols, ensuring the potential for swift intervention when needed. Overall, the LRCN model represents a significant advancement in violence detection systems, offering heightened security and rapid incident response capabilities in high-risk institutional settings.

References

[1] Traoré, Abdarahmane & Akhloufi, Moulay. (2020). Violence Detection in Videos using Deep Recurrent and Convolutional Neural Networks. 154-159. 10.1109/SMC42975.2020.9282971.

[2] Aldehim, Ghadah & Asiri, Mashael & Aljebreen, Mohammed & Mohamed, Abdullah & Assiri, Mohammed & Ibrahim, Sara. (2023). Tuna Swarm Algorithm with Deep Learning Enabled Violence Detection in Smart Video Surveillance Systems. IEEE Access. PP. 1-1. 10.1109/ACCESS.2023.3310885.

[3] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo and A. F. Dragoni, "Deep Learning for Automatic Violence Detection: Tests on the AIRTLab Dataset," in IEEE Access, vol. 9, pp. 160580-160595, 2021, doi: 10.1109/ACCESS.2021.3131315.

[4] S. Jianjie and Z. Weijun, "Violence Detection Based on Three-Dimensional Convolutional Neural Network with Inception-ResNet," 2020 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Shenyang, China, 2020, pp. 145-150, doi: 10.1109/TOCS50858.2020.9339755.

[5] W. -F. Pang, Q. -H. He, Y. -j. Hu and Y. -X. Li, "Violence Detection in Videos Based on Fusing Visual and Audio Information," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 2260-2264, doi: 10.1109/ICASSP39728.2021.9413686.

[6] Jo, Jun-Mo. (2019). Effectiveness of Normalization Pre-Processing of Big Data to the Machine Learning Performance. The Journal of the Korea institute of electronic communication sciences. 14. 547-552. 10.13067/JKIECS.2019.14.3.547.

[7] Li, Ji & Jiang, Xinghao & Sun, Tanfeng & xu, ke. (2019). Efficient Violence Detection Using 3D Convolutional Neural Networks. 1-8. 10.1109/AVSS.2019.8909883.

[8] Setiaji, Hari & Paputungan, Irving. (2018). Design of Telegram Bots for Campus Information Sharing. IOP Conference Series: Materials Science and Engineering. 325. 012005. 10.1088/1757-899X/325/1/012005.