# Enhanced Violence Detection Using Deep Learning.

Praveen Bhawantha*
*Department of Computer Systems Engineering*
*Faculty of Computing and Technology*
*University of Kelaniya*
Sri Lanka
1praveenbhawantha@gmail.com

S. P. Kasthuri Arachchi
*Department of Software Engineering*
*Faculty of Computing and Technology*
*University of Kelaniya*
Sri Lanka
sandelik@kln.ac.lk

***Abstract*— Global violence needs to be stopped to increase public safety. With the increasing number of surveillance cameras, manual monitoring of all surveillance feeds is less practical. Because of that, the development of technology-driven solutions to detect real-time violence and inform authorities to prevent it has become necessary. This study focuses on finding a novel deep learning approach to enhance violence detection, specifically addressing the limitations and complexities of previous studies. Notably, the research utilizes proposed models and techniques to evaluate real-life violence scenarios captured in Closed-Circuit Television (CCTV) footage, overcoming the challenges identified and improving the accuracy of violence detection. Two models were proposed in this research paper. The model architecture consists of a multimodal approach, integrating two deep learning techniques, Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). The proposed model utilizing VGG-16 with CNN layers and LSTM, achieved 89% accuracy on the real life violence situations dataset. This emphasizes the effectiveness of applying multimodal deep learning technique in detecting violence, outperforming similar research in accuracy.**

**Keywords—Violence detection, Deep Learning, CNN, LSTM, VGG-16, CCTV, Real-life Violence**

## I. Introduction

The World Health Organization (WHO) defines violence as the intentional use of physical force or power against oneself, another person, or a group or community, resulting in injury, death, psychological and making harm, maldevelopment, or deprivation [1]. Violence affects society significantly. It can be either be intentional or unintentional. Violence cause 4.4 million deaths and that’s 8% of all deaths happening worldwide in 2021. From age 5-29, 60% of the top causes of death are injury related. Violence costs billions of US dollars in healthcare facilities, law enforcement, and lost productivity [2].

CCTV is a crucial IoT device that has been instrumental in reducing violence and enhancing security. Around 85.95% of the current population have smartphones. These devices help capture and report incidents, enabling authorities to take necessary actions. CCTV cameras also create a sense of safety, encourage responsible behavior, and aid in investigations, making communities safer [3]. But relying solely on human force to identify all the violence incidents in real time seems impractical as it takes valuable time to report these incidents and sometimes the damage is already done before taking an action. Also, it often takes too long to detect, search, and arrest someone after a crime is committed by analyzing the past surveillance recordings to identify instigators and culprits.

To Avoid these problems, implementing automated detection systems powered by AI and deep learning offer a promising solution. Current violence detection methods mainly based on two approaches which are traditional machine learning approach and deep learning methods. Traditional Machine Learning technology has shown good results in violence detection, but there are limitations to address. Some studies focus on general activity detection, while others explore fight scenes, raising concerns about accuracy compared to accurate violence detection from CCTV datasets. Technologies, like Latent Dirichlet Allocation (LDA), Support Vector Machines (SVMs), MoSIFT, HoG + HoF, and Harris corner detection, have been suggested for violence detection, but their limitations such as sensitivity to noise, computational intensity, and real-world scenario handling directly affect the efficiency [4].

Using deep learning techniques showed promising results compared to the Traditional Machine Learning methods, but due to these issues those approaches didn’t seem to be the best option. Limited Dataset Diversity and Realism occurs when using datasets that are staged, hockey fights, movie fights which wouldn’t suit for training the model to detect violence on CCTV. The Insufficiency of Data Quantity and Volume causes having a low accuracy and having less generalized graphs. Due to the complexity of the models that were used such as transformers and 3D-CNN there’s a problem in implementing these technologies into CCTV systems practically which is not good for carbon footprint as well in creating smart cities. And having constrained number of frames and low-resolution during training the data makes the models reduce in accuracy in recognizing the fine details and also the temporal patterns. Because of these a new approach needed to be implemented for better analysis and results.

This study aims to develop a novel violence detection method utilizing real life violence and provide authorities with the necessary details to prevent violent events while promoting sustainability in smart cities. The objectives include finding a better dataset containing CCTV violence footage and real fight scenes, using different preprocessing methods to extract features, creating the best model to detect accuracy, by analyzing existing methods to compare and improve.

## II. Related Work

Traditional Machine Learning methods for detecting violence often rely on machine learning algorithms and handcrafted features. These methods can predict violence occurrences using low-level features like human trajectories and motion information. However, they face limitations in handling complex real-world scenarios and achieving reliable results in rapidly changing environments. Machine learning techniques like SVM, Hidden Markov Model, and Binary Location Motion Patterns can lead to bottlenecks such as limited feature extraction, lack of adaptability to real-world scenarios, and difficulty processing high-dimensional data spaces [5].

Deep learning techniques, such as CNN and LSTM, are effective in capturing spatial and temporal patterns in videos. These techniques, combined with deep learning methods, can result in high accuracies, making them popular for violence detection systems and was used in previous works. These techniques are particularly useful in capturing temporal patterns in video data.

CNN can be considered as a great method to perform image processing task which extracts the intricate features from video frames making it available for efficient violence detection. As versions of CNNs, ResNet50, VGG-16 (Visual Geometry Group with 16 layers), and Inception models are recognized for their distinctive architectures, excelling in capturing spatial information and hierarchical characteristics [5]. CNN performs better at capturing spatial features, but sometimes just using that is often not enough to give better results. Temporal information within sequential data is crucial for violence detection. LSTM networks analyze sequential patterns and long-term dependencies, providing valuable insights into the temporal dynamics of actions and enabling the detection of violent events over time. This comprehensive approach offers a better solution to enhance violence detection [6].

A combination of these deep learning techniques has used in previous works, such as using techniques like ConvLSTM where CNN and LSTM were combined [7]. Using techniques like C3D and YOLO-v3 were also there in previous works which was specifically used for identifying objects furthermore [8]. Using pretrained models such as ResNet50+ LSTM [9], MobileNetV2+Bidirectional LSTM to detect violence was also approached. Some other approaches that were taken are 3DCNN+LSTM and powerful models like transformers that also paved path to new ways of detecting violence.

Previous research have utilized various datasets for violence detection, including movie fights [10], hockey fights [11], and staged fights [12,13]. These datasets have been divided into trimmed and untrimmed collections, with varying video duration and annotation granularity. The BEHAVE dataset, developed by Brunsden et al. [12], features groups of 2 to 5 people engaged in various interactions. The Movies Fight and Hockey Fight datasets contain annotated video snippets from movies and hockey games, respectively [10,11]. Hassner et al.'s [14] Crowd Violence dataset specializes in detecting violence in crowded environments. The RGB-D dataset by Yun et al. [15] for violence detection records human interactions with depth information using the Microsoft Kinect sensor. The RE-DID dataset consists of real-life events taken through cameras and other devices, providing high-quality films with full annotations [16]. The RWF-2000 dataset aims to address the shortcomings of earlier collections by providing a more practical and diversified resource for violence detection research [17]. Some datasets contain natural and surveillance fights, which are comparatively better than previous ones as they are captured from real scenarios.

Smart-City CCTV Violence Detection (SCVD) [5] which contained real life violence situation CCTV footages and Real Life Violence Situation Dataset [14] which contained real life violence situations turned out to be the best datasets when compared to other dataset and was also utilized for testing in this research. Table I. shows the comparison between the datasets used in violence detection.

TABLE I. Comparison Between the datasets used for Violence Detection.

| *Dataset* | *Data Scale* | *Resolution* | *Scenario* |
|---|---|---|---|
| BEHAVE [12] | 171 clips | 640 × 480 | Acted fights |
| RE-DID [16] | 30 videos | 1280 × 720 | Natural |
| Hockey Fight [10] | 1,000 clips | 360 × 228 | Hockey Games |
| Movie Fight [11] | 200 clips | 720 × 480 | Movies |
| SBU Kinect Interaction [13] | 264 clips | 640 × 480 | Acted Fights |
| SCVD [5] | 500 clips | Variable | Real life violence |
| Real Life Violence Situation [14] | 2,000 clips | Variable | Real life violence |

## III. Approach

The proposed violence detection method combines CNNs and LSTM networks to capture spatial and temporal information in video data. The CNN-LSTM Model is a hybrid architecture that captures both spatial and temporal information, making it well-suited for violence detection. The VGG16-LSTM model combines the CNN architecture with LSTM layers for violence detection in video data. The method's input consists of video clips from real-world scenarios, preprocessed to extract consecutive frames. Data augmentation techniques improve model generalizability. The generated frame sequence serves as input data for deep learning models, producing a binary classification label indicating the presence or absence of violence. Two datasets were used for this research which are SCVD dataset and Real Life Violence Situations Dataset [5,14].

### A. Data Preprocessing

The data preprocessing pipeline in CNN-LSTM model involves extracting sequential frames from video clips using the OpenCV library. Data augmentation techniques, such as cropping frames to 64×64 pixels and horizontally flipping frames, enhance the model's robustness. To improve accuracy, 30 frames were extracted and checked. Seed values are set for random number generators to ensure reproducibility. Classed were defined as “violence” and “nonviolence”.

In VGG16-LSTM model, preprocessing starts by extracting sequential frames from the video clips and applying data augmentation such as horizontal flipping. Normalization of pixel values is used to ensure the consistency of all the frames. Each video consists of 20 consecutive frames, and the size of an input frame is 224×224 pixels.

The datasets were split into 80% of training and 20% and was changed slightly to achieve the best accuracy.

### B. Feature Extraction

In CNN-LSTM each frame was enlarged to a consistent size 64×64 and normalized by dividing pixel values by 255. The CNN architecture, consisting of convolutional and pooling layers, extracts spatial patterns from processed frames. The output is sent to the LSTM network, which learns temporal connections from CNN feature maps, capturing motion and context information.

VGG16-LSTM model’s frames are preprocessed by shrinking them to uniform size 224×224 and normalizing pixel values to range [0,1]. The VGG16 model, trained on ImageNet, is used as a feature extractor. The VGG16 network output is routed through to the last fully connected layer (fc2), which extracts high-level abstract features from frames. These features are fed into the LSTM network, which learns temporal relationships, recording motion and context data over time.

### C. Model Architectures

The Convolutional Block of CNN-LSTM Conv2D layer uses 16 filters with 3×3 spatial dimensions, non-linearity, and dropout to process each frame independently, ensuring output feature maps have the same spatial dimensions as the input and the model incorporates a 32-unit LSTM layer to capture temporal dependencies and patterns in the sequence of extracted features from convolutional blocks. Fig. 1. shows the full architecture of the CNN-LSTM Model.

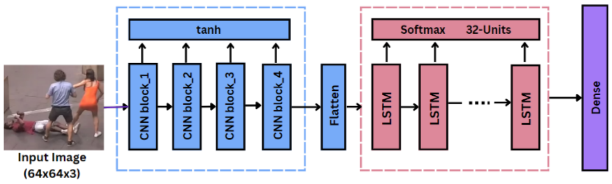

Fig. 1. *Full Architecture of CNN-LSTM Model*

The VGG16 model's Convolutional Block comprises five blocks with convolutional layers and max-pooling layers, reducing input spatial dimensions and increasing filters to capture hierarchical features, reducing channels. The LSTM layer in this model extracts features from the VGG16 model, producing a 3D tensor with dimensions of "(None,512)" and a tensor with dimensions of "(batch size, 512)" for further processing. Fig. 2. shows the architecture of VGG-16 and Fig. 3. Shows the full architecture of the VGG16-LSTM Model.

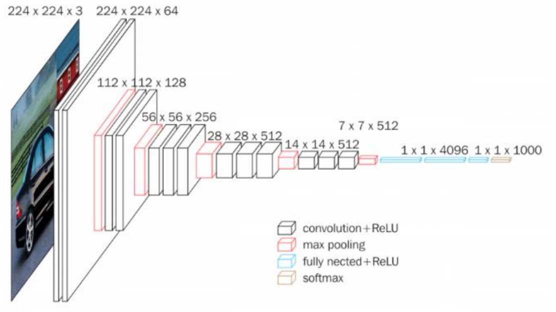

Fig. 2. *Architecture of VGG-16 Model [20]*

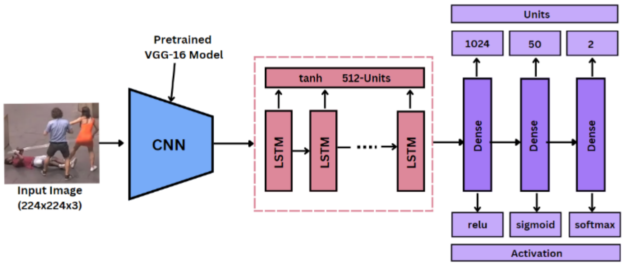

Fig. 3. *Full Architecture of VGG16-LSTM Model*

## IV. Results

The experiments were carried out using different approaches. Firstly, there was only one model and one dataset “SCVD dataset” [5] and the CNN-LSTM model. Due to poor quality of the dataset a new dataset was utilized which was, “Real Life Violence Dataset” [14]. Moreover, another method was approached to check which method gives the best results out of them which is the VGG16-LSTM approach.

The CNN-LSTM model was used with SCVD dataset and the results weren’t good and the main reasons were the quantity and the quality of the dataset were low as well as there were only 500 videos containing violence and nonviolence altogether. The best accuracy obtained from using this dataset was 82% and the loss was 58.16%. And the graphs that were given as the results were not generalized well.

Because of this, the Real Life Violence Dataset was tried out with CNN-LSTM Model, and the main goal was to achieve a good generalized model with better accuracy. Several changes were tried to check the performance and Table II. shows the changes that were done and the accuracy and loss according to that.

TABLE II. Accuracy and Loss with the Changes Made

| *Test* | *Accuracy* | *Loss* |
|---|---|---|
| Initial State | 79% | 47% |
| Running model for 100 epochs | 80.15% | 43.16% |
| Changing preprocessing techniques | 81.75% | 40.64% |
| Changing the frames from 20 to 30 | 86.60% | 47.07% |
| Running for 100 epochs + 30 frames | 87% | 40.42% |

By doing these changes, compared to the previous dataset, better results were achieved. But still, the generalization problem was there. Fig. 4. shows that the graphs are not generalized even at their best accuracy and loss results.

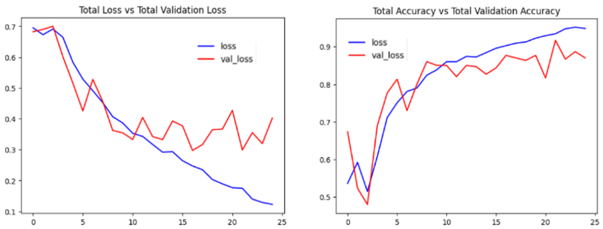

Fig. 4. *Best Results graphs of Real Life Violence Dataset in CNN-LSTM Model.(a) Total Loss graph respective to its validation (b) Total Accuracy graph resective to its validation.*

Since there was an issue with the generalization, another approach was taken to get better accuracy and a better generalized graph. Because of that, VGG16-LSTM Model was tried out. Table III. shows the changes that were made and the accuracies and the losses over the changes in batch size and epochs.

TABLE III. ACCURACY AND LOSS WITH THE CHANGES MADE

| *Batch Size* | *Epochs* | *Accuracy* | *Loss* |
|---|---|---|---|
| 4 | 200 | 85.35% | 12.15% |
| 32 | 200 | 86.55% | 8.16% |
| 256 | 200 | 84.65% | 13.16% |
| 512 | 200 | 82.15% | 16.54% |
| 32 | 500 | 88.50% | 3.43% |
| 32 | 100 | 89% | 3.37% |

Out of the tried methods, the best results were given when the batch size was 32 and in 100 epochs. Figure. 5. demonstrates the best results that were obtained by doing the experimental setup and these graphs were generalized when compared to the previous experiments carried out in CNN-LSTM Model.

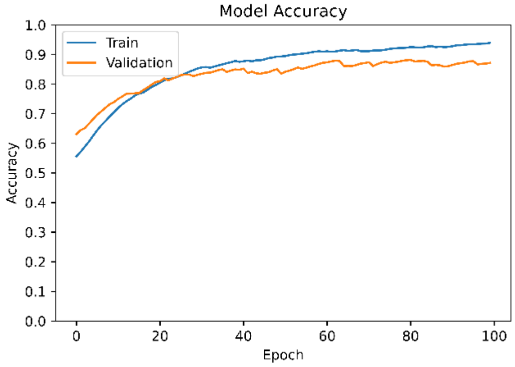

Fig. 5. *(a) Accuracy Graph respective its validation of the Best results of Real Life Violence Dataset in VGG16-LSTM Model.*

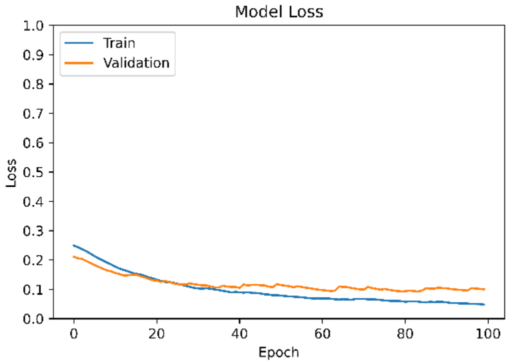

Fig. 5. *(b) Loss Graph respective to its validation of the Best Results of Real Life Violence Dataset in VGG16-LSTM Model*

Finally, after getting the results, the model was tried out by getting sample videos and testing it out, this was shown frame-by-frame violence detection as demonstrated in Fig. 6. and Fig. 7. In Fig. 6. a video with violence is taken and split the frames and shown each frame whether it contains a violent act or not and in Fig. 7., a video of Nonviolence is taken and this is shown as a video in the notebook.

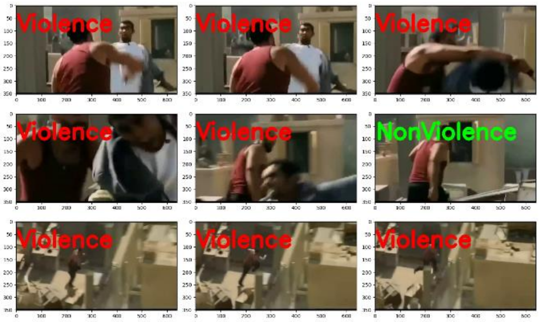

Fig. 6. *Model testing for violence video in VGG16-LSTM.*

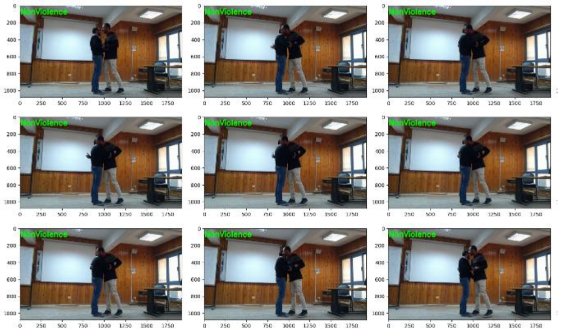

Fig. 7. *Model testing for Non-violence video in VGG16-LSTM*

## V. DISCUSSION

Training the model on SCVD dataset: Using the SCVD dataset was a good approach, however using it with the model CNN-LSTM presented better accuracy from around 80%, but the generated graphs of accuracy and loss respective to the valid loss and accuracy didn't generalize. The arrived conclusions for this could be due to the videos being of less diversity, and the dataset being skewed to generalize a broader range of situations. Data augmentation methods didn't help improve the model as well and it showed that a dataset containing more videos would present better graphs.

Based on the results gained training the model on Real life Violence Dataset, the performance of VGG16-LSTM model was better than the CNN-LSTM model approach, and that is because in comparison the VGG16-LSTM has the power to get a good feature extraction, as VGG16 model is a pre-trained deep CNN that was trained on a large dataset (ImageNet) and it is designed to extract high level of features, and also from the videos the CNN-LSTM trained using 64×64 resolution but VGG16-LSTM was trained in 224×244 resolution. Fig. 8. shows the difference between 224×224 and 64×64 resolution. As 64×64 image is pixelated that would cause issues in giving good results and training.

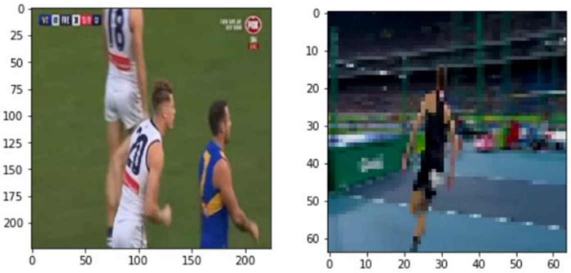

Fig. 8. *Difference between 224×224 and 64×64. (a) shows the 224×224 extraction and (b) shows the 64×64 extraction.*

Also, VGG16 model has multiple layers and there were a huge number of parameters trained using it. Comparing only the VGG-16 Model without adding the LSTM layers it had, 138,357,544 parameters while the 1st model only has 72,994 combined with both CNN and LSTM.

Therefore, it can be concluded that the VGG16 based model is far more complexed in comparison and can present more accuracy and better generalized graphs, however a simpler structure could be easily prone to underfitting and overfitting.

### A. *TensorBoard Utilization*

TensorBoard, an opensource visualization toolkit offered by TensorFlow, played a major role in monitoring and visualizing the performance metrics like accuracy, loss and validation across different epochs in real-time [23]. This saved a lot of time by being able to identify the applied changes over the model behavior so that the training process can be stopped in the middle if the graphs doesn't get good results such as low accuracy, high loss and even not generalizing. Fig. 9. shows one of the accuracy and lost results that was gained using TensorBoard in real-time and since it showed that it's not going to be generalized specially comparing the loss respective to the validation, this helped to conclude that the changes made aren't going to give a good result without waiting till the end of training. Most of the times it was possible to stop the process without waiting till the end thanks to the real-time results.

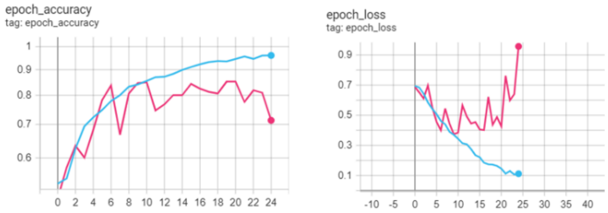

Fig. 9. *TensorBoard visualization of one of the tests carried out. (a) Total accuracy graph (b) Total Loss graph compared to their validations.*

### B. *Comparison with prior Research*

In prior research comparisons, mainly the exact dataset being used was considered. There were researches that approached the same violence detection dataset of "Real Life Violence detection" and a method that used the architecture MobileNetV2+BidirectionalLSTM that achieved an accuracy of 93%. But the model's main problem was the extracted frames, which was 16 frames per video plus the resolution was 64×64. In comparison with my approach, 224×224 enabled a more detailed view of the video frames of the data set. Therefore, despite showing better accuracy rates, the quality of the model accuracy prediction is a problem.

Another approach using the same dataset achieved a highest accuracy of 96% [21]. But the main problem was that its technique uses a method called transformers, and with its need of high computational power it is not suitable to capture CCTV footages live to provide violence detection. Even though they produce great accuracies [18], Implementing such a model will be more costly than using Deep Learning method. Plus, fixing issues will be difficult due to its complexity [19].

When comparing with other researches using different datasets and there were some researches that presented good accuracies after testing despite having a smaller number of videos causing less quality of the prediction, and some datasets only contained movie fights, hockey fights, staged acts which will provide good accuracies but they cannot be compared with real life violence scenarios.

## VI. Future Directions

The search for suitable CCTV violence datasets remains a significant challenge due to ethical and privacy concerns. So, it is a must to use a dataset where, they include real life violence incidents and having the concerning about the ethical and privacy concerns. The pursuit of less complex models that can be effectively implemented in real-time CCTV systems. This could really help in making the implementation easy and expand it due to the low cost and feasibility. This will also help to reduce the carbon footprint in smart cities due to less computational demand. Using methods to detect audio on the environment and training a model according to that would be beneficial as some violence scenarios starts verbally. Integrating algorithms like YOLO will broaden the violence detection capabilities as it will help to identify objects and that would become handy in detecting gun violence.

Fig. 10. shows the implementation of the violence detection system and this can be modified by applying the things mentioned previously.

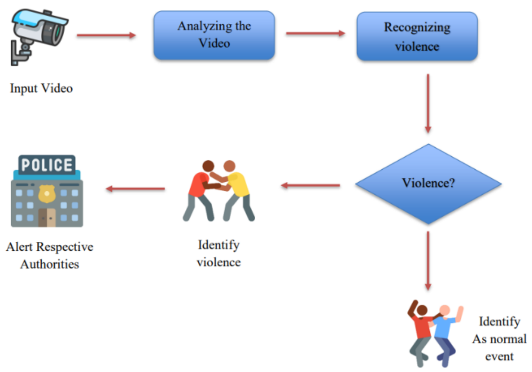

Fig. 10. *Model testing for Non-violence video in VGG16-LSTM*

## VII. Conclusion

Initially, two distinct deep learning architectures were analyzed; CNN LSTM and VGG16 LSTM. The VGG16 LSTM model merges VGG16 architecture with LSTM layers utilizing pre trained image recognition capabilities to enhance violence detection. In terms of accuracy and generalization across scenarios, the VGG16 LSTM model outperformed the CNN LSTM model.

A crucial aspect of our research involved preprocessing and augmenting the CCTV footage dataset as ensuring data quality and diversity is essential, for training models.

Using CCTV footage, violence identification can be advanced and future studies can explore the fusion of model data sources like audio and text to gain deeper insights for analysis. And using these techniques and implementing a solution will help to eradicate violence from occurring and create a better society.

## Acknowledgement

I am thankful to everyone who provided invaluable support in the development of my research article. I'd like to convey my heartfelt gratitude to my dedicated supervisor, Dr. S. P. Kasthuri Arachchi, whose direction, constructive input, and unwavering encouragement helped shape the course of this study. Special acknowledgment is given to the founders and maintainers of the datasets used in this study. I also would like to acknowledge the Faculty of Computing and Technology at the University of Kelaniya for creating an environment that promotes academic exploration. This research paper stands as a testament to the collective efforts and support of everyone involved.

## References

[1] S. Hamby, “On defining violence, and why it matters,” Psychol. Violence, vol. 7, no. 2, pp. 167–180, 2017, doi: 10.1037/vio0000117.

[2] “Injuries and violence.” https://www.who.int/news-room/fact-sheets/detail/injuries-and-violence (accessed Aug. 07, 2023).

[3] “New Technology and the Prevention of Violence and Conflict - Stability: International Journal of Security and Development.” https://stabilityjournal.org/articles/10.5334/sta.cp (accessed Aug. 07, 2023).

[4] M. Chen and A. Hauptmann, “MoSIFT: Recognizing Human Actions in Surveillance Videos,” CMU-CS-09-161, Jan. 2009.

[5] T. Aremu, L. Zhiyuan, R. Alameeri, M. Khan, and A. E. Saddik, “SSIVD-Net: A Novel Salient Super Image Classification & Detection Technique for Weaponized Violence,” 2022, doi: 10.48550/arXiv.2207.12850.

[6] N. Mumtaz *et al.*, “An Overview of Violence Detection Techniques: Current Challenges and Future Directions.” arXiv, Sep. 21, 2022. Accessed: Aug. 07, 2023. [Online]. Available: http://arxiv.org/abs/2209.11680

[7] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, “Violence recognition from videos using deep learning techniques,” in 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), 2019, pp. 8085.

[8] S. Vosta and K.-C. Yow, “A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras,” Appl. Sci., vol. 12, no. 3, p. 1021, Jan. 2022, doi: 10.3390/app12031021.

[9] M. Sharma and R. Baghel, “Video surveillance for violence detection using deep learning,” Advances in Data Science and Management, pp. 411–420, 2020. [Online]. Available: https://link.springer.com/chapter/ 10.1007/978-981-15-0058-9 35

[10] “Hockey Fight Detection Dataset,” Academic Torrents. https://academictorrents.com/details/38d9ed996a5a75a039b84cf8a13 7be794e7cee89 (accessed Aug. 07, 2023).

[11] “Movies Fight Detection Dataset,” Academic Torrents. Accessed: Oct.01,2023.[Online].Available: https://academictorrents.com/details/70e0794e2292fc051a13f05ea6f5 b6c16f3d3635

[12] S. Sandıkcı, S. Zinger, and P. H. N. de With, “Detection of Human Groups in Videos,” in Advanced Concepts for Intelligent Vision Systems, J. Blanc-Talon, R. Kleihorst, W. Philips, D. Popescu, and P. Scheunders, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2011, pp. 507–518. doi: 10.1007/978-3-642-23687-7_46.

[13] K. Yun, J. Honorio, D. Chattopadhyay, T. Berg, and D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning. 2012, p. 35. doi: 10.1109/CVPRW.2012.6239234.

[14] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, “Violence Recognition from Videos using Deep Learning Techniques,” 2019 Ninth Int. Conf. Intell. Comput. Inf. Syst. ICICIS, pp. 80–85, Dec. 2019, doi: 10.1109/ICICIS46948.2019.9014714.

[15] K. Yun, J. Honorio, D. Chattopadhyay, T. Berg, and D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning. 2012, p. 35. doi: 10.1109/CVPRW.2012.6239234.

[16] P. Rota, N. Conci, N. Sebe, and J. M. Rehg, “Real-life violent social interaction detection: IEEE International Conference on Image Processing, ICIP 2015,” 2015 IEEE Int. Conf. Image Process. ICIP 2015 - Proc., pp. 3456–3460, Dec. 2015, doi: 10.1109/ICIP.2015.7351446.

[17] T. Hassner, Y. Itcher, and O. Kliper-Gross, Violent flows: Real-time detection of violent crowd behavior. 2012, p. 6. doi: 10.1109/CVPRW.2012.6239348.

[18] “What Are Transformers in NLP: Benefits and Drawbacks.” https://blog.pangeanic.com/what-are-transformers-in-nlp(accessed Aug. 19, 2023).

[19] W. Lin, “Drawbacks of Transformers,” Apr. 2023.

[20] “VGG-16 | CNN model,” *GeeksforGeeks*, Feb. 26, 2020. https://www.geeksforgeeks.org/vgg-16-cnn-model/ (accessed Aug. 07, 2023).

[21] A. R. Abdali, “Data Efficient Video Transformer for Violence Detection,” in *2021 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, Jul. 2021, pp. 195–199. doi: 10.1109/COMNETSAT53002.2021.9530829.

[22] S. M, G. S, J. R. Fenitha, and S. R, “Fight Detection in surveillance video dataset versus real time surveillance video using 3DCNN and CNN-LSTM,” in *2022 International Conference on Computer, Power and Communications (ICCPC)*, Dec. 2022, pp. 313–317. doi: 10.1109/ICCPC55978.2022.10072291.

[23] “TensorBoard,” TensorFlow. Accessed: Oct. 03, 2023. [Online]. Available: https://www.tensorflow.org/tensorboard.